Prediction of tissue-specific cis-regulatory sequences: application to the ascidian Ciona intestinalis and the anterior neurectoderm

Maximilian Häussler

Université Paris XI

Discipline *Biologie Cellulaire et Moléculaire*

École doctorale *Gènes, Génomes, Cellules*

Thèse

pour obtenir le grade de

Docteur de l'Université Paris XI

Soutenance prévu le 15. Juillet 2009

par Maximilian Häussler

Prédiction des séquences cis-regulatrices tissu-spécifiques: application à l'ascidie *Ciona intestinalis* et au neurectoderme antérieur

Prediction of tissue-specific cis-regulatory sequences: application to the ascidian *Ciona intestinalis* and the anterior neurectoderm

---

Jury

President	M. Pierre Capy
Rapporteurs:	M. Nicolas Pollet
	M. Sebastian Shimeld
Examinateur:	M. Elia Stupka
Directeur de thèse:	M. Jean-Stéphane Joly

Summary

The detection and annotation of cis-regulatory sequences is a difficult problem. There is currently no generally applicable experimental procedure or computational algorithm to identify the non-coding regions of the genome that serve to activate gene expression in a given cell type. The only indicator of cis-regulatory function is the conservation of a sequence in other genomes. Regions can then be tested one-by-one in transgenic assays but this is time-consuming in vertebrates. Only a limited number of these already validated cis-regulatory sequences have been curated in biological databases. One of the main advantages of the model organism Ciona intestinalis is that cis-regulatory tests can be conducted very easily and the result is observable after one day while the animal follows the chordate body plan. However, a sequence found to be active in this organism can currently not be mapped to genomes of other animals.

In my thesis, I first established a procedure to rank combinations of short sequence motifs by their distribution around a set of genes. The better a combination matches around genes expressed in a certain tissue, the higher is its score. I applied this to an already characterized enhancer of C. intestinalis expressed in the anterior neurectoderm which had been found by systematic mutations to be composed of a duplicated structure. The results of my procedure indicated that duplicated GATTA-sites are an essential feature of cis-regulatory elements active in the anterior neurectoderm. Searching the genome for matches to this signature resulted in putative enhancers that drive a reporter gene in 50% of the cases in the anterior neurectoderm. This is a relatively high proportion compared to much more complex prediction approaches reported in the literature.

In addition, I tried to improve the curation of already published cis-regulatory elements by extracting them automatically from the full text of the biological research articles. Thanks to the thriving open access publishing model and the improvement in experimental assays, more and more of this data is becoming available.

Finally, I showed that in the absence of sequence alignments between vertebrates and C. intestinalis, one can nevertheless find a handful of loci with a very unusually conserved gene order. In these cases, the cis-regulatory search space is reduced to a set of introns, some of which were recently shown to harbor enhancers. Many of these loci have not been analyzed yet.

Together, these computational approaches should lead to a better characterization of cis-regulatory sequences and pave the way for further experimental validations.

Acknowledgments

The great thing about human language is that it prevents us from sticking to the matter at hand
Lewis Thomas, "The Lives of a Cell", p 95:

As most PhD students, writing this part just before handing the file over to the printing office, I am surprised that this thesis actually got finished. Like many, I often thought about canceling the project. Research, on ascidians or anything else, is first and foremost based on the interaction between ideas of very different human beings. The pile of printed paper that you are reading would have not been possible without the support of...

- Jean-Stephane, who had the idea of the project and took the risk of recruiting a foreigner with little French and no biological background: For his passion for science and permanent efforts of organizing very good funding for his group and a confused PhD student like me
- Florian, who taught me Molecular Biology: for his realism on science and the many hours he spent to teach a total beginner all the molecular biology essential for the completion of the article and this thesis, while receiving very little support himself and being only acknowledged here.
- Yan, who actually did most of the Molecular Biology in our article: For his calm, even during times when all experiments had failed and his insistence afterwards, often with success
- Helene, the third student, who become much more than a collegue in the end: For her ability to quickly and calmly solve many problems and helping me to get through the last part of my thesis

I would not be working in biology, live in Paris or speak French without Ute although she is probably not aware anymore of her influence in all of this and hopefully happily cycling somewhere in Ukraine today.

Being always able to count on one's parents is the biggest resource on which a risky endeavor like a thesis is built. Starting as a foreigner in a new country and a new scientific discipline changes the support network in a way that makes the work environment and family far away more important than under “normal” circumstances. I relied on friends from the lab, for all non-biology related parts of life, like renting an apartment, sports, cooking, shopping etc. So I am happy to have met Kei, the only female football player in all the institutes in and around Gif, who is not afraid of anything and amazingly never seems to be homesick, unlike me. Thanks to Joana for one apartment, Charlotte, the best flatmate ever, for another apartment. Aurelie and Matthieu for her imperturbable and intensive commitment to research. Also to Stephane and his football team who accepted me as the only player, who never scored a single goal in four years but nevertheless played every week. Thanks also to the people at INRA Muriel Mambrini and Micaela Galozzi, for yet another apartment and one good example of how to finish a thesis in exactly three years, by the day, which I didn't achieve. Our elders, Marylin, Alessandro, Pierre and Jean-Michel could have taught me how not get let down by research and the environment, but I still have a lot to learn there.

Part of this thesis would have not been even started without my main collaborator Casey Bergman, who provided a lot of motivation and always professional guidance on these projects from Manchester. I also thank the two referees of this text, Sebastian Shimeld and Nicolas Pollet, as well as the other jury members, Elia Stupka and Pierre Capy, for their motivating remarks and corrections.

Table of Contents

Chapter 1: Introduction

> *Many years ago Prof. Goodsir perceived that the lancelet presented some affinities with the Ascidians, which are invertebrate, hermaphrodite, marine creatures permanently attached to a support. They hardly appear like animals, and consist of a simple, tough, leathery sack, with two small projecting orifices. They belong to the Molluscoida of Huxley—a lower division of the great kingdom of the Mollusca; but they have recently been placed by some naturalists amongst the Vermes or worms. Their larvæ somewhat resemble tadpoles in shape, and have the power of swimming freely about. Mr. Kovalevsky has lately observed that the larvæ of Ascidians are related to the Vertebrata, in their manner of development, in the relative position of the nervous system, and in possessing a structure closely like the chorda dorsalis of vertebrate animals; (…)*
>
> *Charles Darwin, On the origin of species, p 159, 2nd ed, 1881*

This chapter introduces:

- the model organism *C. intestinalis* and its development
- some particularities of its genome
- a literature overview on the screening of cis-regulatory sequences. (This section is written in a way to be submitted to a journal as a review after the thesis defense)

1.1 The model organism C. intestinalis

The Urochordates in the tree of life

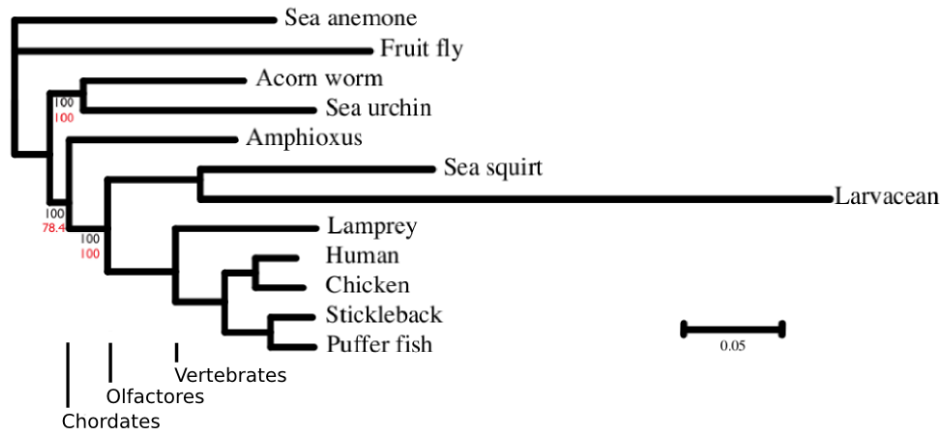


Figure 1: Annotated phylogenetic tree from (Holland2008), based on 1090 orthologous genes.

The bulk of biological research is conducted on vertebrates like mice and rats and to - a limited extend but increasingly - on fish. *Drosophila* and *C. elegans* have an accepted place in the lab but their developmental patterning processes bear little resemblance to vertebrates. To tackle questions on the mechanism of body patterning, one has to use other organisms. One of them are urochordates, more similar to humans than flies or nematodes, yet still easy to manipulate in the lab.

In the tree of life, the parent phylum of vertebrates are the chordates. Present-day chordates include vertebrates, amphioxus (cephalochordates, lancelet) and sea squirts (called urochordates or tunicates). Instead of a backbone, one of their common features is a stiff notochord in the dorsal part of the animal. It can be restricted to the tail (greek:uro) or also extend into the head (greek:cephalo). According to recent molecular comparisons, urochordates are the taxon phylogenetically closer to vertebrates than cephalochordates (Delsuc2006) and are sometimes grouped with them into the “olfactores” (Figure 1). Urochordates can also be called “tunicates” due to their “tunic”, a protective outer layer, on top of the epidermis, made of cellulose that is produced by an enzyme probably acquired by horizontal

gene transfer from bacteria (Nakashima2004)(Matthysse2004). They can be classified into larvacea, thaliacea and ascidiacea: the first two lead a free-swimming planctonic existence, the latter are sessile and comprise the lion's share, 2300 of the 3000, urochordate species (Satoh1994). These live in mostly shallow water all around the world. After fertilization of their eggs, they develop within 12 hours to some days into swimming tadpole-like embryos (without a mouth), that soon use their palps to attach to a solid substrate like rocks, shells or ship bottoms, to metamorphose into a barrel-like shape and start filter feeding.

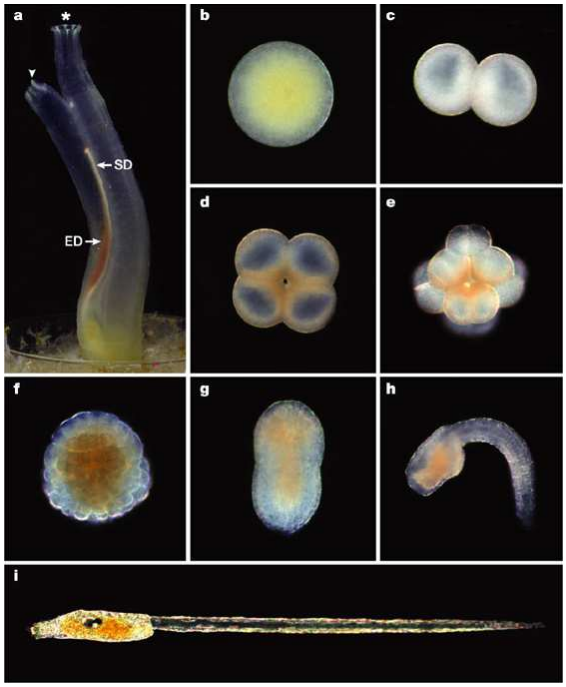


Figure 2: Developmental stages of C. intestinalis, copied from (Satoh2003): b) egg, b) 2-cell, d) 4- cell e) 16-cell, f) gastrula g) early-tailbud h) mid-tailbud i) tadpole larvae a) adult animal

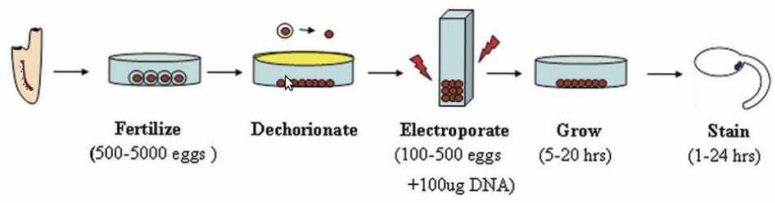


Figure 3: Overview of the protocol to electroporate eggs of C. intestinalis (from (Shi2005))

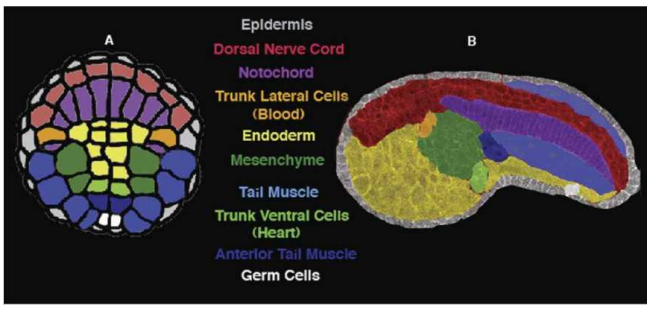


Figure 4: The main tissue types distinguishable at the tailbud stage in Ciona intestinalis, from (Shi2005)

Development of C. intestinalis

The most common and cosmopolitan species *C. intestinalis* has some obvious advantages as a model organism: at the facility level, the infrastructure to keep the animal consists of merely a refrigerated bucket of seawater and eggs develop into a swimming larvae within a day ((Satoh1994),

p4-8). Such a quick succession of developmental stages in a transparent body (Figure 2) has obvious advantages for the observation of developmental processes of chordates. In addition, *C. intestinalis* was among the first invertebrate organisms with a sequenced genome, DNA can be introduced into several hundred eggs in parallel by means of a relatively simple electroporation procedure (Figure 3) to over- or mis-express genes or trace the expression of promoters with standard reporters. Morpholino injections can suppress mRNAs and one successful application of siRNAs has been reported (Nishiyama2008). Thanks to the combined work of several Japanese and French research groups, >30.000 RNA in-situ hybridization images for more than 2500 genes at different stages can be downloaded from websites (Satou2005)(http://aniseed-ibdm.univ-mrs.fr).

On the other hand, ascidians are certainly rather derived chordates, their reproduction in the lab compared to other model organisms without sea water access is not straightforward (Liu2006)(Joly2007), egg production is seasonal (Joly2007) (as in *Ascidiella aspersa* (Chabry1887)) and reproductive capacity reached only at the age of 1-2 months. But, for the study of processes that are similar between ascidians and vertebrates, *C. intestinalis* represents a simple animal model with protein sequences relatively similar to vertebrates and a rich set of molecular tools.

The most interesting developmental time for inter-species comparisons is the “tailbud stage” (18 hours at 18°C) where *C. intestinalis* resembles a

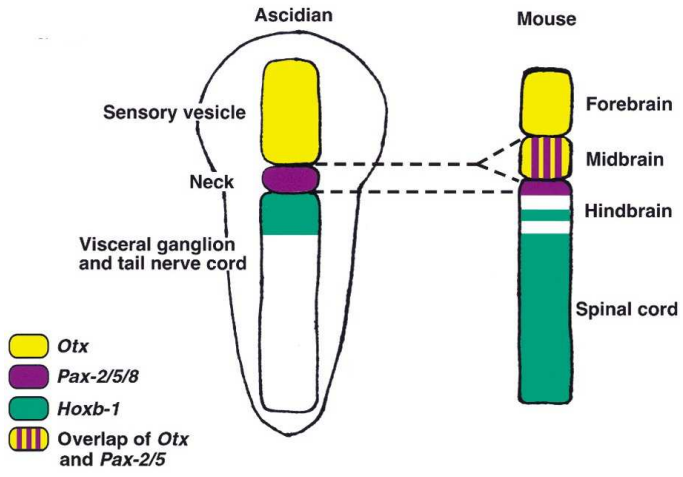


Figure 5: Expression pattern of key transcription factors along the anterior-posterior axis in ascidians and mouse, copied from (Corbo2001)

vertebrate tadpole but consisting of only ~2600 cells, as opposed to millions of cells in a frog embryo. At this stage, a handful of tissues are distinguishable with a binocular microscope (Figure 4). They include the tripartite central nervous system, and the notochord, flanked laterally by muscle and dorsally by the nerve chord. When assayed by in-situ hybridization, the expression pattern of key transcription factors follows the vertebrate scheme: muscle tissue is determined by Tbx6-genes (Yagi2005), and the heart field precursors are specified by Mesp, though possibly with different upstream activators (Satou2004) (Christiaen2009). Brachyury is sufficient and necessary for notochord induction, as in mouse embryos (Corbo1997) (Satoh2003), and nervous system tissue development depends on the expression of Otx (Wada1996) (Hudson2001). Whereas the upstream part of the early gene regulatory network of neuronal and notochord cells is quite different (Figure 6), the spatial arrangement of key transcription factors in the nervous system resembles vertebrates (Figure 5).

The case of the mouth, an evolutionary important feature for chordates as they are deuterostomes, is complicated by the fact that ascidian tailbud embryos are not feeding and completely lack a digestive system. There are rudiments, however: at the larval stage, the tip of the head includes a small invagination which gives rise to the oral siphon (Chiba2004) and expresses the stomodeal marker gene Pitx specifically. Therefore, the invagination is called stomodeum or oral siphon primordium (Christiaen2002). It develops

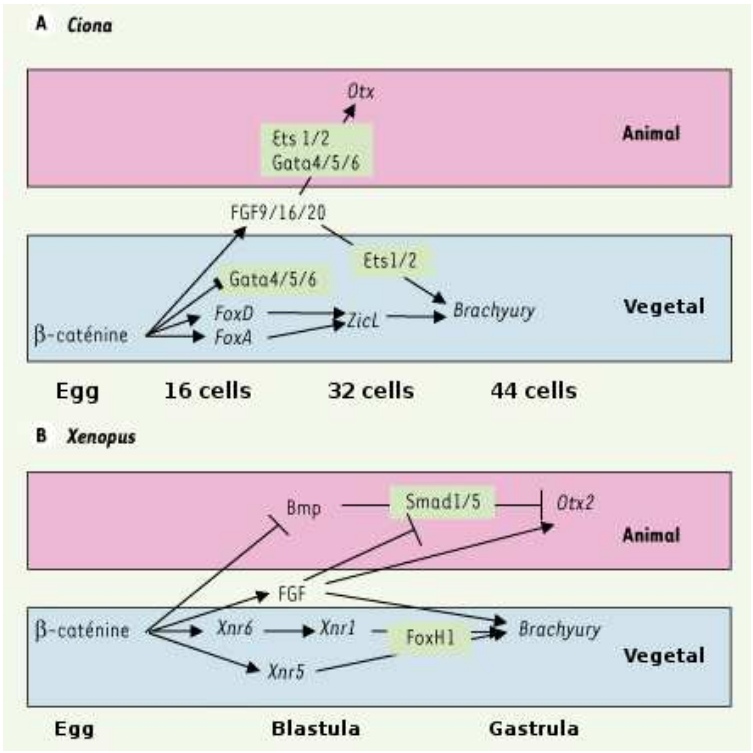


Figure 6: Gene regulatory network leading to notochord (rose) and anterior neural tissue specification (blue) in Ciona intestinalis and Xenopus, copied and translated to English from (Lamy2008)

from three cells at the mid-tailbud stage, located at the boundary of the anterior-most tip of the nervous system and the epidermis. This origin, in combination with the PITX expression pattern, suggests a similarity with the anterior neural ridge/boundary (ANB) in vertebrates. Thus the structure has been called ANB in ascidians at the mid-tailbud stage.

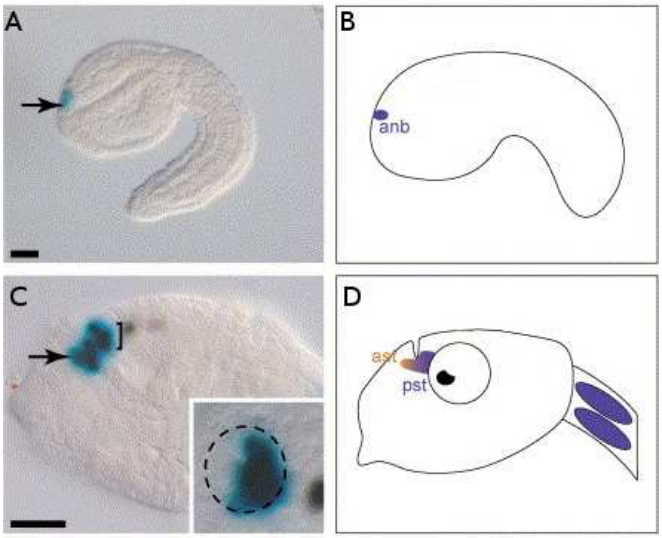


Figure 7: The ascidian anterior neural boundary marked by a transgenic construct of a 188bp enhancer cloned into the lacZ expression vector PCES, at the mid/late-tailbud stage (A, B). The stomodeum is marked at the larval stage (C,D). anb - anterior neural boundary, ast – anterior stomodeum, pst – posterior stomodeum

(Christiaen2005) have isolated a 188bp enhancer of the gene PITX that drives LacZ expression in the ANB (Figure 7). This was the only enhancer of this type at the time. Given its small size and the restricted expression pattern, we were interested to find similar cis-regulatory regions in the genome of *C. intestinalis.*

1.2 Ascidian genomes

1.2.1 A great diversity of genome assemblies

All large-scale non-coding sequence analyses require a genome sequence. The ascidian ones show some particularities which lead to different competing versions, created by various groups. To my knowledge, no other research model organism has seen a similar diversity of assembly and annotation approaches. In the following, I want to resume my experiences in annotating them during the last three years and give an overview of the literature on these genomes that has not been reviewed elsewhere.

Two *Ciona* species have been sequenced until now, *C. intestinalis* and *C. savingyi*. The first project was started independently by two different efforts, one in Japan (NIG) and one by the JGI. NIG sequenced three individuals collected at Onagawa with 5X coverage, the JGI sequenced one individual from the Half Moon Bay area with 8.5X coverage, accompanied by two BAC libraries and one cosmid library from other specimens. The initial assembly (JGI1, 2001) was published in 2002 (Dehal2002), but did not include the Japanese sequence reads. Consistent with previous work (Schmidtke1980), it reported a high heterozygosity rate, affecting around 1.2% of the nucleotides in the genome that differ between the two haplotypes of the single specimen that was sequenced. For all other genome projects at the time, either inbred laboratory lines were available (*Drosophila, C. elegans*) or the animals had a low population size (human, mouse, fugu). For this reason, heterozygosity was a rather new problem in 2001 (Vinson2005). The two strategies to cope with this problem were copied from the fugu genome: an increased overlap tolerance in the assembly process and combined with a final removal of duplicates (Dehal2002).

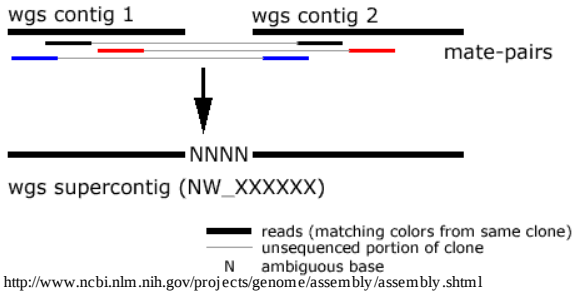
The next major version of the genome (JGI2, 2005) included the Japanese reads into the assembly, to a total 11X coverage, and also added BAC mapping data by FISH to place 65% of the resulting sequence scaffolds onto the 14 chromosomes (Shoguchi2005), resulting in much longer scaffolds. However, it was soon found that several known loci were missing from this assembly: Among them, Troponin I and the basal forkhead promoter, both examined in several publications. This raised the concern, pronounced at

the 2007 Tunicates meeting by Patrick Lemaire, that JGI2 can not be really called an improvement and that Ensembl should rather annotate the original version. While it is difficult to estimate the quality of an assembly in the presence of a very heterozygous population, one reasonable estimator of quality is how well the genome sequence fits existing gene data, as represented by cDNA/EST in Genbank.

A Short Genome Assembly Glossary (based on (Scheibye-Alsing2009)):

At most 1000bp of a DNA molecule can be sequenced at once. Therefore, a genome is sheared into small fragments of a defined size. Their 5' and 3' ends are sequenced and assembled by software:

If two of these **reads** overlap, they can be joined into a longer sequence. By repeating this, longer fragments, called **contigs,** are obtained. The puzzle cannot continue at regions where too there is too much overlap (repeated regions). Therefore, in the next step of the assembly process, **paired reads** (aka **mate-pairs**) obtained from DNA fragments longer >1kb, are used to order contigs into **scaffolds** or **supercontigs** (see illustration below). This leaves gaps between the individual contigs, but their size can be estimated from the fragment size when the DNA was sheared. In the final stage, scaffolds can be ordered by supporting evidence into yet longer sequences (sometimes called **ultracontigs,** for C. savingyi: **reftigs)**, by combining with data from separate projects such as the sequenced ends of cosmids or BACS, fully sequenced cDNAs (e.g. as found in Genbank) or from genetic linkage maps. These can then be assigned to chromosomes based on genetic markers or FISH assays. The human genome relied on cDNAs for a long time, the Medaka genome used mostly a genetic map, while the Fugu genome is staying at the scaffold stage. The different versions of the *Ciona* genomes used all combinations of these strategies (see text).



Assembly quality is expressed by the **N50** metric, the average scaffold size of the longest scaffolds which amount to 50% of the genome.

To this end, I queried the aligned NCBI sequences of UCSC genome browser database. The known and published genes missing from JGI2 amount to almost 6% of the genes in Genbank. They include FGF3/7/10/22, Smad2/3a, frizzled receptor, dead ringer homolog, and 46 others. According to all measures, JGI1 indeed corresponds better to the known cDNA data (See Table 1 which by and large corresponds to the data in (Satou2008)). It is unlikely that this is due to the origin of the ESTs as the majority was not

collected in the USA (30% are from animals harvested in Roscoff and the rest was obtained almost exclusively from Japanese animals, according to NCBI DbEST).

The root of the problem is probably the high rate of heterozygosity in combination with mixed DNA from several Japanese animals which confused the assembler in the case of JGI2. Even more as the assembly process did not model the at least 14 different genomes (two American and five Japanese animals) explicitly. Although a assembly of the two haplotypes of *C. intestinalis* separately was noted in the genome paper and deferred to a follow-up article, the data were never published. The idea was eventually put into practice by (Kim2007b), who unfortunately did not compare the results with the older assemblies and neither tried to improve the quality of JGI1 based on the two haplotype genomes.

Assembly Version	Cint JGI1	Cint JGI2	Cint KH	Csav1	Csav2
Total Assembly Size	117 MB	173 MB	115 MB	164Mb	174MB
N50 Size	0.187 MB	2.6 MB	5.2 Mb	1.05 MB	1.8 MB
Mappable Known Genes (RefSeq mRNA)	898	859			
Available ESTs	1,205,674	1,205,674	1,205,674	84,302	84,302
Alignable ESTs	1,103,805	1,059,959			
EST: Avg. identical nucleotides per EST	613	589			
EST: Avg. mismatches per EST	6.97	7.34			

Table 1: The three different versions of the C. intestinalis Genome and some statistics on the quality of matching sequences from NCBI. As a comparison for the N50 sizes: The initially reported N50 values of T. rubripes, D. melanogaster, M.-musculus and H. sapiens were 40kb, 14.5 MB, 16.9 MB and 4 MB (Jaffe2003)

Due to the problems of JGI2, the two model organism databases Aniseed and Ghost still work mainly with the first version of the genome. After the completion of a BAC mapping project (Shoguchi2006), the authors could manually link scaffolds to longer chromosome sequences, leading to an intermediate assembly, still accessible in the Ghost database as “Ciona chromosome browser”. This paved the way for another improvement, the inclusion of cDNA data to join more scaffolds. In the newest genome assembly, which the authors call “KH” (Kyoto Hoya), scaffolds are partially ordered into longer sequences if they are overlapped by the same cDNA or BAC end sequence and then mapped onto chromosomes with FISH assays. The resulting “KH” version, the original assembly from 2001 with additional evidence, seems to be the best currently available ascidian genome.

For *C. savignyi,* the problem of heterozygosity was known and in addition the genome was sequenced at 12.7 X, with DNA from one single animal. The assembler *Arachne* was modified to take heterozygosity into account: at an increased overlap stringency, it first produced almost two genomes, one for each haplotype, which were then merged to include each locus just once (Vinson2005). The heterozygosity rate was reported to be 4.6%, one out of every 20 base pairs, even higher than in the amphioxus genome (where the initial assembly had to be cleaned of duplicates with a similar strategy) and roughly 50 times higher than the human genome (Small2007). For the second version of the *C. savignyi* genome (Hill2008), the merging procedure was improved in a way such that, among others, one haplotype genome could correct assembly errors of the other. A genetic linkage map permitted to resolve assembly errors and join scaffolds, further improving this genome sequence.

1.2.2 Genomic polymorphism

The uniquely high heterozygosity rate of ascidians, while a disadvantage for bench work, can be also seen as a advantage for the study of polymorphism in general. Re-sequencing of the same species can uncover subtle sequence features with a high turnover, like individual transcription factor binding sites, that usually pass under the radar of traditional sequence alignments:. Preliminary data from the Sidow lab from one locus presented at the Tunicate Meeting 2007, Villefranche, suggested that 1000 sequenced individuals would be needed to obtain the exact level of constraint on each base pair. With these, an assembly sequence would become a distribution, where for every position in the genome the probabilities to observe one of the four nucleotides could be calculated. This goal has become more realistic thanks to two new developments: first, the efficient mapping of *C. savignyi* next-generation sequencing reads onto the genome (Rumble2009), and second new sequencing machines that are generating at least 6GB data per run (Ondov2008) and have shown to produce up to 50 GB recently (ABI Solid), although many reads are unusable (Harismendy2009). Therefore, sequencing 1000 individual specimen of a model organism with a small genome and high polymorphism rate, like *Ciona*, *Drosophila (DGRPWebsite), C. elegans* and *Arabidopsis thaliana* (http://1001genomes.org) will be possi-

ble very soon with a handful of sequencing runs and will result in a completely different view and analysis of genomes, perhaps more than resequencing of the human genome (http://1000genomes.org), with its very low polymorphism. During the next years, bioinformatics groups will have to find ways to handle these data in databases and to visualize them on genome browsers.

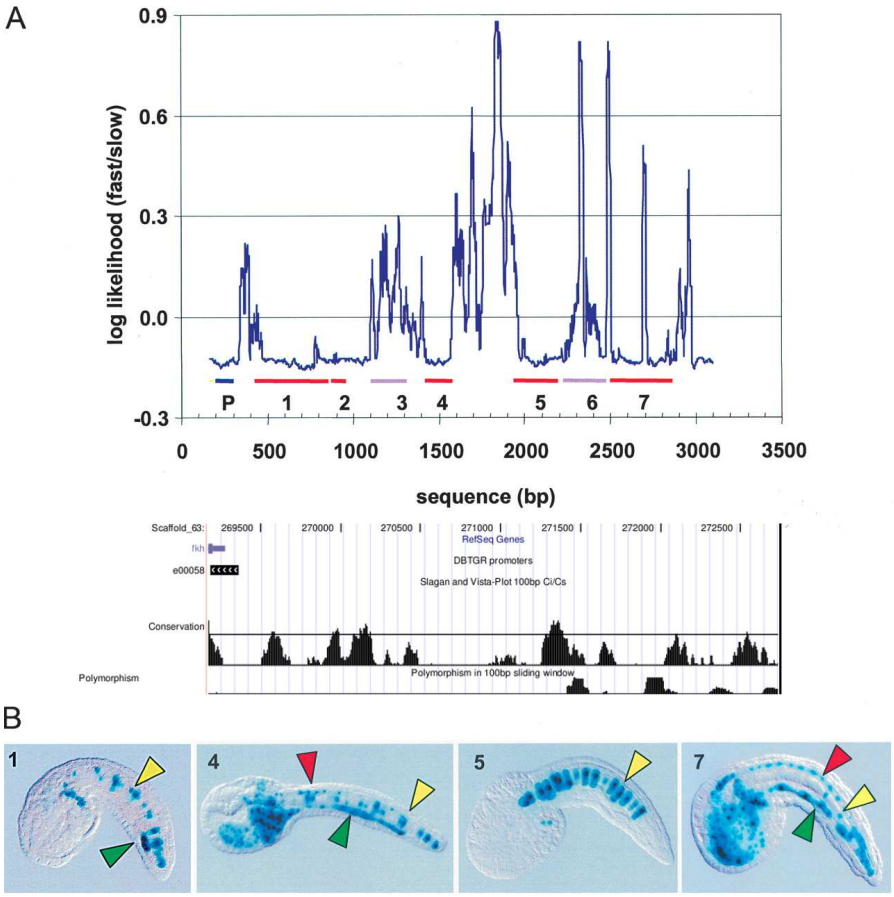


Figure 8: Based on (Boffelli2004): *A: Polymorphism in 33 C. intestinalis specimen from various countries 5' of the forkhead gene, lower part: the Ci/Cs VISTA conservation plot from our UCSC browser installation. Both conservation measures do not completely agree, e. g. region 4 would have been missed based on inter-species conservation alone. As can be seen from the UCSC plot, the information derived from polymorphism in the Japanese genome traces is not sufficient to distinguish between the different regions. B: Expression patterns of subfragments 1,4,5 and 7 from A, cloned upstream of a reporter gene*

However, *C. savignyi* is a relatively uncommon model, so rendering this strategy applicable to *C. intestinalis* (that appears 10 times more often in Pubmed abstracts) might be of more interest to the ascidian research community. (Boffelli2004) have shown that while the haplotype heterozygosity

rate might be lower in *C. intestinalis,* the polymorphism rate between specimens from different populations is very high: by re-sequencing ~40 animals from around the globe, they could gather enough information to distinguish functional from non-functional non-coding regions with a simple alignment conservation plot (See Figure 8). This suggested that the populations of *C. intestinalis* might be more divergent that previously acknowledged. At the extreme ends of this spectrum, two main sub-populations have been found, with an estimated divergence time of 20 MYA, based on genomic (Suzuki2005) and mitochondrial sequences (Nydam2007) (Iannelli2007). They can be hybridized (Suzuki2005), bis with infertile progeny (Caputi2007).

There is one huge sequence resource that has not been used until now for polymorphism analyses: the Japanese genome sequencing reads (which belong to the same sub-species as the American ones at Half Moon Bay). When Takeshi Kawashima designed a custom micro array for *C. intestinalis,* used e.g. by (Azumi2007) and (Christiaen2008), he had to align them onto the genome to make the probes compatible with cDNA from Japanese animals. He kindly sent me his files and I converted them into UCSC format. We observed a polymorphism rate of 5.5% in these alignments. This is higher than in one *C. savignyi* specimen, the data is readily available and could be directly used for genomic analyses, e.g. biased codon substitutions as observed in *C.savignyi* (Donmez2009) or regional polymorphism differences. However, it is no surprise that 5.5% of variability is not enough to distinguish functional from non-functional regions (Figure 8). But it seems likely that - if the right populations from *C. intestinalis* are selected - less than 1000 sequenced genomes could be sufficient to capture a conservation profile on a single base pair level.

1.2.3 Genome annotation

Genomes are commonly annotated with a mixture of composition-based predictions and cDNA sequencing data. On the non-coding side, a computational pipeline predicted thousands of conserved sequences that bear some resemblance with structured mRNAs (Missal2005). Results from a similar algorithm were mostly confirmed to be transcribed (Norden-Krichmar2007). High-throughput sequencing of miRNAs lead to the observation that they were often flanked by other miRNAs which (Shi2009) termed “miRNA-offset”, moRNAs, a type of sequence only observable currently in *C. intestinalis*. Different types of retrotransposons have been found in the genome (Permanyer2003). Among them, the P transposon (Kimbacher2009), with the bizarre observation that the insertion sequences are more similar between *C. intestinalis* and *D.melanogaster* than between the two sequenced ascidians. On the whole, non-coding gene annotation in ascidians is currently rather limited.

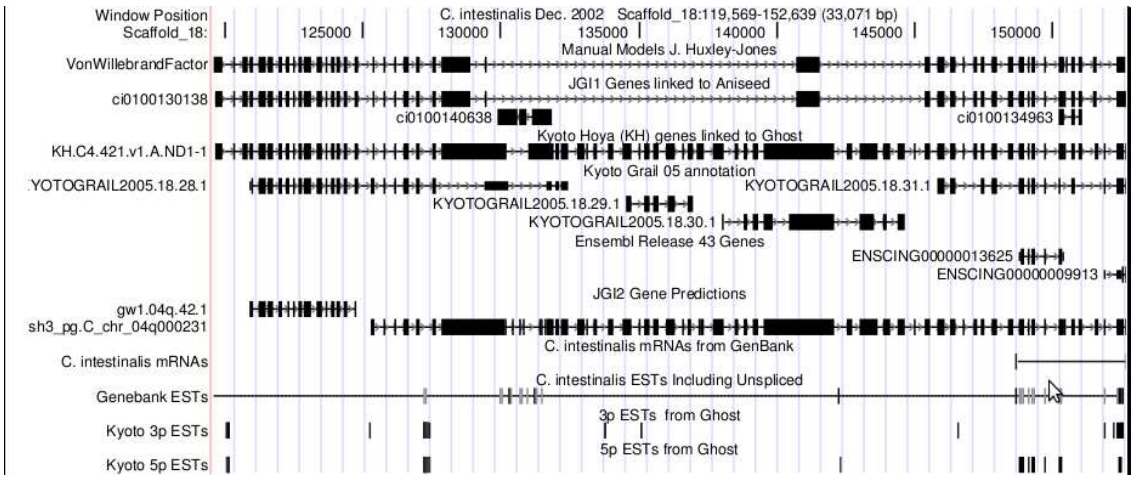


Figure 9: The locus of the Von-Willebrand-Factor, with all annotated gene models. The first model has been manually curated by Julie-Huxley Jones (Huxley-Jones2007) based on mammalian sequences, all other sequences were mapped with BLAT. It can be seen that the intron-exon structure is very different between the different gene models and on this locus, KyotoGrail and JGI2 models are completely off the mark. Also note the big difference between KH and the manual curated model. According to the author of the curated model, given a vertebrate/ciona alignment of the different gene models, the one from JGI1 is the most likely (Julie Huxley-Jones, personal communication)

However, *C. intestinalis* is the invertebrate with by far the most EST data (e.g. four times more than *C. elegans*) and with more ESTs per gene

than chicken (NCBI DBEST May 2009). Based on these 1.2 million end-sequenced cDNAs an automatic, high-quality gene prediction should be possible. In general, gene model pipelines resemble the whole-genome assembly process: they join overlapping ESTs into longer gene models, with some exons predicted from just the sequence composition of the genome. Supporting evidence based on protein matches is added from databases like Swiss-Prot. For *C. intestinalis,* each JGI assembly was accompanied by a new gene set (JGI1-Genes and JGI2-Genes). A Japanese group established their own, separate gene set (KyotoGrail) every year based on the software GrailExp (Satou2005). As ascidians are among the few invertebrates included in the Ensembl genome browser, the Ensembl pipeline also predicts a new gene set for both Cionas every six months. Finally, in an effort to improve these gene models, (Satou2008), manually selected the best model for all 15254 loci from all evidence available in 2008. This resulted in 3330 completely new genes and fused 1779 JGI1 genes into longer ones. Figure 9 shows how all these gene models can differ, on a example of the von Willebrand-locus where published curated data by a gene-specific expert is available (Huxley-Jones2007).

1.2.4 Whole-genome alignments

Conserved non-coding regions are the commonly used predictor of cis-regulatory function (Johnson2004). With the two Ciona sequences available, a whole-genome alignment of them is needed to uncover these region. Algorithms consist usually of three steps: first, a list of alignable genomic fragments are established (anchors). As some are not uniquely assignable but match several other genomic locations, their neighboring matches are compared to select a set of alignable anchors that are consecutive in both genomes. The longest set of them, after an optional refinement, is then output. SLAGAN (Brudno2003) and the UCSC toolchain (Kent2003) from the Mouse Genome Project are the two main algorithms that implement such a procedure. SLAGAN uses CHAOS to obtain a list of anchors, searches for chainable segments of these, extends the borders and then applies the LAGAN global aligner on them. The UCSC toolchain, however, runs BLASTZ on both genomes and then filters and merges the output in three stages (chain-

ing, netting, maffing) to find the best syntenic blocks of locally alignable sequence which is much faster.

Although a global alignment is considered more sensitive in general, the difference between SLAGAN and BLASTZ amounts to 1-2% when benchmarked on one human/mouse chromsome and when run on the whole genome, BLASTZ is even slightly more sensitive (Brudno2003). The main advantage of SLAGAN for ascidian biologists is the VISTA website, which uses a sliding window of usually 100bp to generate colorful and publication-quality graphs. The pairwise SLAGAN algorithm itself is rarely used anymore, since for most animals more than one close genome is available and a multiple alignment is needed. As a result, the main vertebrate genome browser conservation tracks (UCSC, Ensembl, dcode.org) are based on BLASTZ which is also faster. They added many useful tools to post-process the alignment files (filter, split, overlap, etc) and allow the inspection of these on a base-pair level, with additional annotations superposed onto them. Unfortunately, they have not aligned the two ascidian genomes.

The VISTA and JGI websites display alignments, but are limited to the genomic annotations already present in their databases. As a result, in 2006, one had to juggle between four different genome browsers (JGI1, JGI2, Ensembl, Ghost browser, Vista browser) to find out if a given conserved region is really non-coding. When switching between them, JGI1 gene numbers are still the only ones universally accepted by all sites. This poses a problem when one wants to screen conserved sequences, as it is time-consuming and error-prone to manually find the flanking genes for a high number of alignment blocks and to validate whether they are really non-coding. This could be done by programs but then all the different databases would have to be converted into a common format.

1.2.5 A comprehensive genome annotation database for ascidians

For these reasons, I decided to create a local copy of the UCSC genome browser at the CNRS with both the 2001 and the 2005 version of the *C. intestinalis* genome (the KH assembly was not published yet) and do the annotation myself. I started with basic data copied from UCSC (EST alignments, Refseq sequences, JGI1), then converted and added all available gene mod-

els (Ghost, TIGR, Ensembl, JGI2, KH), some cis-regulatory regions from the literature, 254 cis-regulatory regions from the Aniseed database, 84 from DBTGR and direct links to insitu expression patterns. For the whole-genome alignments, I converted SLAGAN Vista alignments to display them in a VISTA-like format. Based on the source code and documentation of the UCSC BLASTZ whole-genome alignment pipeline, which is tailored to the "kilocluster" at Santa Cruz, I was able to run the programs on the Vital-it cluster of the SIB at Lausanne.[1] As a result, it is the only database which contains BLASTZ and Vista alignments, all known gene models and ESTs and can display them next to each other. As a matter of fact, my local web server is known to and used by very few researchers[2] but constitutes a convenient resource to select and analyze conserved non-coding regions. Error: Reference source not found illustrates this on the example of the HOX3 regulatory region. The underlying database of my UCSC browser presents an ideal starting point for whole-genome analysis of non-coding sequences. One is left with the problem of how to establish the link with gene expression databases and how to screen the predicted sequences afterwards which will be treated in the next section.

1 My documentation http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto is currently the only detailed description of the UCSC pipeline on the web. It has been repeatedly used by the UCSC genome browser staff to illustrate the individual steps, from BLASTZ files to the multi alignment block, on their public mailing list

2 Web server reports show that the site has been accessed ~ 900 times during the last year, of which half originated from my own colleagues at the DEPSN and the other half from Berkeley University (Lionel Christiaen), around 40 connections originated from other US states

1.3 When needles look like hay: How to set up a tissue-specific enhancer screen

Maximilian Häussler, Jean-Stéphane Joly

Abstract:

One important tool to investigate tissue-specific processes and genes are cis-regulatory elements. As they do not bear a distinctive sequence signature researchers have to identify them by elaborate *in-vivo* screens . Here, we give an overview where and how these elements have been characterized in the literature. We discuss enhancer distances, promoter specificity and inter-species conservation, and derive general guidelines for a cis-regulatory screen. Experimental improvements from different model organisms are added. We also resume results from computational predictions based on short binding site motifs which can be a useful filter, given adequate controls.

Apart from this advice, we summarize the most convincing explanations for several puzzling questions raised by cis-regulatory analyses. Non-conservation of elements predicted by ChIP, for instance, can be partially explained with the low tissue-specificity of these assays. The existence of long conserved sequences despite much shorter binding sites could be due to the overlap of adjacent sites. The observed redundancy of elements and the biased distribution of non-coding conservation might not be surprising as presented, since they rather resemble phenomenons known from protein-coding sequences.

The current mental model of cis-regulatory elements is derived from a vast body of literature ranging from human genetics and transgenic animals to high-throughput cell culture assays and computational sequence analyses. Our comprehensive overview of this changing field should represent a helpful guide when preparing a cis-regulatory screen.

Introduction

Activating tissue-specific cis-regulatory elements - called "enhancers" (Banerji1981) - trigger gene expression in a given cell type, at the right developmental time and in the necessary quantity. They are tools of fundamental importance in diverse domains of biology. Cloned upstream of a fluorescent reporter gene, they allow to track cell fate during embryogenesis with laser-scanning microscopes and to automatically sort dissociated cells. They permit the analysis of essential genes by limiting the effect of functional assays to targeted cell populations: Ectopic or over-expression of a gene, knock-down with RNAi or dominant-negative proteins or activation of Cre/Lox constructs can be performed in a tissue-specific manner. Finally, sequences of cis-regulatory elements can give clues on the trans-activating factor, helping to identify tissue-specific selector genes (Hobert2008). However, relatively few of these elements, especially for embryonic structures, have been described until now and even fewer of them can be found in reviews or databases of cis-regulatory elements. Researchers are therefore often obliged to dissect the cis-regulatory landscape of a gene themselves.

Many reviews of the different types of cis-regulatory elements and trans-acting factors are already available, e.g. (Arnosti2003) (Maston2006). When one is searching for an element with a specific expression pattern of interest, different strategies can be adopted. In the following, we provide guidelines for an enhancer screen. We summarize how various improvements can be integrated in order to simplify the *in-vivo* testing of tissue-specific cis-regulatory elements with transgenic model organisms. We argue that, given the predictive power of conserved non-coding sequences and results from whole-genome transcription binding assays, most conserved enhancers are expected to consist of dozens of overlapping binding sites. Some algorithms are available that predict the expression pattern from these sequences. We point out their common characteristics and possible limitations in the context of an enhancer screen and highlight some topics deserving further investigation, notably silencers and the curation of validated elements.

Experimental screening to find cis-regulatory elements

The importance of the proximal promoter region

Where are enhancer elements found? As a matter of fact, they are usually sought in non-coding genomic regions. A handful have been mapped to 5' UTRs (e.g. in the first exon of Pax6, (Zheng2001), IGF-1 (McLellan2006) and TH (Arányi2005)). We know of only three examples located in translated exons (Hoxa2 (Tümpel2008)(Lampe2008) and Adamts5 (Barthel2008)) but there are probably more to discover, as suggested by excessive conservation of synonymous base pairs (Woltering2009) (Chen2007). The most natural place to look for cis-regulatory control is the region just upstream of the gene's transcription start site. Genomic fragments <10kb can be simply cloned into plasmids containing a reporter gene. In the standard *in-vivo* assay, the resulting DNA is then injected (mouse, fish, sea urchin, flies, nematodes) or electroporated (ascidians, chicken) into fertilized eggs or embryos. For invertebrates with small genomes like *C.elegans, Drosophila and C. intestinalis*, this very often reproduces the gene expression pattern faithfully (Boulin2006) as a large part of the complete upstream region fits into one plasmid. The approach is sometimes also successful in vertebrates (e.g. (Wang2002) (Park2000) (Yoshikawa2007)). However, with long-range regulatory elements missing from the construct, many proximal regions recapitulate only a part or none of the wild-type expression pattern of a gene. But they are easy to clone and already contain the basal promoter which otherwise has to be added to the reporter gene.

It is often assumed that basal promoters are rather ubiquitous, merely directing the polymerase to the start of transcription (Frith2008). This general role might explain why genomic sequence analyses are in roughly one third of the cases successful in establishing a link between the direct upstream sequence and the cell type where a gene is expressed (Roider2009) (Smith2007). It could also be the reason why the direct upstream region is less conserved than distal CNEs (Tsuritani2007)(Blanchette2006), as the sequence itself is less important. However, basal promoters can exhibit more tissue-specific activities than anticipated. In invertebrates not every promoter plays well with every enhancer: In Drosophila, enhancers of gsb,

gsbn, ant, bx require a certain type of promoter (DPE or TATA) (Li1994) (Ohtsuki1998) (Butler2001) , a mutation of the yellow or oaf promoter can change the contacted upstream enhancer (Lee2006) (Merli1996), the Hsp70 promoter might direct weak salivary gland expression in flies (Markstein2008) and a neural motif is not active when combined with some non-neural promoters in *C. elegans* (Wenick2004). In an extreme case, a tissue-specific element in sea urchin showed two different expression patterns, depending on the basal promoter used (Kobayashi2007).

This can lead to problems in medium-scale enhancer screens that test CNEs genome-wide, flanking many different genes. In these experiments, standardized promoters have to be used, typically pHsp or pBeta-globin. These might introduce a bias as compared to detailed single-locus analyses using the endogenous promoter. In Drosophila, this problem motivated the development of the artificial *Super Core Promoter*, a mix of several different sequences with the goal of high enhancer compatibility and high expression levels (Juven-Gershon2006). In vertebrates, one enhancer element tested with six different basal promoters in zebrafish did not change the expression pattern, according to the authors, though quantitative differences are visible (ß-globin, Ngn1, Hsp70, Hs-Sox3, Atp11c (human and zebrafish), Dr-Gata2 (Navratilova2009)). In general, various studies have found a similar ratio of enhancers although they used very different promoters (see Table 1). Though we are not aware of a clear proof for promoter incompatibility in vertebrates, an endogenous sequence should be preferred. This is another reason for testing the direct upstream region of a gene first.

Long range control and position effects

Sometimes proximal elements do not drive expression in the right cell type, a BAC is not specific enough, or short enhancer elements are needed. This motivated the use of larger vectors, cosmids and BACs (Long2007) which are more difficult to handle than plasmids. Thanks to optimized protocols and better selectable markers, they can now be efficiently modified within one week (Sharan2009) (Tursun2009) (Smith2008) (Venken2009) (Ejsmont2009). Protocols and reagents are available free of charge from the National Cancer Institute at Frederick (NCICRF). Instead of screening individual DNA fragments to find the cis-regulatory element of interest, a BAC-

clone with the gene replaced by a fluorescent protein should usually be sufficient to mark a cell type for subsequent analyses (Bouchard2005). Mouse lines for 800 BACs with a GFP knock-in can be ordered through the GENSAT consortium (Geschwind2004).

If the exact location of the enhancer is required, the radius of a regulatory element search can be extensive: Analysis of chromosomal rearrangements in human disease and large vector tests showed that enhancers can be located up to 1 MB away from their target gene in vertebrates (Lettice2003)(reviewed by (Long2007)). Taking into account the smaller genome sizes, long distances in invertebrates have also been reported (Jack1991) (Dorsett1993) (Conradt1999) (Smith2008). These long-range interactions were unexpected: Since the early 1980s, the common model of chromatin-loopings that permit cis-regulatory contacts is based on observations from experimental data on the β-globin locus (reviewed by (West2005)), supported by chromatin conformation capture assays (Dekker2002) and chromatin bound by tagged RNAs (Carter2002). It was found that β-globin enhancers contact each other (Patrinos2004) and basal promoters over long distances. The necessary DNA looping is induced and anchored by transcription factors like GATA1 (Vakoc2005) and accompanied by chromatin modifications of the globin enhancers (Li2006). Contacts like these can even reach out to other chromosomes (Chen2002) (Simonis2006) (Lomvardas2006) (Ronshaugen2004).

Apart from BAC-based experiments, one simple way to reduce the search space for cis-regulatory elements could be genome synteny comparisons. Several authors have independently argued that long-range regulation limits possible chromosomal rearrangements and maintain some exceptionally long and well-conserved syntenic blocks. (Kikuta2007) (Santagati2003) (Goode2005) (Lee2006#104) (Engström2007) (Ahituv2005) (Wang2007#349) (Hufton2009). Following this model, synteny breakage could be used to delimit the boundaries of enhancer action: If a given region is not flanking the ortholog in a related species, the enhancer is less likely to be located there. The genome browsers of UCSC and Ensembl provide a DNA-based synteny view for this (“UCSC Net Tracks” and “Ensembl Multicontigview”). Metazome (www.metazome.net) tracks only genes, which makes it easier to use but less sensitive and the tool Synorth (Dong2009)

combines genome and gene tree view. Figure 2 shows an example of the gene SALL1 based on the UCSC Browser where synteny with *X. tropicalis* suggests that most enhancers concentrate in a 1.5MB segment around the gene.

Some enhancers have been shown to regulate several genes. They are often called “global control regions” or “locus control regions”. Well-known loci include, apart from the alpha- and beta-globins, the interleukins, and the EVX2-HOXD locus (reviewed in (Spitz2008)). Regions of the genome under the influence of global control regions have been called "gene expression neighborhood" (Oliver2002) or "regulatory landscape" (Spitz2008). They might also explain non-random placement of co-expressed genes along the chromosomes, as observed in *D. melanogaster*, *C. elegans*, zebrafish and many other organisms (e.g. (Ng2009), reviewed by (Hurst2004)). Therefore, the experimenter has to be prepared to screen up to 1 MB of flanking sequence around the gene of interest, even beyond neighboring genes or within their introns.

Long-range control can lead to problems cis-regulatory tests, when sequences and reporter genes are randomly inserted into the genome. The "position effect" (Spradling1983) describes expression pattern variations between transgenic animals due to the influence of the genomic context around the construct. In Drosophila, the effect between different genomic insertion sites can be 100-fold and RNAi constructs lead to very different wing phenotypes depending on the insertion site (Markstein2008). A common counter measure is to report only the pattern common between several transgenic embryos. An often-proposed alternative is the addition of flanking insulators around the reporter construct (Potts2000) (Markstein2008). In mice, the knock-in of constructs into the transcriptionally "neutral" locus HPRT, now aided by a set of readily available plasmids (Yang2009), should completely eliminate position effects. This is useful when one strives to quantify the effects of small changes in known cis-regulatory sequences (Ahituv2007b) but is too laborious in the context of a screen.

Insulators and repressors

Not all elements are responsible for gene activation. Some of them separate genes expressed in different tissues and are thought to place limits

around enhancers. Insulators in the drosophila bithorax complex and the yellow locus have been analyzed in detail for many years (reviewed by (Akbari2006) (Maeda2007)). They are bound by CTCF in vertebrates, the only currently known vertebrate insulator protein and are thought to place limits around enhancers (Kim2007). Sometimes, currently only shown in flies and sea urchin, tissue-specific tethering- or "anti-insulator" elements can bypass insulators in some tissues (Akbari2008), reviewed by (Dorsett1999)) (Zhou1999) (Lin2004) (Calhoun2002)(Irvine2008).

Some enhancers exert no influence on the expression of neighboring elements (Visel2009), but enhancer interactions have been found: Some elements have a repressor (Conte2007) or amplificator effect on their environment (Yuh1998) (Irvine2008) or both at the same time (Kulkarni2003). In one case, the endogenous expression pattern of the gene Shh could only be recreated with a certain combination of elements, not any individual one (Ertzer2007). Therefore, inactive elements should be preferably tested in combination with others before concluding that they are non-functional.

The high price of *in-vivo* testing

Organism	Delivery	Avg. time from experiment to observation	Price transgenesis, academic rate	Source
D. melanogaster	injection	1 day	$250.00	thebestGene.com
C. elegans	injection	1-2 days	No core	
C. intestinalis	eletroporation	18-24 hours	No core	
Zebrafish	injection	1-2 days	350 EUR	Amagen Core
Chicken	electroporation (not all cells)	1 day	No core	
Mouse	injection	7-13 days	2200$	OSU mouse core

Table 1: Animal models, DNA delivery techniques and the respective cost of testing a single cis-regulatory fragment in transgenic animals based on commercial or academic service core facilities

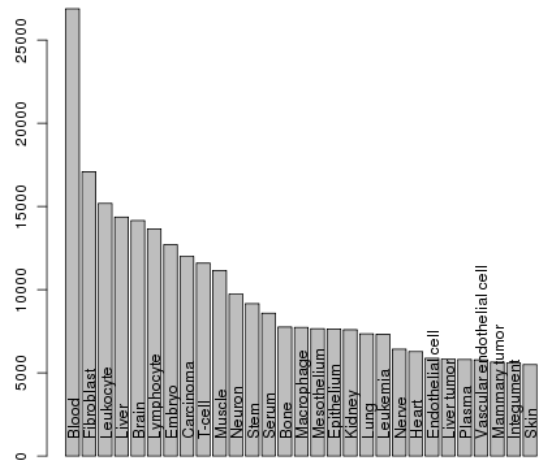


Figure 1: Number of abstracts found when querying Pubmed for a list of synonyms for "cis-regulatory element" and one of the 1208 tissues annotated by SwissProt (only the best 30 tissues are shown). As can be seen, most cis-regulatory information is available from tissues with cell lines. Note that muscle is the main model tissue for computational predictions (see main text) but not the one with the most cis-regulatory information.

To validate active individual regulatory elements within long regions, many small sub-fragments have to be cloned into plasmids one by one and tested for their activity. As a result, elements of tissues with available cell

cultures are the ones best described in the literature (Figure 1). *In vivo* however, current experimental techniques do not allow to screen large regions efficiently for their cis-regulatory potential at kilo base pair resolution. Full testing of all randomly sheared fragments within a genomic region is only feasible in simple model organisms such as ascidians or sea urchins (Keys2005) (Cameron2004). Nevertheless, protocols for other animals have been streamlined during the last years: Observation of F_0 embryos in mice is often sufficient (Loots2008) and in zebrafish and *C. elegans*, cloning can be avoided altogether by injecting PCR products (Woolfe2005) (Hobert2002), though with an increase in mosaicism. In zebrafish, the number of assays can be reduced by using genomic DNA from *Takifugu rubripes*, which is four times more compact while assumed to harbor similar regulatory elements (Barton2001). These experiments are still expensive in vertebrates, ranging between several hundred dollars per tested element in flies and fish to thousands in mice (Table 1). Given the comparable expression patterns of mouse/fish-conserved sequences when tested in fish (Aparicio1995) (Navratilova2009) (Suster2009) (Kimura-Yoshida2004) and the lower cost of these animals, a time-saving strategy might include an initial screen in fish followed by transgenic mice with selected positive elements.

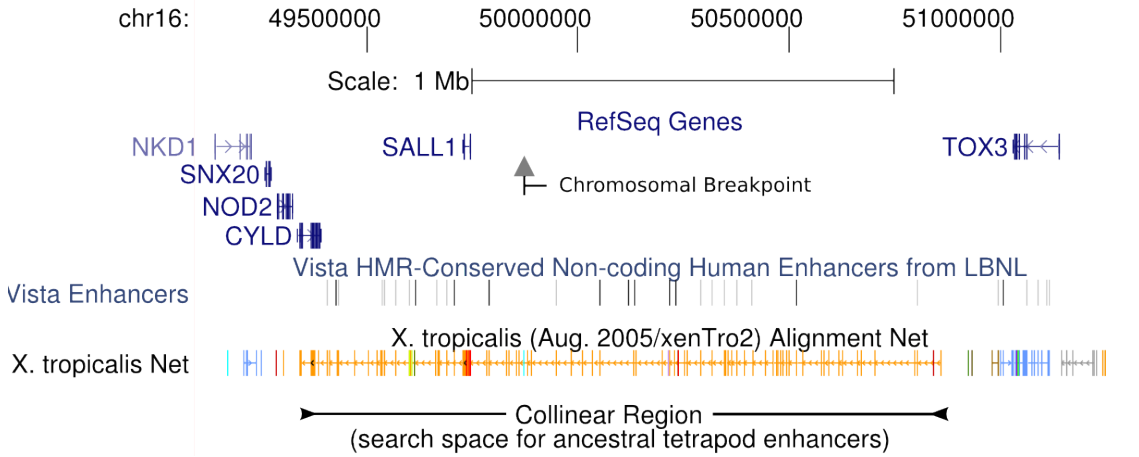


Figure 2: Screenshot of the UCSC Genome Browser showing the SALL1-locus (chr16:49086985-51296445, Mar 2006) with all neighoring genes (Track" RefSeq Genes"). A grey triangle marks the chromosomal breakpoint of a non-coding translocation causing the same effect as a mutation in the gene SALL1 (Ahitu-v2005) (Marlin1999). Several conserved non-coding sequences were already tested in mice (Track "Vista Enhancers", sequences that direct LacZ expression at E11.5 are colored in black).

Same-colored stretches in the net-track do not represent genes but connect consecutive alignable non-coding sequences. The orange collinear region with X. tropicalis is limited to a 1.5MB fragment around the gene SALL1. If one is following the model of "genome regulatory blocks" then an enhancer screen should concentrate on this 1.5 MB region, including the introns of the CYLD-gene.

Non-coding conservation and its implications

The biggest help in finding short tissue-specific enhancers in megabase-sized regions are genomic alignments with non-coding sequences from other species. Since the first analyses of human/mouse alignments in the β-globin locus (Hardison1993) and later the mouse genome project (Hardison1997) (Waterston2002) surprisingly many of these alignable sequences have been found. They are not simply mutational cold-spots but have been shown to be under selection (Drake2006) (Casillas2007) (Sakuraba2008).

Table 2 shows a selection of studies from the literature that tested non-coding conserved elements. It can be seen that most (80%) of the inter-species conserved elements showed a cis-regulatory effect and that the most common criteria is human/mouse conservation. The expression pattern of these elements varies a lot; the bigger screens describe them at a lower resolution. Only few binding sites within these sequences have been further characterized and the most common promoters were Hsp68 and β-globin.

Conservation depth of CNEs

Today, standard genome browsers allow the identification of the different types of conserved noncoding elements (CNE) with a mouse click. Depending on the filtering applied, these regions bear different names: *conserved noncoding sequences* (CNS, >X% identity over Y bps) (Dubchak2000), *deeply conserved elements* (human/fish) (Attanasio2008), *ultraconserved* (200bp identical human/mouse/rat) (Bejerano2004), *extremely conserved* (Pennacchio2006) or *extremely highly conserved sequences* (de_la_Calle-Mustienes2005), *hyperconserved sequences* (more than 5 nucleotides in five species (Guo2008)) and many more (reviewed by (Woolfe2008)). Researchers have been concentrating on these during the last years when searching for tissue-specific enhancers (**Table 2**) and this approach has been very successful. Please note that while one single medium-scale program at the LBL has uncovered more enhancers than all other laboratories together, it is currently lacking a detailed annotation of the tissues stained by the reporter gene.

Locus	Organism (DNA/org)	Sequence conserved with	Tissue or cell type	Tested Enh-ancers	Confirmed Enhancers	Position relative to Gene	Trans-acting Factor determined?	Promoter	Publication
Sall1	chicken	human	anterior neural ridge	5	1	intron	No	thymidine kinase	(Izumi2007)
Sox2	chicken	human	Di/mesencepalon, Nasal and otic placodes, Rhombencephalon, Neural induction, Head ectoderm Mesencephalon Spinal cord, Late lens, Dorsal root ganglia	25	10	50kb 5'	No	Herpes virus thymidine kinase	(Uchikawa2003)
Eya1	chicken		Hensen's node, neural tube, migrating neural crest cells, otic vesicle, olfactory placode, cranial ganglia, trigeminal ganglia	29	10		many (match)	Herpes virus thymidine kinase	(Ishihara2008)
Dach1	mouse	fugu	fore/mid/hindbrain, retina, limb buds, neural tube, genital eminence	9	7	<870kb	No	Hsp68	(Nobrega2003)
Dlx1/2	mouse	zebrafish	anterior entopeduncular area, subventrivular zone, parvalbumin-, calretinin-, neuropeptide Y, and other interneurons	4	4	<12kb	No	Hsp68	(Ghanem2007), (Ghanem2003)
Flt4, PD FFrβ, Ece1, Nrp1, Foxp1	mouse	human?	endothelium	10	5	?	FoxC2, Ets (ectopic expr/KO)	β-globin	(De_Val2008)
Gata2	mouse	human	rostral urogenital system, caudal urogenit. system	4	2	3', 1MB	No	Gata2	(Khandekar2004)
Hoxb4	mouse, fugu/mouse		rhombomer 7/8, anterior mesoderm, neural tube	3	3	intronic	No	Hsp68,Hoxb4	(Aparicio1995)
Hob2	mouse, chicken	bat, chicken	rhombomere 4	1	1	introns	HoxB1, Prx, Prep1 (emsa, mut, overexpr)	β-globin	(Maconochie1997)
Mbp	mouse		oligodendrocytes at different stages	4	4	15kb 5'	Nkx (mut)	Hsp68	(Farhadi2003)
nicotinic acetylcholine receptors	mouse	human	adrenal gland, superior cervical ganglion, pineal gland, SCG neurons,	1	1	30kb 5'	No	None (BAC deletion)	(Xu2006)
Nkx2-5	mouse	human	heart common atria, common ventricle, aortic sac, distal stomach region, tongue,	3	3	27kb 5'	Gata/Smad (mut)	Hsp68	(Chi2005)
Otx2	fugu/mouse, fugu/zebrafish	mouse	roof of dienc., medio-caudal telenc., ventral dienc., ZLI, cephalic mesenchyme, trigeminal ganglions, cranial nerves, dorsal dienc., rhombenc.,nasal pits, first branchial groove	7	7	60kb	No	Otx2	(Kimura-Yoshida2004)
Pax6	mouse	human	late eye development, diencephalon (auto), heart, rhombencephalon	4	3	intronic	Pax6 (emsa)	Hsp68	(Kleinjan2004)
Pax6	human/zebrafish (same)	human	left and right habenulae, roofplate, pineal, medial habenulae	8	6	~ 300kb 5' and 3'	No	Gata2, Hsp70, Ngn1, Atpc11, Atpc11, Sox3	(Navratilova2009)
Shh	zebrafish/ mouse	mouse	embryonic shield, hypothalamus, zli	3	3	introns	No	Gata2	(Ertzer2007)
Sox10	mouse	chicken	otic vesicle, oligodendrocytes neural crest, peripheral nervous system, adrenal gland, sympathetic ganglia, neural crest	7	5	65kb	Sites for Sox/lef/Pax/Ap2 (EMSA)	Hsp70	(Werner2007)
Sox21, Pax6, Hlxb9, Shh	zebrafish	human	approx annotation: nervous sys., sens. organs, notochord, muscle, blood, heart, skin	25	23	various	No	β-globin	(Woolfe2005)
Sox3	human/zebrafish	zebrafish	brain, epiphysis, floor plate, inner ear, cerebellum	8	6	300kb 3', 100kb 5'	No	Gata2 + 5 others	(Navratilova2009)
Various	zebrafish	human	Rough classification into 6 tissues, quantitative	16	10	various	No	cMLC2, luciferase	(Shin2005)
Various	mouse	human	Rough classification in fore/mid/hindbrain	1083	497	None		β-globin	(Pennacchio2006)

Table 2: A selection of studies that describe tissue-specific elements identified by non-coding conservation. If chicken sequences are not counted (tissue-dependant eletroporation), out of 117 conserved non-coding sequences, 93 drove a tissue-specific expression pattern (80%). In the biggest screen in mouse embryos which were fixed a E11.5, only 497/1083 CNEs were active (45%).

Non-conserved enhancers

As an extension of the human genome project, the ENCODE pilot study characterized “functional elements in 1% of the human genome” (Birney2007), which included conserved as well as non-conserved regions with high-throughput chromatin immunoprecipitation assays. These assays promised to identify cis-regulatory elements much faster than traditional *in vivo* injections.

Subsequent computational analysis of the resulting fragments considered functional showed that they were not significantly enriched in regions under constraint in cross-species non-coding alignments (King2007) (Zhang2007). This seems to contradict the publications from **Table 2** that concentrated with 80% success onto conserved sequences. Several factors can explain this result: For technical reasons, ENCODE had to be based on immortalized cell lines like HeLa and HL60 which are already differentiated. Second, the transcription factors targeted by antibodies were mostly ubiquitous, like Sp1, Pol4, E2F1/4 and Taf1, and not tissue-specific. When chromatin immunoprecipitation is directed to a cofactor implicated in tissue-specific elements and uses cells dissected from an animal, an enrichment of conserved sequence was indeed found (Visel2009#206). Third, a region might be bound by a factor but this does not necessarily reflect a function which is under selective pressure (Li2008). Chromatin studies rather predict function and their results need to be confirmed by in-vivo tests. Fourth, selective pressure seems to vary a lot depending on the function of the regulated gene (King2007) and the element itself, so a signal biased towards developmental regulators might be invisible on a whole-genome level.

To our knowledge, the three main techniques that are based on nuclear chromatin have mostly been applied on nuclear extracts from cell cultures. The first is DNaseI digestion for the detection of nuclease hypersensitive sites (Gross1988), the second one chromatin immunoprecipitation to find regions bound by antibodies against modified histones (Heintzman2009) or transcription factors. A third assay, chromatin conformation capture, uses proximity ligation to identify and quantify contacts between cis-regulatory sequences like promoters and enhancers. (Dekker2002)(Dostie2006)

But results obtained from cell culture assays do not seem to expose tissue-specific elements (Attanasio2008) (Göttgens2000). Some of the enhancers predicted from cell cultures can become repressors when changing the cell context (Voth2009). Replacing cultured cells with ones manually dissected from animals can remedy this, but this depends on the size of the tissue: (Heintzman2009) had to isolate forebrains from 150 mouse embryos, for instance. The alternative, automatic cell sorting, requires a already available cis-regulatory element to mark the cells with fluorescence, to select only e.g. blood or neurons (Long1997) (Cerda2009) (Jiang2008). Both approaches still depend on big amounts of nuclear extract, on the order of 10^7-10^8 cells, a problem that will become less critical with recent technical improvements of the immuno-precipitation procedure (Dahl2008) and the replacement of microarrays with DNA sequencing (Wederell2008).

With the “traditional” cloning and testing approach, non-conserved fragments have been shown to direct expression: examples from the vertebrate loci PHOX2 (McGaughey2008), REST (Fisher2006) and invertebrates (Hare2008#75) (Hare2008) (Wratten2006) (Romano2003) revealed basic tissue-specific elements that were completely absent from mammalian/fish alignments (reviewed by (Nobrega2004)). Still, despite the ENCODE results and individual examples of the contrary, the current literature rather suggests that while not all functional regulatory elements are alignable among vertebrates, the more an element is conserved, the more likely it is to have some tissue-specific regulatory function. (Cheng2008) (Pennacchio2006). Nevertheless, with current protocols, although they represent the future of cis-regulatory in-vivo analysis, high-throughput assays are not yet applicable to a limited number of cells, as those from small embryonic fields or brain substructures and therefore tissue-specific elements are still painstakingly identified by transgenic in-vivo assays.

Main features of CNEs

Many conserved non-coding elements are present in vertebrate genomes. Depending on their definition (Visel2007), one can find between several hundred (ultraconserved), several thousands (human/fish) to several hundred thousands (mammals).. Their analysis give hints how tissue-specific

elements are distributed in the genome and how their sequences are conserved:

- Many of them are better conserved than most protein coding sequences (Bejerano2004)(Dermitzakis2003). Their relative share compared to conserved coding elements increases with organism complexity from yeast, worms and insects to vertebrates (Siepel2005).
- They have a "short lifetime" and are mostly phylum-specific: Only 56 of the vertebrate sequences can be found in a cephalochordate, the amphioxus (Putnam2008). No single CNE is conserved between vertebrates on the one hand and flies, worms or ascidians on the other (Bejerano2006), most alignable elements are found within vertebrates, flies, ascidians and plants. Most mammalian CNEs seem to have emerged during the early tetrapod history (Stephen2008) and have been strongly retained during mammalian evolution (McLean2008)
- The best-conserved primate regions correspond to the best-conserved mammalian alignable regions (Prabhakar2006) (Wang2007)
- CNEs show a biased A/T distribution with 6% more A/T than in the flanking regions, in vertebrates, worms and plants (Walter2005) (Vavouri2007) (Li2009).
- In five regions conserved in sea urchins, insertions >20bp are almost absent (Cameron2005) and one 16bp-insertion into one of the best conserved enhancers in the genome, in the Dach locus did not change the expression pattern (Poulin2005).
- Some CNEs are alignable between paralogous genes: After a segmental or whole-genome duplication, paralogs can retain a limited number of essential cis-regulatory elements (McEwen2006) (Woolfe2007) (Li2009) (Tsang2009), which are very likely to represent enhancers. But even in fish genomes that have undergone an additional whole-genome duplication, these "duplicated conserved noncoding elements" (dCNEs) are quite rare (~124, in list established by (McEwen2006))

- Compared to genes, the lengths of CNEs (Retelska2007) are relatively well conserved in vertebrates compared to flies. The distances between CNEs (Sun2006) are better conserved than distances between genes or between exons.
- Around genes, vertebrate CNEs are evenly distributed between the 5' and 3' end. The regions farther away from the gene are denser in conserved elements (Blanchette2006), 12% of duplicated CNEs (for these, a target gene is clearly assignable) are located farer than 1MB from their target (Vavouri2006)
- The distribution of CNEs in the genome is strongly biased. In vertebrates, flies and worms, they concentrate around transcription factors (Sandelin2004) and are under-represented around housekeeping genes (Farré2007) (Vavouri2007). The initially reported over-representation of nervous system genes (Bejerano2004) was merely a result of their longer flanking regions (Taher2009).
- CNEs are four times more common in "gene deserts", defined as >640kb without a protein-coding gene, making up 25% of the human genome (Ovcharenko2005)(Siepel2005). The longest of these regions flank some well-known developmental regulators like OTX2, DACH, SALL1 or SOX2 (see also Table 1)
- The share of functional elements increases with the phylogenetic distance of the animals where they are alignable (Pennacchio2006) and also with the density of surrounding elements (Prabhakar2006)

These findings have important implications when selecting candidate enhancer sequences: As conserved regions are unevenly distributed, there is currently no “optimum” combination of genomes to find them but preference should be given to the most conserved regions in a given locus. Experiments on invertebrates are a lot simpler, but current alignment algorithms cannot identify homologs of CNEs in vertebrates. To identify vertebrate CNEs researchers currently use a combination of various species and rather simple cutoffs (see Table 1), although primate sequence comparisons are reported to be sufficient. Although the upstream part of genes is the most common place to look for cis-regulatory elements, CNE-distribution suggests that elements can be located just as well in the 3' region.

Guidelines for enhancer screens:

- Invertebrates are the cheapest organisms to manipulate but their sequences can not be mapped to vertebrates with current alignment algorithms. Assaying fragments in fish instead of mice can accelerate the assays. In some cases, non-coding alignments between paralogs, cloning DNA from close organisms with smaller genomes and injection of raw PCR fragments can simplify the experiments.
- The proximal upstream region should be tested first, it could also give rise to an endogenous promoter
- The endogenous basal promoter should be preferred if possible, otherwise there is little evidence for the necessity of a “Super Promoter” in vertebrates so far.
- One should be cautious when basing the strategy on large-scale chromatin data from cell-cultures although more and more of them are becoming available.
- CNEs that are to be tested can be located up to 1MB away, skipping neighboring genes. The synteny of the locus can be taken into account when delimiting the search space.
- CNEs with a conservation across the highest phylogenetic distance should be tested first and transcribed sequences are not to be excluded. Essential genes like transcription factors and those expressed in the nervous system are flanked by more and better conserved elements, so the "best" phylogenetic distance depends on the gene of interest, it can be human/chicken in one case (Uchikawa2003), fish/human in another (Shin2005)or the best-conserved primate alignments (Prabhakar2006).
- Partial redundancy is expected and negative elements can be further characterized by combining them with others, as they might repress or modulate the activity of others
- The number of proteins binding to a conserved cis-regulatory element should not be underestimated. This can make interpretations of non-coding mutations difficult to interpret,
- Sequence-based predictions heavily rely on the available data about the tissue of interest. They should be taken with a grain of salt if they make general assumptions on the composition of cis-regulatory elements but can be tested on control gene sets (e.g. derived from in-situ screen databases)

Redundancy of regulatory elements

Expression patterns of enhancers in a single locus often seem to overlap (See Table 1). Hong et al (Hong2008) recently coined the term "shadow enhancer" for this phenomenon. They reason that the resulting redundancy protects essential developmental processes against mutations. While redundancy is often observable at early developmental stages, we have not come across two enhancers active at advanced developmental stages with an exactly identical pattern, although they are often overlapping (see Table1, in particular data on Nkx, Six3, Sox2, Otx2, Gdf6 and Shh).

But redundancy might explain that no phenotypic effect was observable by researchers in a laboratory environment when mega bases of non-coding sequence, highly conserved elements or previously characterized enhancers for Engrailed2, Fgf4, Gata1 and Myod were knocked out in mice (Visel2007) (Nóbrega2004) (Ahituv2007) (Li_Song2000) (Iwahori2004) (Guyot2004) (Chen2004). However, directed mutations of tissue-specific elements have shown a clear phenotypic effect in the loci of Shh, Shox, Meis1, Hoxc8, Dhand2 and Bmp2 (Lettice2003) (Sabherwal2007) (Xiong2009) (Juan2003) (Yanagisawa2003) (Dathe2009), even when they involved just single base pairs (Papachatzopoulou2007) (Lettice2008) (Rahimov2008). In the case of TCR-gamma, two elements have to be deleted in combination to produce a visible effect (Xiong2002).

Taken together, the redundancy of regulatory elements resembles the redundancy of genes. It brings to mind a controversy on the exact function of HOX paralogs that started 15 years ago. Several of them were knocked out, some in combination, with the conclusion that redundancy is apparent in some tissues, some genes, and not in others (Horan1995)(Condie1994). Therefore, partial co-expression of essential cis-regulatory sequences is usually expected for many essential processes, just like in genes. For a screen of putative elements, this increases the chance of the experimenter to find activating sequences in the tissue of interest but can render analysis by deletion (knock-out in genome or BACs) difficult to interpret.

The origin of conserved non-coding sequences

CNEs as non-coding RNAs

A puzzling question remains: How can an enhancer be conserved over 200bp without a single base pair mutation between human and mouse (Bejerano2004), if a transcription factor binding site is only 4-8 base pairs long? Why do the nucleotides between the binding sites not mutate? Some of the well-conserved CNEs are derived from transposable elements (Nishihara2006) (Xie2006) (Bejerano2006) but this does not account for the selective pressure on their sequence. One explanation could be a double function of the elements, if they serve as enhancers and a regulatory RNA at the same time. Pervasive transcription of non-coding sequences (Birney2007) supports this, as well as the overlap of non-coding transcripts with chromatin boundaries (Akbari2006) (Rinn2007), a conservation profile resembling structured RNAs (Washietl2007) and various examples where non-coding RNAs regulate directly the transcription process (Amaral2008). Indeed, some CNEs of interleukin and IRX genes are validated enhancers and are also transcribed into RNA at the same time (Jones2005) Su(de_la_Calle-Mustienes2005), two transcribed enhancers even cooperate with transcription factors that bind to them (Feng2006) (Sanchez-Elsner2006). Although transcription seems to play a rule in cis-regulation, we are not aware of a general mechanism of how these RNA molecules are linked to regulatory sequences and why. In any case, experimenters should not eliminate transcribed sequences from an enhancer screen.

CNEs as dense clusters of binding sites

Apart from RNA another explanation for the high conservation of long sequences is the overlap of neighboring binding sites. An elegant example of this has been recently found in the enhancer of interferon-β (Panne2007). The authors combined several crystal structures of transcription factors that bind to the 50bp enhancer and form a complex called "enhanceosome". The 3D model shows a general absence of protein interactions but instead a strong overlap of the different binding sites which do not always correspond to the consensus motif **(Fig 1).** Such dense chains of proteins with contacts

at every single base pair of the DNA could explain the high conservation of enhancers. Estimations based on one example are certainly daring, but if this is similar for all other conserved regions, then all conserved enhancers, i.e. most conserved non-coding sequences, might be bound by tightly overlapping transcription factors. E.g. a sequence that is conserved with mouse over 500 bp, should be bound by around 100 proteins. It is easy to imagine that in pleiotropic enhancers essential for proper animal development and expressed in several tissues, the number of proteins bound could be much higher. Some of them might play rather a minor role and show a quantitative effect when tested. Others will influence the spatial expression pattern of the genes. These are of main interest for developmental studies.

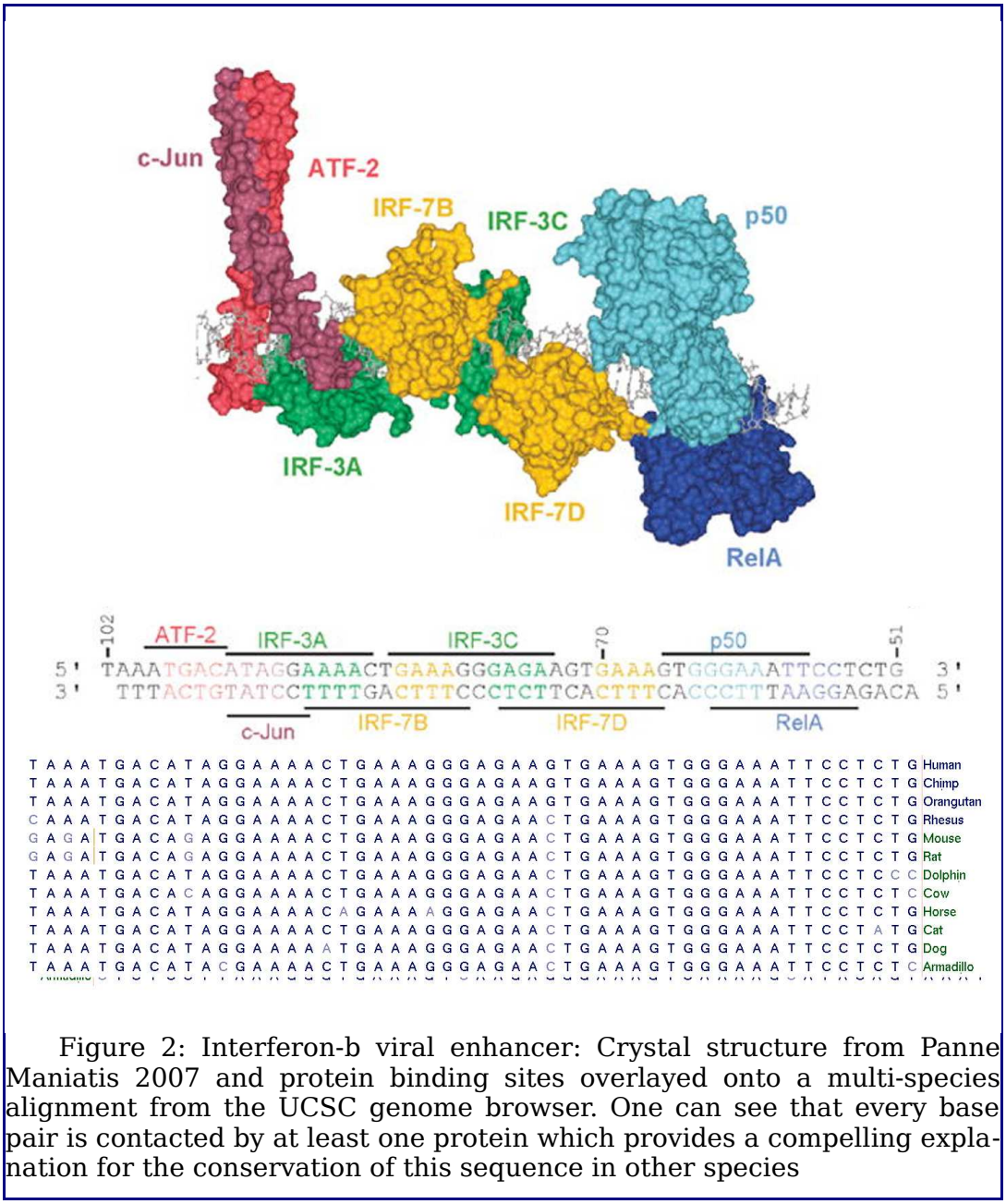


Figure 2: Interferon-b viral enhancer: Crystal structure from Panne Maniatis 2007 and protein binding sites overlayed onto a multi-species alignment from the UCSC genome browser. One can see that every base pair is contacted by at least one protein which provides a compelling explanation for the conservation of this sequence in other species

Predicting tissue-specificity from nucleotide sequences

Proposed distinctive features of binding sites

The guidelines from Box 1 should maximize the number of any type of positive enhancers from a screen but they cannot select elements that are specific for a certain tissue. To tackle this question, one has to find a link between the sequence and the function of conserved elements. The basic idea

to use the nucleotide sequence of cis-regulatory elements to predict their expression tissues is not new (Fondrat1994). Its success depends on the detection of functional binding sites, a complex topic which has been reviewed elsewhere in detail (Wasserman2004) (Vavouri2005) (Elnitski2006). The main difficulty here is that a degenerate motif equivalent to 4-6 base pairs (Maston2006) occurs virtually anywhere in the genome. This leads to the "futility theorem", which states that "essentially all predicted TFBS will have no functional role" in the cell (Wasserman2004) although the purified protein domain often binds the predicted oligo-nucleotide in gel shifts (Tronche1997). Therefore, in order to discriminate functionally valid sites from spurious sequence matches, several additional features have been proposed.

One of them is *helical spacing* between them, with preferred distances between sites, as a complete turn of the DNA stretches over about 10 base pairs: This is clearly supported by experimental data for certain transcription factors ((Makeev2003) and references therein). Nevertheless, (Berman2004), for instance, did not observe helical spacing in a list of *Drosophila* enhancers and in yeast, similar observations have been corrected recently (Yuan2007), noting a very weak link between site distances and the expression pattern.

The second criterion involves the strength of the match: As transcription factors recognize degenerate sequences, sites can correspond more or less to the consensus. In some high-throughput assays, regions that lack the E2F consensus motif can be bound very well (Rabinovich2008), while Su(H) recognized mostly optimal consensus sequences (Adryan2007). In the case of Foxa and Rest (Gaudet2002) (Bruce2009), the affinity of the site to the factors seems to correspond to the biological function of the enhancers.

The third proposed property is “homotypic clustering”, binding sites that occur in several, possibly degenerate, copies. This is thought to increase the thermodynamic probability of binding while transcription factors track along the chromatin (Gorman2008). A filter based on this criterion led to the identification of new enhancers when searching the *Drosophila* genome (Berman2002) (Markstein2002) and is a general feature of enhancers involved in fly blastoderm patterning (Rebeiz2002) (Lifanov2003) (Segal2008) . In mammalian genomes, it lead to non-random predictions by whole-

genome scan for predicted binding sites for factors like p53 or Rest (Zhang2006). However, other studies, on Su(H) (Adryan2007), Stat5 (Pena2005) and *C. elegans* interneuron enhancers (Wenick2004) observed that just a single binding site with no additional copies was sufficient for expression. This suggests a situation like in yeast, where only some types of binding sites tend to occur in homotypic clusters (Harbison2004). We note that the more recent experimental data and algorithms favor thermodynamical models that evaluate all matches in a certain window and as such a homotypic cluster of several weak copies and a single strong match can obtain the same score (Gertz2009) (Roider2007).

Based some of these general rules, software predictions do exist that try to detect all sequences with any cis-regulatory potential in the genome (Pierstorff2006) (Taylor2006). However, they are tested on a limited set of enhancers, from a certain type of experiment, so their results risk being biased towards the tissues that the models are trained on. In total, evidence for homotypic clustering, spatial constraints or protein affinity depends very much on the type of transcription factor analyzed. Therefore, it is difficult to derive a rule to distinguish functional from spurious binding site matches valid for all factors, tissues and organisms.

General approach of the algorithms

On the other hand, some rules have been found in examples from certain tissues. We searched the literature for studies that predicted tissue-specific enhancers, followed by a screen of the resulting DNA fragments and found 15 publications on various model organisms (see **Suppl Table**). They share a common setup: The starting point is either a collection of previously described and co-expressed enhancers from which common motifs are extracted *de novo*, without any knowledge of the factors that bind them (for reviews on this step see (Sandve2006) (Tompa2005) (MacIsaac2006)). An alternative is a set of well-known tissue-specific transcription factors and their DNA-specificities, like Dorsal in the case of dorsal-ventral patterning. The newly discovered or already known short DNA motifs are then used to search the genome or around some genes of interest for similar sequences. The crucial part is to define the "similarity" of a sequence, in the absence of BLAST-statistics that require longer alignable sequences . Do two weak

matches score higher than one strong match? How many binding sites are necessary to signal a match? Does one Kruppel site score as well a two Twist sites? Researchers have answered these questions very differently.

Enhancer prediction based on short sequence motifs

Most studies confirmed that the tissue-specific factors do leave a trace in the non-coding sequences of the gene they regulate. But they do not allow to point out a clearly superior search algorithm as the particular benchmark sets and cell types have little in common. Ahab (Rajewsky2002) and the similar but faster Cluster-Draw (Papatsenko2007) based on thermodynamical foundations obtain convincing results in the case of Drosophila patterning and the programs are available and easy to run. But they do not take into account conserved regions. EEL is the only program that focuses on conserved regions, has been validated in experiments and can be run on any computer (Palin2006). It is also the only one based on the assumption that binding site order has to be conserved.

Detailed protocols on the practical application of enhancer prediction tools have been published (Smith2008) (Papatsenko2005) (Palin2006). Most of these tools have been trained on muscle or blastoderm patterning but can be easily applied on genes that are not related to these cell types. They promise to reduce the number of *in-vivo* tests by filtering out sequences that do not fit to the model. Before validating these predictions with experiments, one should consider if other ways to benchmark the results. How many of the key transcription factors in the tissue of interest are already validated known? If not, is there a control set of known enhancers, perhaps extractable from the literature? Even in the absence of a list of known enhancers, predictions can be assessed by checking the genes flanking predicted enhancers and their expression annotation (Papatsenko2005). Gene lists for a given tissue can be downloaded from in-situ expression databases that are available now for many model organisms (See (Armit2007) for a list of resources).

The simplest score was the number of exact binding site matches within a certain window size, e.g. three dorsal binding sites within 400 bp (Markstein2002). The most complex approach took into account the affinity of the

DNA sequences to the transcription factor, competition between sites, their distances, order and the conservation in a second species (Hallikas2006). The sequences are then scanned according to this model, regions that exceed a minimum score are reported and highlighted if they are already known from the literature.

Validating predictions

There are two ways to analyze the predicted regions: Researchers can either determine the expression patterns of the flanking genes or test the predicted enhancers themselves with a basal promoter and reporter. Like all predictions, these are error-prone and unlikely to achieve 100% accuracy. The most interesting performance measure in the context of an enhancer screen is the enrichment relative to a background: If 30% of all genes in the organism are expressed in a tissue (background or random rate) but the positive share increases to 60% among the predictions, then this corresponds to a two-fold enrichment. In the following we will add binomial p-Values to this score to indicate if the enrichment values are significantly different from the background. Obviously, for enhancer tests with a reporter gene, background rate and p-Value are difficult to determine, as the total number of active enhancers for a given tissue is not known.

What can we conclude from the studies summarized in Supplemental Table 1? First of all, the majority focuses on invertebrate model organisms. The reason is probably experimental advantages, compact upstream sequences fitting into a single plasmid and development times measured in days. Furthermore, the approaches have been focusing mostly on two examples: *Drosophila* blastoderm patterning and muscle cells. The latter is one of the best-described models of transcriptional regulation in animals, with many characterized enhancers as training data. Cell cultures of muscle lead to abundant literature and one of the first and most-cited enhancer sequence analyses made data available in a convenient format (Wasserman1998). For both tissues, upstream transcription factors had been identified by previous studies, their binding sites could be searched and validated against the known data which in turn motivated experimental validations. Therefore, the only algorithm (Schroeder2004), where all predicted fragments were really enhancers, could build on decades of research on

Drosophila patterning and searched for known binding sites of nine well-known transcription factors. (Wang2006) base their prediction on only one transcription factor (GATA1), but its expression had been shown to directly lead to terminal differentiation of blood cell precursors.

We note that the complexity of the prediction algorithm seems to be less important than the type of the cells and previous knowledge about them: One of the highest rates of correctly predicted genes is achieved by a straightforward single-motif scan, based on genes expressed by two individual interneuron cells in *C. elegans*. Muscle gene identification starting with several previously completely uncharacterized motifs leads to merely a 2-fold enrichment, which is still some improvement compared to random selection. It depends on the particular gene if enrichment values of 2-4 are high enough to justify the risk of missing the essential enhancers by focusing on predictions or if it is preferred to test all conserved regions in a locus.

Perspectives

The preceding paragraphs explored the options for enhancer screens, resumed in the box on page 45. Many questions remain that have attracted little interest until now. Silencers are one of them, as negative results often do not encourage further study. But many of the validated enhancers drive expression in several tissues. The simplest, and possibly only way to restrict them specifically to a single cell type could be the addition of appropriate silencers. However, we are not aware of any silencer screens in an *in-vivo* context. It would be interesting to test some of the putative silencers from Table 2 in combination with a well-known enhancer and measure the effects. In addition, some of them might have small effects that are difficult to measure with GFP and LacZ reporter genes. An in-vivo luciferase assay like (Shin2005) would allow to quantitatively measure the effects of cis-regulatory elements onto others, as in the example of the Endo16 enhancer in sea urchin. (Yuh1998).

On the computational side, some of the presented tools make searches for short motifs in conserved cis-regulatory elements easy to use on whole genomes. However, it is surprisingly difficult the link of the resulting matches then with the already known gene data. Simple tasks like the anno-

tation of flanking genes still require programming and the extraction of other tissue-specific genes from in-situ databases is far from trivial. In addition, programs like EEL, Ahab and Clusterdraw allow to scan only one set of motifs at a time, mandating a “trial and error” approach (Palin2006), although control data sets of tissue-specific genes would permit automatic optimization of all parameters.

Both computational algorithms and wet-lab users would benefit from better curation of published studies. The first need training and benchmarking data to tune their algorithms. The latter have difficulty finding already validated enhancers that drive in the right tissue but might have been isolated in a different locus and scientific field, thereby lacking the necessary keywords in the abstract. Although more and more cis-regulatory analyses are available, vertebrate model organism databases currently do not curate transgenic sequences at all (MGI) or just expression patterns for some of them (Zfin) from publications. Third-party projects like Oreganno (Griffith2008) curate only sequences but not the expression pattern, as they lack the species-specific knowledge. It is in the interest of the scientific community working on vertebrates that model organism databases start to annotate sequences and expression patterns of enhancers, as it is current practice in the invertebrate models like *Drosophila* (Halfon2008), *C. intestinalis* (Sierro2006) (Tassy2006) or *C.elegans* (Lee2005). Then, with more and more identified enhancers, more general guidelines should emerge that will help to identify other cis-regulatory sequences .

Supplemental Table 1: Software used for enhancer prediction in the literature and their performance

Year / Authors	Number of Predictions	Benchmark predictions on known data or	Prediction Benchmark Results	Flanking gene assay (#positives/#tested)	Expected result on random flanking	Success of Enhancer Test (#positive/#tested)	Basal Promoter for enhancer	Comments
2004 Wenick Hobert				41/57 (72%)	4-8%		endogenous promoters;	no repressors found; basal promoters from muscle or gut did not work
2004 Guhathakurta	Ranked list	Mann-Whitney on tests on control gene sets	All highly significant	high ranked motifs (rank4-60): 9/10 checked with GFP fusions	low ranked genes (rank 4800-17000) 1/10	Gene not in test set: mlc-2, site mutations reduced expression to 30-60%, Double and triple mutations reduced to 7%	endogenous promoters	Control genes were: 1) Training set 2) 1200 genes from microarray data (Roy et al) 3) C. Briggsae genes 4)
2007 Zhao et al	Used top 198 to draw random genes	a) known muscle genes b) 25 regions from literature	Sens 88%, PPV 65% (random genes: 49% muscle, 36%)	Randomly 8 unannotated genes, 6 are expressed in muscle (GFP) + 1 tested module (minimal promoter)	1014 out of 2576 (39.3%)	2 by deletion, 1 with minimal promoter, all are in muscle	PES-10	50% of muscle genes with 12% of negatives detected
2002 Halfon et al	647	Flybase insitus	p = 0.0001			1 of 7	hsp	
2002 Markstein et al	15	training set		2 of 2		1 of 1	eve	Pmid 12464180 found one more enhancer
2004 Markstein et al	7 (3 were already in training set)					2/2 lacZ + 1/1 from same locus in Anopheles (no estimation of background rate)	eve	
2002 Berman et al	152, 37 with higher density, 28 are unknown	19 regions literature	14/19	28 clusters, flanking 49 genes, 49 insitus, 10 / 28 cluster are flanked	3.2% according to Berkeley In-situ collection	1 tested, giant 1.1kb	eve	
2004 Berman	37	See Berman 2002	See Berman 2002	See Berman 2002	See Berman 2002	9/27 active (3/9 do not correspond to flanking gene)	eve	Positive enhancers contain mostly sites conserved in D. pseudoobscura
2004 Schroeder et al	52	22 known out of 52 known	very unlikely to find 22 by chance, p=10^-8			13/16 (around genes with known expression), 2/5 from non-predicted elements	eve	pmid 12398796 describes AHAB alg. And the benchmark data
2006 Phillipakis et al	?					4 of 12 (no background rate?)	eve	pmid 15759656 desc. ModuleFinder software; try they all different motif combinations
2007 Goltsev et al	2 or 3					5 fragments tested, the enhancer is located in the one with the highest score	eve	pmid 17308342 describes the cluster-draw score in detail
2005 Johnson et al	269 (519 without the conservation	training set coverage & overlap score				7 out of 23 (but not specified how they were selected from the 269	Mix: forkhead and brachyury	pmid 15297614 desc. CisModule algorithm, two websites on CisModScan &
2006 Wang et al	?				Of 31 negative elements, 6 drove expression	transient + stable transfections validated in at least one assay 26/44, Chip: 10/12, 1 silencer	HBG1	Well-conserved sites are more active, one fragment is an enhancer AND a silencer
2006 Hallikas et al	42 for GLI sites, 132 for TCF sites	3 known enhancers	2 out of 3 detected	10/16 "relatively restricted" situs, 5/12 insitus in tail bud or AER, four other with tail expression	BLAST can find 1/3 enhancers, EEL 2/3	GLI: 3/best 4 drive lacZ, completely different patterns // // 4/6 diverse patterns on c-Myc/NMyc locus	TK (Goldhamer 95)	NB: Drosophila eve enhancers can be located without any alignment, see Berman 2002
2006 Pennacchio et al	Ranked list, Top30 reported					4 of 23 (17%) (while 4 of 77 (4%) of non-predicted fragments drive in forebrain	hsp68	P-value for this results is is 0.08

Chapter 2: Results

> *At present there are some hundreds of applications of computer being made in the biomedical sciences. Most of these are the work of relatively isolated research workers, who are, with few exceptions, people having extensive cross-disciplinary backgrounds.*
>
> *Report on the use of computer in biology and medicine*
> *NIH Washington, 1960*

This chapter includes three different results:

- The prediction of cis-regulatory sequences expressed in the anterior neurectoderm from duplicated GATTA motifs (submitted to Plos Biology),
- The automatic curation of cis-regulatory sequences from fulltext scientific articles (published in Genome Biology)
- An unpublished analysis of genes that keep their flanking homologs in human/ascidian comparisons, partially hold together by embedded cis-regulatory sequences

2.1 Prediction of anterior neurectoderm elements

In our lab, Lionel Christiaen had previously identified a short enhancer of the gene Pitx, expressed in a territory he called the “anterior neural boundary” (ANB), which later develops into the stomodeum and finally the oral siphon. He had hypotheses which transcription factors might bind to it, supported by mutations. It was unclear, however, how enhancers with a similar expression pattern could be found in the genome based on these ideas.

As explained in the preceding chapters, any gene-based analysis of these cells has to be based on *in-situ* expression annotation, as there is no microrray data available for the ANB. Concerning the sequences to search for, the best population of elements with cis-regulatory function is preferable: I therefore concentrated on conserved non-coding sequence alignments and on sites that are perfectly conserved within these, as suggested by (Berman2004). In order to quantify the quality of these matches, I adopted a simple binomial score, like e.g. (Schroeder2004), as genes can only be either present or absent in the target territory and there are no quantitative expression values that could be taken into account (unlike microarrays). In addition, the binomial score is quick to calculate, easy to explain and makes no assumption about the clustering or composition of binding sites: As described in the introductory chapter, there are many different models in the literature of how and where binding sites are preferentially located, but in the case of anterior neurectoderm enhancers, we have no reason to prefer one of them.

The enhancer predictions in the literature usually start with a set of motifs (sometimes automatically derived from the positive examples with motif prediction software) and then search for enhancers that fit their model. The novelty of our approach is less the score, nor the inclusion of *in-situ* data but rather the calculation of the score for all possible motif-combinations against the data. This exhaustive search is possible for several reasons, some of them due to radical simplifications: First, our score is relatively fast to calculate. Second, we use consensus sequences and even in their simplest form, the list of all non-degenerate pentamers. Third, thanks to the data

from targeted mutations, we can limit the combinations to identical pairs of motifs. All of this is reducing the number of possibilities to check.

Altogether, the system is simple enough that it allows to search the genome for all pairs of the selected motifs, count for each how often the matches flank anterior nervous system genes, calculate a P-value of this count and rank the motif by P-Value. This rather straightforward approach resulted in a pentamer that fits well into the accepted model of anterior nervous system patterning.

2.1.1 A cis-regulatory signature for chordate anterior neurectodermal genes

Maximilian Haeussler[1*], Yan Jaszczyszyn[1*], Lionel Christiaen[1,2], Jean-Stéphane Joly[1]

*These authors contributed equally to this work

[1] INRA group, UPR2197, DEPSN, Institute of Neurosciences, CNRS, 1 Avenue de la Terrasse, 91 198, Gif-sur-Yvette, FRANCE

[2] current address : Department of Molecular and Cell Biology, Division of Genetics, Genomics and Development, Center for Integrative Genomics, University of California, Berkeley, CA 94720, USA.

Background:

One of the striking findings of comparative developmental genetics was that expression patterns of core transcription factors are extraordinarily conserved in bilaterians. However, it remains unclear whether *cis*-regulatory elements of their target genes also exhibit common signatures associated with conserved embryonic fields.

Results:

To address this question, we focused on genes that are active in the anterior neurectoderm and non-neural ectoderm of the ascidian *Ciona intestinalis*. Following the dissection of a prototypic anterior placodal enhancer, we searched all genomic conserved non-coding elements for duplicated motifs around genes showing anterior neurectodermal expression. Strikingly, we identified an over-represented pentamer motif corresponding to the binding site of the homeodomain protein OTX, which plays a pivotal role in the anterior development of all bilaterian species. Using an *in vivo* reporter gene assay, we observed that 10 of 23 candidate *cis*-regulatory elements containing duplicated OTX motifs are active in the anterior neurectoderm, thus showing that this *cis*-regulatory signature is predictive of neurectodermal enhancers.

Conclusion:

These results show that a common *cis*-regulatory signature corresponding to K50-*Paired* homeodomain transcription factors is found in non-coding sequences flanking anterior neurectodermal genes in chordate embryos. Thus, field-specific selector genes impose architectural constraints in the form of combinations of short tags on their target enhancers.

This could account for the strong evolutionary conservation of the regulatory elements controlling field-specific selector genes responsible for body plan formation.

Introduction

The concept of "selector genes" was introduced 30 years ago by Garcia Bellido to define genes that interpret a transient regulatory state and specify the identity of a given developmental field (Garcia-Bellido 1975). The question of how embryos execute distinct and unique differentiation programs using these selector genes can be tackled by focusing on how gene expression is encoded in *cis*-regulatory elements and their cognate field-specific *trans*-acting factors (TF).

This concept was more recently extended to terminal selector genes that coordinate the expression of differentiation genes to determine a given cell type (Hobert 2008). In vertebrates, examples include the *Crx* TF that synergizes with another TF to control the expression of target genes in rod photoreceptors (Chen et al. 1997; Blackshaw et al. 2001; Hsiau et al. 2007). In vertebrates as well as in flies, *Crx* and its *Drosophila* homolog *Otd* act through a small *cis*-regulatory motif overrepresented in the elements flanking the target genes (Nishida et al. 2003; Tahayato et al. 2003; Alon 2007; Koike et al. 2007; Ranade et al. 2008). In addition to this evolutionary conserved network, many others in *Caenorhabditis elegans* and *Drosophila melanogaster* have shown that cell specific enhancers contain a common ‘‘tag’’ corresponding to a specific *cis*-regulatory motif, and that this motif is linked to one or a few terminal selector genes (McDonald et al. 2003; Wenick and Hobert 2004). In contrast, during early development, very few studies have reported how a set of region-specific *cis*-regulatory elements responds to field-specific selector genes. In insects, one of the best characterized sets of functionally related *cis*-regulatory elements responds to the gradient of nuclearized *dorsal* TF in the early *Drosophila* embryo (Zinzen et al. 2006; Hong et al. 2008). However, the regulatory mechanism of dorsal-ventral patterning is not enough conserved in chordates to allow comparative studies of the regulatory logics.

A more general character of bilaterians is the tripartite organization of the nervous system along the antero-posterior axis (Denes et al. 2007). In the posterior part (hindbrain and nerve cord), *Hox* genes are expressed in a colinear order. In the domain anterior to the *Hox* genes, several striking similarities in the relative expression patterns of other transcription factors

have been noted in bilaterians (Davidson 2006); (Lowe et al. 2003; Chiori et al. 2009). The *OTX*-like homeobox transcription factors (*otd* in insects) are expressed in the anteriormost part of animals as diverse as cnidarians, insects, annelids, urochordates and vertebrates (Williams and Holland 1996; Bruce and Shankland 1998; Hudson and Lemaire 2001). In chordates, OTX has a sustained expression in the anterior neurectoderm and in derivatives of anterior ectoderm such as placodes, stomodeum (Hudson and Lemaire 2001; Schlosser 2006). In mice, null-mutants of this gene lack various head structures (Acampora et al. 1995). These results suggest that OTX-like proteins belong to a conserved developmental control system operating in the anterior parts of the brain, different from the one encoded by the *Hox* complexes (Acampora et al. 2001).

Many homeodomain proteins bind to the core DNA sequence ATTA, but several subfamilies have longer binding specificities around this core (Noyes et al. 2008). OTX homeodomain proteins contain a lysine at position 50 which confers them additional specificity to guanines 5' of the ATTA motif, resulting in a core recognition sequence of GATTA/TAATC (Hanes and Brent 1991). The DNA binding domains of homeobox gene families are highly similar over large evolutionary distances and cross-species experiments have demonstrated that the OTX proteins can be exchanged between flies, mice and human without major developmental defects (Acampora et al. 1998; Acampora et al. 1999), and more recently between ascidians and mice (Acampora et al. 2001; Adachi et al. 2001).

For studies of anterior nervous system development, the ascidian *Ciona intestinalis* offers the advantage of a simple chordate body plan with the canonical tripartite brain along the antero-posterior axis (Wada et al. 1998). In addition, the genome is small, with short intergenic regions which can be aligned with another ascidian species, thus simplifying the identification of *cis*-regulatory elements (Satoh and Levine 2005). Moreover, complete expression patterns have been determined for thousands of genes and are readily available in public databases (Satou et al. 2001; Imai et al. 2004; Tassy et al. 2006). Therefore, *Ciona intestinalis* constitutes an ideal model system for combining whole genome bioinformatics and experimental *cis*-regulatory analyses.

Here, we first focus on one single anterior ectodermal enhancer in *Ciona intestinalis*. Its detailed analysis points to an internal tandem-like structure and underscores the key role of the selector gene *Otx*. We then examine if duplicated putative binding sites for OTX preferentially flank anteriorly expressed genes in the genome.

Results and discussion

D1 mediates the initiation of *Ci-pitx* expression in the anterior neural boundary (ANB)

We have previously described an enhancer sequence (called "D1", 323bp) that controls expression of the *Ciona intestinalis Pitx* gene in a sub-region overlapping the neural and the non neural ectoderm called the anterior neural boundary (ANB) (Christiaen et al. 2005). For the sake of simplicity, and although ANB has a dual origin, we label it as a derivative of the neurectoderm and call the region composed of epineural epidermis, ventro-anterior sensory vesicle and ANB, the “anterior neurectoderm” (see figure 4A). Here, we used a minimal 206 bp fragment of D1 that is sufficient to drive reporter gene expression in the ANB. We divided the remaining fragment into five parts (D1a-e) for further analysis. Deletion of the first 16pb (D1a, Fig. 1A) led to ectopic reporter gene expression in the epineural epidermis (ene) and ventro-anterior sensory vesicle (vasv) (e.g. Fig. 1E and data not shown). This indicates that D1 responds to neurectodermal *trans*-activating factors that are not restricted to the ANB.

We tested whether D1bcde controls the onset of *Ci-pitx* expression in the ANB. Endogenous *Ci-Pitx*-gene expression was not detected in ANB cells before the initial tailbud stage (Boorman and Shimeld 2002; Christiaen et al. 2002), suggesting that it starts at this stage. To test whether D1bcde recapitulates the temporal pattern of *Ci-Pitx* expression, we assayed reporter gene expression by either X-gal staining or lacZ *in situ* hybridization on the same batch of electroporated embryos fixed at successive stages. The rationale is to take advantage of the delay in β-galactosidase protein synthesis (e.g. (Bertrand et al. 2003), which should produce a marked difference between X-gal and *in situ* staining shortly after the onset of reporter gene expression. We could detect neither lacZ RNAs nor β-galactosidase activity *before* the initial tailbud stage. At this stage, however, *lacZ* transcripts could be detected in 55.4% (n=46 of N=71) of the embryos

while only 7% (n=5 of N=83) showed positive ANB cells after X-gal staining (Table S1). Hence, D1bcde-driven transcription starts at the same time as the endogenous *pitx* gene, which indicates that the D1bcde enhancer element triggers the initiation of *Ci-pitx* expression in ANB cells.

Short blocks of conserved nucleotides are required for D1 enhancer activity

Conservation between *Ciona intestinalis* and *savignyi* genomic sequences is not uniformly distributed throughout conserved non coding elements (CNEs) but rather concentrated in short blocks of identical nucleotides, which point to candidate transcription factor binding sites (TF-BS; Figs. 1A, S1A). We identified four classes of putative TF-BS based on nucleotide composition and by querying binding site databases (Matys et al. 2003; Bryne et al. 2008). One of them matches the OTX/K-50 *paired* homeodomain consensus sequence (sites O1 and O2, Fig. 1A). Other sites, called T (T/A-rich), G (G/C-rich) and M, bear resemblance to Forkhead, Smad and Meis family factors, respectively (Figs. 1A and S1A). Notably, each class of these candidate binding sites was represented at least twice in the minimal D1bcde element. The function of candidate TF-BS was tested by introducing point mutations in the corresponding blocks of conserved sequences, followed by reporter gene expression assay (Protocol S1). With the exception of mutations disrupting the “M” sites, modifications of all O, T and G sequences strongly reduced reporter gene expression in the anterior neurectoderm derivatives (Fig. S1B). Taken together, these observations indicate that D1 enhancer activity requires at least two copies for each one of three distinct classes of conserved putative TF-BS. (Fig S1).

A tandem organization of binding sites is required for D1 activity

The aforementioned observation that the essential putative binding sites occur several times in the enhancer led us to investigate whether the structure of D1 bears functional significance to its enhancer activity. Notably, the 54-bp D1(ab) element contains the three previously mentioned conserved motifs O, T and G in addition to a putative Pax binding site (P), but D1(ab) is not sufficient to enhance reporter gene transcription (Fig. 1C). Since each of the critical sites is represented at least twice in the full length en-

hancer, we asked whether D1 enhancer activity relies on this tandem-like repetition of essential binding sites.

To this aim, we created artificial enhancers containing multiple copies of D1(ab) and found that as little as two copies of D1(ab) were sufficient to drive strong lacZ expression in the anterior neurectoderm (88% of 167 tailbud embryos (Fig. 1D, E)).

To test whether enhancer activity of the D1(ab) dimer relies specifically on the duplication of O, T and G sites, we introduced point mutations in the second D1(ab) copy. Each of these mutations strongly reduced enhancer activity (Fig. 1D). These observations are reminiscent of the requirement for multiple copies of *bicoid* binding sites for target gene activation during *Drosophila* head development (Lebrecht et al. 2005). Our results demonstrate that duplications of critical binding sites are essential for D1 enhancer activity and do not constitute mere redundancy.

We next asked whether the distance between the duplicated 54bp elements influenced the activity of the artificial D1(ab) dimer. To this aim, we designed sequences that are not predicted to bind any characterized transcription factors from the Uniprobe database (see Materials and Methods) and inserted 25, 50, 75 and 150bp spacers between the D1(ab) duplicates. Overall, enhancer activity of these constructs is reduced compared to the original D1(ab) dimer and almost completely abolished with the 75bp and 150bp spacers (Fig. 1F). Similar structural constraints were reported in the *Drosophila knirps* enhancer, which was shown to require a specific arrangement of duplicated *bicoid* binding sites for activation (Ma et al. 1996; Fu et al. 2003). Similarly, *even-skipped* enhancers contain a conserved structure of paired binding sites (Hare et al. 2008), that duplicated and relatively distant (30-200bp) TFBS are necessary for a correct activity of the SV40 enhancer (Ondek et al. 1988) and the *lac* operon (Friedman et al. 1995). Taken together, our observations demonstrate that D1 enhancer activity relies on the clustering of duplicate short conserved sequences.

Ci-Otx function is required for D1 enhancer activity

Among D1(ab) essential putative binding sites, the GATTA/TAATC “O” sequences correspond to the consensus for K50-*Paired* homeodomain proteins. In ascidians, this family includes Goosecoid, Pitx and Otx, which is the sole trans-activator expressed in

the right time and place to account for D1 enhancer activity in the anterior neurectoderm in *Ciona* (Hudson and Lemaire 2001).

A functional study using morpholino antisense oligonucleotides in *Halocynthia roretzi* –another ascidian species- showed that the *Hr-Otx* knockdown strongly perturbs anterior neurectoderm development, mostly because it is required for early specification events in the gastrula (Wada et al. 2004). To avoid this early effect, we used targeted expression of dominant-negative and hyper-active versions of the *Ci*-OTX protein to interfere with its endogenous activity specifically after gastrulation. To this aim, we engineered protein chimeras between the *Ci*-OTX homeodomain and the *Drosophila* engrailed repressor peptide or the VP16 *trans*-activation domain to create dominant-negative (OTX:EnR) or hyper-active (OTX:VP16) forms, respectively. We then used the *Ci-Six3 cis*-regulatory DNA to drive expression of these fusion proteins in a region that encompasses the ANB (Fig. S2). These constructs were co-electroporated with the *Ci*-Distal-*Pitx* reporter plasmid, which contains the D1 enhancer with the two essential O1 and O2 K50-*Paired* binding sites (Christiaen et al. 2005), and the number of anterior neurectodermal cells expressing the reporter gene was scored at the mid-tailbud stage (Fig. 2). In control embryos expressing a *Ci-Six3*:Venus construct, an average of 2.78 anterior neurectodermal cells per embryo activated the *Ci-Pitx* reporter construct, which can be accounted for by the mosaic incorporation of the transgene in the four ANB cells (Fig. 2A,C). In contrast, targeted expression of *Ci*-OTX fusion proteins significantly altered *Ci-Pitx* reporter gene expression in the anterior neurectoderm: the engrailed fusion inhibited ANB expression, while OTX:VP16 produced ectopic activation in surrounding neurectodermal cells (Fig. 2B-D). These observations strongly suggest that *Ci*-OTX trans-activating inputs are required for D1 enhancer activity in the anterior neurectoderm. In addition, widespread expression of *Ci-Otx* in the anterior neurectoderm contributes to the broad D1 trans-activation potential that encompasses the ANB, epineural epidermis and anterior ventral sensory vesicle and is probably defined in D1 by the conserved GATTA/TAATC duplicated sequences.

Tandems of OTX binding sites preferentially flank anterior neurectodermal and ectodermal genes

The observation that the transcriptional response to the broadly expressed head field-selector gene *Otx* is mediated by duplicated and well-characterized GATTA motifs led us to investigate whether this regulatory architecture was overrepresented in neurectodermal genes at early tailbud stages. The basis of our approach is to compare *in situ* gene expression patterns to whole-genome sequences.

We first obtained whole mount *in situ* hybridization data for 1518 genes showing tissue-specific expression. We selected genes that are expressed in the central nervous system (CNS) and the ANB and classified them into different territories according to their expression along the antero-posterior axis: following previous reports (Wada et al. 1998; Imai et al. 2002; Dufour et al. 2006), the ascidian visceral ganglion and the nerve cord were considered as “posterior” CNS whereas the whole sensory vesicle, including the ANB, constitute the “anterior” nervous system. This lead to a detailed annotation of nervous system expression patterns for 258 genes (Table S2). From this list we retained only those 100 genes that are specifically expressed in the anterior and not the posterior parts of the CNS. Finally, we obtained annotations for 904 additional genes expressed in tissues like muscle, epidermis or notochord, from the database ANISEED (http://aniseed-ibdm.univ-mrs.fr/). This latter set of genes was used as negative controls, which allowed for background definition for further statistical analyses.

We then aimed at studying the distribution of duplicated short DNA motifs around the 904 genes to find those that show a bias towards genes expressed in the anterior or posterior nervous system, muscle, epidermis or notochord. We concentrated on conserved non-coding elements (CNEs), as these have been shown to be enriched in developmental enhancers (Woolfe et al. 2005; Pennacchio et al. 2006). To obtain these elements for the genome of *Ciona intestinalis*, we run a whole-genome alignment of it with *Ciona savignyi* (Kent et al. 2003) and removed aligned positions in transcribed regions. This results in 168306 CNEs with an average length of 143 bp.

Then, we searched for duplicate matches to all 512 possible pentamers within 125 bp of all CNEs in the *Ciona intestinalis* genome and subsequently calculated the number of tissue-specific neighboring genes associated to each duplicated conserved pentamer and tissue. The rationale for using consensus and not matrix based searches was that all sub-

classes of homeodomain proteins have well characterized binding sites that resemble pentamer motifs without degenerate positions (Berger et al. 2008; Noyes et al. 2008). For the window size parameter, we observed from our case study that the sites had to occur in duplicates with a maximum distance of about 125bp, which was the total length of the fragment between both OTX-sites in the 75bp spacer construct. The score we chose was inspired by (Yoseph Barash 2001). This “motif-tissue-score” is the negative logarithm of the binomial probability to obtain a certain number of annotated genes from a given tissue by chance and therefore reflects the association of individual pentamer motifs with specific tissues.

Our first observation was that a duplicated OTX (GATTA) motif within 125 basepairs appears among the motifs with the highest score in the anterior CNS region (Table S3). For instance, genes containing duplicated GATTA motifs within 125bp in their flanking conserved genomic DNA are more likely to be expressed in the anterior nervous system than in any of the other tissues used in this analysis including the posterior CNS (26% versus 12% or less, Table 1).

We then set out to assess the robustness of this analysis to variations of all three parameters: copy-number, window size and gene annotation. We varied the number of motif-duplicates from one to four and still obtained the highest motif-tissue scores in the anterior region with two copies. Increasing the window size from 25bp to 300bp did not change the scores to a large extent and the relative order between the anterior nervous system and other tissues always remained the same. The influence of errors in the manual annotation process was investigated by a simulation: we randomized 10% of all gene annotations and repeated this procedure 100 times. The 95% confidence intervals from these are small compared to the total differences between the tissues (Fig 3).

These results indicate that a biased distribution of GATTA motifs in CNEs supports the model of anterior ectodermal expression based on D1 enhancer analysis. We conclude that the presence or absence of multiple OTX binding sites in a CNE is a common regulatory signature of tissue specificity and of regionalized expression along the AP axis.

Duplicated GATTA-motifs identify functional anterior ectoderm enhancers

We then sought to test whether conserved sequences containing duplicated GATTA motifs act as enhancers in the anterior neurectoderm. We cloned 23 CNEs with at least

two conserved GATTAs in a 125 bp window and inserted them into a lacZ expression vector. After electroporation, we observed that ten of them are active enhancers in the anterior head at the tailbud stage (Fig. 4, Fig. S3 & Table S4). Most of the remaining non-coding regions were inactive or drove non-specific expression in the mesenchyme, as is often observed in electroporated ascidian embryos (Corbo et al. 1997; Harafuji et al. 2002). This ratio of positive elements is high compared to a previously published enhancer screen of random DNA fragments (5 active enhancers out of 138 tested fragments) (Harafuji et al. 2002) and similar to a prediction based on binding site occurences in *Drosophila* muscle founder cells (6 out of 12 tested elements) (Philippakis et al. 2006). Thus, this study shows the importance of the duplicated GATTA regulatory architecture as a predictive tag for the identification of anterior enhancers in chordates.

Could this signature also be predictive in vertebrates? (Pennacchio et al. 2006) reported that the GATTA motif is over-represented in forebrain enhancers and used it as one of six motifs to predict forebrain enhancers in the mouse genome. We found other overrepresented motifs in anteriorly expressed genes (see Supporting Table S3). Therefore, as determined experimentally with the D1 element, additional complexity must supplement the duplicated GATTA sites to achieve cell-specific expression. Similar approaches performed in *Drosophila* and *Caenorhabditis* have identified several binding sites, which correspond to factors that specify a particular fate or behaviour in a combinatorial fashion, such as the myogenic factors (Halfon et al. 2002; Philippakis et al. 2006). However, our study identifies for the first time a *cis*-regulatory signature that determines the transcriptional response to a "master" homeobox gene in a simple chordate and establishes a model for genome-wide predictions of tissue-specific enhancers.

Materials and methods

Animals

Adult Ciona intestinalis were purchased at the Station de Biologie Marine de Roscoff (France) and maintained in artificial sea water at 15°C under constant illumination. Eggs and sperm were collected from dissected gonads and used in cross fertilizations. Electroporations, using 70 µg of DNA, and LacZ stainings were performed as previously described (Christiaen 2005). Embryo staging at 13°C were done according to

(Christiaen et al. 2007; Hotta et al. 2007). Images were taken on a Leica DMR microscope.

Artificial enhancers

Plasmids with artificial enhancers were designed by cloning inserts into the pCES2::lacZ vector that contains the basal *Ci-Fkh/FoxA* promoter (Harafuji et al. 2002). Insert D1(ab) was generated by cloning two long complementary primers with XhoI/XbaI cohesive ends into pCES2. Inserts (abde), (abd), (ab)(ab-P^{del}), (ab)(ab-O^{mut}), (ab)(ab-T^{mut}), (ab)(ab-G^{mut}) were generated by cloning a second insert consisting of another couple of long complementary primers into the XbaI/BamHI site of D1(ab). The insert of D1(ab)x5 was designed *in silico*, synthetized by Genecust Europe (Luxembourg) and cloned into pCES2::LacZ between XhoI and BamHI. To obtain D1(ab)(ab), we cut out the first two parts of D1(ab)x5 with SalI/XhoI and ligated them into pCES2. The spacer sequence between both (ab) parts of D1(ab)-xx-(ab) constructs was created *in silico* by avoiding all octamers bound by homeodomain factors from a large-scale DNA-protein binding assay (Berger et al Cell 2008). We recursively added random nucleotides to an unbound sequence and backtracked if the new sequence contained an octamer with PBM enrichment score >0.3 from the UniProbe database (Newburger and Bulyk 2009). These constructs, D1(ab)-xx-(ab) are also derived from D1(ab), but the insert was synthesized by GeneScript Corporation (Piscatway, NJ, USA). We amplified spacers of the appropriate length by PCR from the longer fragment and cloned them between the two duplicated (ab) fragment by restriction/ligation

OTX fusions

A pSix3:Venus plasmid was digested by BamHI/EcoRI to eliminate the Venus/YFP reporter. VP16 fusion: the OTX_{HD} fragment was amplified by PCR from tailbud *Ciona* cDNA using OTX_{HD}-F (CGGGATCCACAATGGTATACAGTTCGTCTAGAAAA) and OTX_{HD}-R (AAACCATGG GTTGTTGCACTTGTTGGCGACA) oligos and digested by BamHI/NcoI. The VP16 domain was amplified with VP16-F (AAGATATCGACAAACCATGGTGCAGCTGGCACCACCGA CCGATGTCAG) and VP16-R (AACAGCTGGAATTCTTAGATATCCCCACCGTACTC GTCAATTC) oligos, and digested by NcoI/EcoRI. Both resulting fragments were ligated into the linearized pSix3 driver to obtain the pSix3:OTX_{HD}:VP16 construct

EnR fusion: the OTX_{HD} fragment was amplified by PCR from tailbud cDNA using OTX_{HD}-F (CGGGATCCACAATGGTATACAGTTCGTCTAGAAAA) and OTX_{HD}-R (AAACCATGG GTTGTTGCACTTGTTGGCGACA) oligos. The enR repressor domain was amplified with enR-F (CTCGAGGCCCTGGAGGATCGC) and enR-R (CGAATTCTATACGTTCAGGTCCT) oligos. Both fragments were fused by additionnal rounds of PCR using oligos that overlap the 3' part of OTX_{HD} and the 5' part of enR (enR(OTX)F:TGTCGCCAACA AGTGCAACAACTCGAGGCCCTGGAGGATCGC, OTX_{HD}(enR) R: GCGATCCTCCAGGGC CTCGAGTTGTTGCACTTGTTGGCGACA). The resulting product was digested by BamHI/EcoRI and ligated into into the digested pSix3 driver to obtain the pSix3: OTX_{HD}:enR construct.

Constructs for the enhancer screen

Plasmids containing non-coding elements were created with the Gateway Technology System (Invitrogen Carlsbad, CA, USA). We cloned an AttR3/AttR4 Gateway Cassette from (Roure et al. 2007) into the XhoI/XbaI-site of pCES2 and called the resulting construct AttR3R4-pCES2. Predicted fragments were first amplified by primers including part of the flanking AttB3/AttB4-sequences and then extended by a subsequent PCR to the full length sequences of AttB3/AttB4. These fragments were recombined with BP clonase into the P3/P4-donor Vector (Roure et al 2007) and the resulting entry vectors recombined with LR clonase into AttR3R4-pCES2 producing expression vectors.

In Silico Methods

Computational methods are described in Supporting Protocol S2. Programs that were used for whole-genome analyses are accessible at http://genome.ciona.cnrs-gif.fr/scripts/.

Acknowledgements

We wish to thank Laurent Legendre, Aurélie Heuzé and Charlotte Bureau for technical assistance; Fabrice Daian and Patrick Lemaire for *in situ* hybridization annotations; Martha Bulyk for advice on the generation of neutral spacer sequences with Uniprobe; Hiram Clawson, Angie Hinrichs, Olivier Mirabeau and Matthieu Defrance for advice on computational methods and mathematical analysis; Casey Bergman and Helene Auger for critical reading of the manuscript.

Supporting Information

Table S1. Timing of D1bcde-driven reporter gene expression.

Table S2. Annotation of expression patterns for 258 genes at early-/mid-tailbud stages.

Table S3. Motif-Tissue-Scores of the ten highest-scoring motif duplicates in the anterior and posterior nervous system.

Table S4. Overview of enhancer screen CNEs and stained territories after electroporation.

Figure S1. Mutational analysis of D1bcde reveals at least three necessary binding sites for its activity.

Figure S2. The pSix3 driver.

Figure S3. Images of LacZ-stained embryos and genomic context of the CNEs.

Figure S4. Motif-tissue scores of the 2xGATTA *cis*-regulatory signature depending on window size.

Protocol S1. Mutations of D1bcde.

Protocol S2. Computational methods.

Author contributions

MH, YJ, LC and JSJ conceived the project, designed the experiments, analyzed the data and wrote the manuscript. MH, YJ and LC performed the experiments.

Author information

Authors declare no competing interest

Correspondence and requests for materials should be addressed to JSJ

Financial disclosure

This work was supported by INRA and CNRS, the French GIS Institut de la Génomique Marine, the Marine Genomics Network of Excellence (EU-FP6 contract no. GOCE-CT-2004-505403), the ANR projects CHOREGNET and CHOREVONET, and the Pluri-genes STREP project LSHG-CT-2005-018673, MH was a recipient of a Marie Curie Early Stage Research Training Fellowship (MEST-CT-2004-504854)

References

Acampora D, Gulisano M, Broccoli V, Simeone A (2001) Otx genes in brain morphogenesis. Progress in neurobiology 64(1): 69-95.

Acampora D, Mazan S, Lallemand Y, Avantaggiato V, Maury M et al. (1995) Forebrain and midbrain regions are deleted in Otx2-/- mutants due to a defective anterior neuroectoderm specification during gastrulation. Development (Cambridge, England) 121(10): 3279-3290.

Acampora D, Avantaggiato V, Tuorto F, Barone P, Reichert H et al. (1998) Murine Otx1 and Drosophila otd genes share conserved genetic functions required in invertebrate and vertebrate brain development. Development (Cambridge, England) 125(9): 1691-1702.

Acampora D, Avantaggiato V, Tuorto F, Barone P, Perera M et al. (1999) Differential transcriptional control as the major molecular event in generating Otx1-/- and Otx2-/- divergent phenotypes. Development (Cambridge, England) 126(7): 1417-1426.

Adachi Y, Nagao T, Saiga H, Furukubo-Tokunaga K (2001) Cross-phylum regulatory potential of the ascidian Otx gene in brain development in Drosophila melanogaster. Development genes and evolution 211(6): 269-280.

Alon U (2007) Network motifs: theory and experimental approaches. Nature reviews 8(6): 450-461.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133(7): 1266-1276.

Bertrand V, Hudson C, Caillol D, Popovici C, Lemaire P (2003) Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors. Cell 115(5): 615-627.

Blackshaw S, Fraioli RE, Furukawa T, Cepko CL (2001) Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. Cell 107(5): 579-589.

Boorman CJ, Shimeld SM (2002) Pitx homeobox genes in Ciona and amphioxus show left-right asymmetry is a conserved chordate character and define the ascidian adenohypophysis. Evolution & development 4(5): 354-365.

Bruce AE, Shankland M (1998) Expression of the head gene Lox22-Otx in the leech Helobdella and the origin of the bilaterian body plan. Developmental biology 201(1): 101-112.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic acids research 36(Database issue): D102-106.

Chen S, Wang QL, Nie Z, Sun H, Lennon G et al. (1997) Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. Neuron 19(5): 1017-1030.

Chiori R, Jager M, Denker E, Wincker P, Da Silva C et al. (2009) Are Hox genes ancestrally involved in axial patterning? Evidence from the hydrozoan Clytia hemisphaerica (Cnidaria). PLoS ONE 4(1): e4231.

Christiaen L, Bourrat F, Joly JS (2005) A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. Dev Biol 277(2): 557-566.

Christiaen L, Burighel P, Smith WC, Vernier P, Bourrat F et al. (2002) Pitx genes in Tunicates provide new molecular insight into the evolutionary origin of pituitary. Gene 287(1-2): 107-113.

Christiaen L, Jaszczyszyn Y, Kerfant M, Kano S, Thermes V et al. (2007) Evolutionary modification of mouth position in deuterostomes. Seminars in cell & developmental biology 18(4): 502-511.

Corbo JC, Levine M, Zeller RW (1997) Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Development 124(3): 589-602.

Davidson E (2006) The regulatory genome. San Diego: Academic.

Denes AS, Jekely G, Steinmetz PR, Raible F, Snyman H et al. (2007) Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. Cell 129(2): 277-288.

Dufour HD, Chettouh Z, Deyts C, de Rosa R, Goridis C et al. (2006) Precraniate origin of cranial motoneurons. Proc Natl Acad Sci U S A 103(23): 8727-8732.

Friedman AM, Fischmann TO, Steitz TA (1995) Crystal structure of lac repressor core tetramer and its implications for DNA looping. Science 268(5218): 1721-1727.

Fu D, Zhao C, Ma J (2003) Enhancer sequences influence the role of the amino-terminal domain of bicoid in transcription. Molecular and cellular biology 23(13): 4439-4448.

Garcia-Bellido A (1975) Genetic control of wing disc development in Drosophila. Ciba Foundation symposium 0(29): 161-182.

Halfon MS, Grad Y, Church GM, Michelson AM (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res 12(7): 1019-1028.

Hanes SD, Brent R (1991) A genetic model for interaction of the homeodomain recognition helix with DNA. Science (New York, NY) 251(4992): 426-430.

Harafuji N, Keys DN, Levine M (2002) Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. Proc Natl Acad Sci U S A 99(10): 6802-6805.

Hare EE, Peterson BK, Eisen MB (2008) A careful look at binding site reorganization in the even-skipped enhancers of Drosophila and sepsids. PLoS genetics 4(11): e1000268.

Hobert O (2008) Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. Proc Natl Acad Sci U S A 105(51): 20067-20071.

Hong JW, Hendrix DA, Papatsenko D, Levine MS (2008) How the Dorsal gradient works: insights from postgenome technologies. Proc Natl Acad Sci U S A 105(51): 20072-20076.

Hotta K, Mitsuhara K, Takahashi H, Inaba K, Oka K et al. (2007) A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. Dev Dyn 236(7): 1790-1805.

Hsiau TH, Diaconu C, Myers CA, Lee J, Cepko CL et al. (2007) The cis-regulatory logic of the mammalian photoreceptor transcriptional network. PLoS ONE 2(7): e643.

Hudson C, Lemaire P (2001) Induction of anterior neural fates in the ascidian Ciona intestinalis. Mechanisms of development 100(2): 189-203.

Imai KS, Satoh N, Satou Y (2002) Region specific gene expressions in the central nervous system of the ascidian embryo. Mech Dev 119 Suppl 1: S275-277.

Imai KS, Hino K, Yagi K, Satoh N, Satou Y (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Development 131(16): 4047-4058.

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 U S A 100(20): 11484-11489.

Koike C, Nishida A, Ueno S, Saito H, Sanuki R et al. (2007) Functional roles of Otx2 transcription factor in postnatal mouse retinal development. Mol Cell Biol 27(23): 8318-8329.

Lebrecht D, Foehr M, Smith E, Lopes FJ, Vanario-Alonso CE et al. (2005) Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 102(37): 13176-13181.

Lowe CJ, Wu M, Salic A, Evans L, Lander E et al. (2003) Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. Cell 113(7): 853-865.

Ma X, Yuan D, Diepold K, Scarborough T, Ma J (1996) The Drosophila morphogenetic protein Bicoid binds DNA cooperatively. Development 122(4): 1195-1206.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic acids research 31(1): 374-378.

McDonald JA, Fujioka M, Odden JP, Jaynes JB, Doe CQ (2003) Specification of motoneuron fate in Drosophila: integration of positive and negative transcription factor inputs by a minimal eve enhancer. Journal of neurobiology 57(2): 193-203.

Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic acids research 37(Database issue): D77-82.

Nishida A, Furukawa A, Koike C, Tano Y, Aizawa S et al. (2003) Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. Nature neuroscience 6(12): 1255-1263.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133(7): 1277-1289.

Ondek B, Gloss L, Herr W (1988) The SV40 enhancer contains two distinct levels of organization. Nature 333(6168): 40-45.

Pennacchio L, Ahituv N, Moses A, Prabhakar S, Nobrega M et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature.

Philippakis A, Busser B, Gisselbrecht S, He F, Estrada B et al. (2006) Expression-Guided In Silico Evaluation of Candidate Cis Regulatory Codes for Drosophila Muscle Founder Cells. PLoS Computational Biology 2(5): e53.

Ranade SS, Yang-Zhou D, Kong SW, McDonald EC, Cook TA et al. (2008) Analysis of the Otd-dependent transcriptome supports the evolutionary conservation of CRX/OTX/OTD functions in flies and vertebrates. Dev Biol 315(2): 521-534.

Roure A, Rothbacher U, Robin F, Kalmar E, Ferone G et al. (2007) A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos. PLoS ONE 2(9): e916.

Satoh N, Levine M (2005) Surfing with the tunicates into the post-genome era. Genes & development 19(20): 2407-2411.

Satou Y, Takatori N, Yamada L, Mochizuki Y, Hamaguchi M et al. (2001) Gene expression profiles in Ciona intestinalis tailbud embryos. Development 128(15): 2893-2904.

Schlosser G (2006) Induction and specification of cranial placodes. Developmental Biology 294(2): 303-351.

Tahayato A, Sonneville R, Pichaud F, Wernet MF, Papatsenko D et al. (2003) Otd/Crx, a dual regulator for the specification of ommatidia subtypes in the Drosophila retina. Developmental cell 5(3): 391-402.

Tassy O, Daian F, Hudson C, Bertrand V, Lemaire P (2006) A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. Curr Biol 16(4): 345-358.

Wada H, Saiga H, Satoh N, Holland PW (1998) Tripartite organization of the ancestral chordate brain and the antiquity of placodes: insights from ascidian Pax-2/5/8, Hox and Otx genes. Development 125(6): 1113-1122.

Wada S, Sudou N, Saiga H (2004) Roles of Hroth, the ascidian otx gene, in the differentiation of the brain (sensory vesicle) and anterior trunk epidermis in the larval development of Halocynthia roretzi. Mechanisms of development 121(5): 463-474.

Wenick AS, Hobert O (2004) Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. Developmental cell 6(6): 757-770.

Williams N, Holland P (1996) Old head on young shoulders. Nature 383(6600): 490-490.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS biology 3(1): e7.

Yoseph Barash GB, Nir Friedman. Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 28-31, 2001, Proceedings. In: Moret OGaBME, editor. Lecture Notes in Computer Science; 2001 2001; Aarhus, Denmark. Springer. pp. 278-293.

Zinzen RP, Cande J, Ronshaugen M, Papatsenko D, Levine M (2006) Evolution of the ventral midline in insect embryos. Developmental cell 11(6): 895-902.

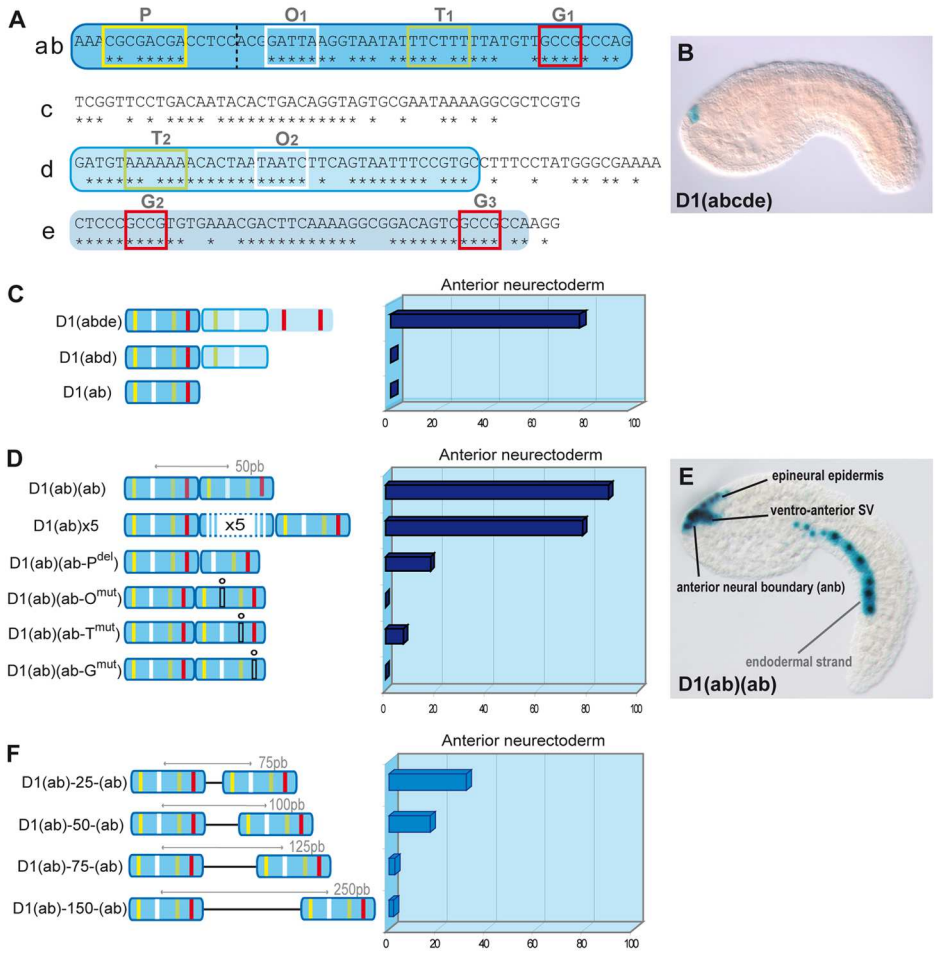


Figure 1. Artificial enhancer constructs reveal a tandem-*like* structure. (**A**) D1(abcde) *Ci-Pitx* enhancer. Stars show conserved positions with *Ciona savignyi*. The element has been divided into five parts (a, b, c, d, e). (ab) fragment is in dark blue, (d) in light blue with blue outline, (e) in light blue. Conserved nucleotides stretches contain putative transcription factor binding sites (TF-BS). O1 and O2 sites (white box) correspond to the BS for K-50 *Paired* homeodomain proteins and P (yellow), T1, T2 (green) and G1, G2, G3 (red) resemble to Pax, Forkhead and Smad protein consensus BS respectively. (**B**) Side view of an early-tailbud embryo electroporated with pD1abcde:CES2:lacZ: expression in the anterior neural boundary (ANB). In some cases, ectopic expression occurs in the mesenchyme and the tail muscles (not shown). (**C**) Expression of artificial enhancers in the anterior neurectoderm of mid-tailbud embryos. LacZ expression was observed after two hours of staining. The D1(abde) construct drives lacZ expression in the anterior neurectoderm (ANB), ventro-anterior sensory vesicle (vasv) and epineural epidermis (ene) in 77.9% of developed embryos (n=57). Deletions of (e) or (de), but not (c), abolish LacZ expression (D1(abd), D1(ab)). (**D**) Two (D1(ab)(ab)) or five (D1(ab)x5) copies of the 54bp D1(ab) drive expression in most of the embryos (88% (n=167) and 77% (n=72), respectively). Only 17% of the embryos express lacZ following the deletion of the second P site (D1(ab)(ab-P_{del}), n=90). Mutations of O, T and G sites in the second copy of (ab) strongly decrease lacZ expression. (D1(ab)(ab-O^{mut}): 0% (n=137), D1(ab)(ab-T^{mut}): 7% (n=84), D1(ab)(ab-G^{mut}): 0% (n=118)). (**E**) Mid-tailbud embryo electroporated with pD1(ab)(ab):CES2:lacZ. Expression

is visible in the ANB, the vasv, the ene and less frequently in the endodermal strand. (**F**) Introduction of spacer regions of 25/50/75/150 bp between the two D1(ab) fragments strongly decreased the activity of the tandem constructs. From 88% (D1(ab)(ab)) to 31.6% (n=76), 16.9% (n=154), 2.5% (n=79) and 1.9% (n=106), respectively. Scores were obtained after one week of LacZ revelation.

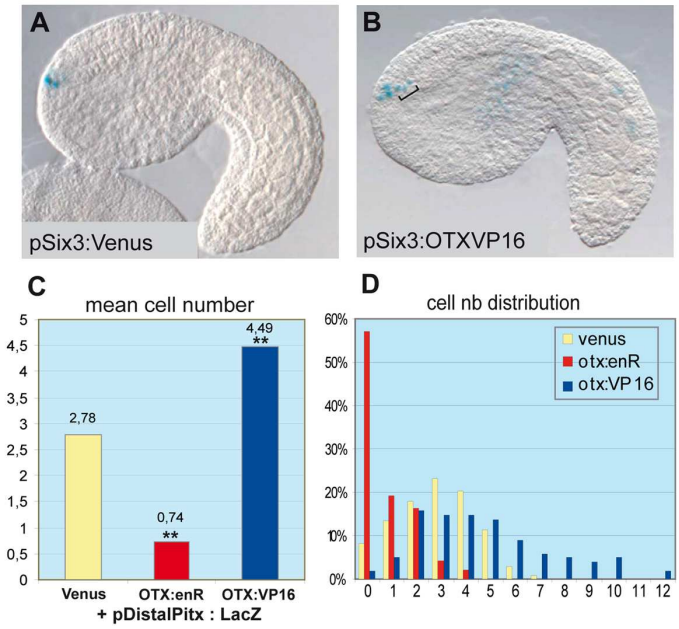


Figure 2. OTX fusions influence the activity of the *Ci-pitx cis*-regulatory element. Co-electroporation of *Pitx* full length distal region (pDistal*Pitx*:lacZ, 5.3kb, containing D1), respectively with pSix3:Venus (control), with pSix3>OTX$_{HD}$::enR (dominant negative OTX) and pSix3>OTX$_{HD}$::VP16 (hyper-active OTX). (**A**) Side view of an embryo co-electroporated with pDistal*Pitx*:lacZ and pSix3:Venus. Three positive cells can be detected in the ANB. (**B**) Co-electroporation of pDistal*Pitx*:lacZ and pSix3>OTX$_{HD}$::VP16. In addition to the expression in the ANB, ectopic expression is detected in the ASV cells (bracket) where OTX:VP16 is produced under the control of p*Six3*. (**C**) Numbers of lacZ expressing cells decrease with the OTX$_{HD}$::enR protein (2.78 to 0.74 cells) and increase with the OTX$_{HD}$::VP16 protein (2.78 to 4.49 cells) The distributions differ significantly from the control in both groups according to two *Wilcoxon–Mann–Whitney* two-sample rank-sum tests: Control/ OTXenR (U_{OTXenR} = 3891, n_{OTXenR}=139, n_{ctrl}=131, P=2.536e-07 two-tailed) and control/ OTXVP16 ($U_{OTXVP16}$ = 15582.5, $n_{OTXVP16}$=98, n_{ctrl}=131, P < 2.2e-16 two-tailed). (**D**) Distributions of cell numbers in the ANB and ASV after co-electroporation of Distal*Pitx*:lacZ and OTX fusions (yellow: control, red: enR fusion, blue: VP16 fusion; X-axis: cell numbers, Y-axis: proportions of embryos).

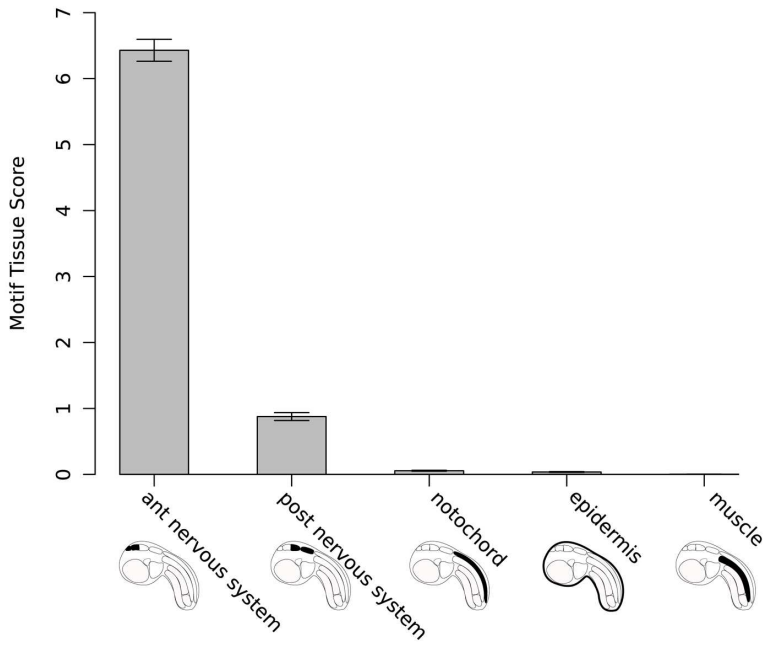


Figure 3. Motif-tissue scores for the motif 2xGATTA/125bp against genes expressed in various tissues. These territories are also visualized on schematic representation of an ascidian tailbud embryo. To illustrate that changes in gene annotation are very unlikely to affect the overall ranking, we shuffled 10% of the gene-tissue assignments, repeated the procedure 100 times and plotted 95%-confidence intervals with error bars.

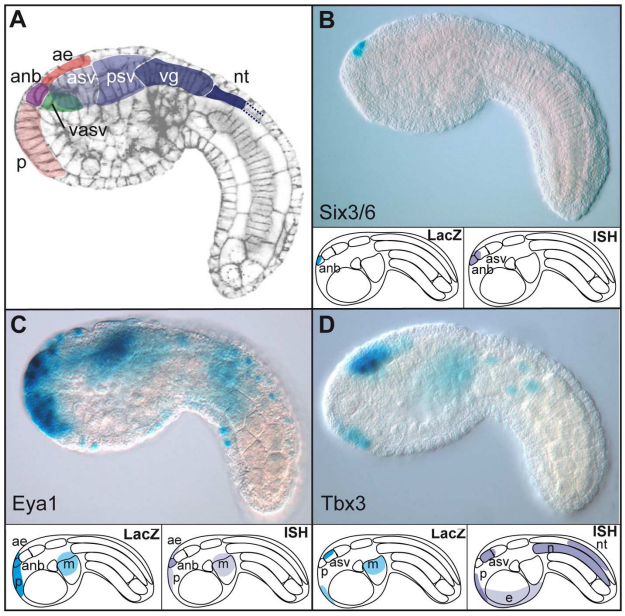


Figure 4. Enhancers with duplicated GATTA are active in the anterior region of the ascidian embryo.

(**A**) Schematic representation of the main regions of gene expressions in a mid-tailbud *Ciona intestinalis* embryo. Cell cortices are stained with Alexa-phalloidin (Christiaen et al. 2007). (**B-D**): expression domains of three enhancers, respectively from *Ci-Six3/6*, *Ci-Eya1* and *Ci-Tbx3* after electroporation and X-

Gal staining at mid-tailbud stage. Lower panels show a schematic representation of the LacZ expression driven by the enhancer (left) and endogenous gene expression as assayed by *in situ* hybridization (ISH) (right). Enhancers can be subdivided into different classes following their expression domains: very restricted expression only in the ANB while the gene expression domain is slightly larger (Six3, (**B**)); broad anterior expression recapitulating more or less the endogenous expression pattern (Eya, (**C**)); only the most anterior expression domains are driven by the enhancer (Tbx3 (**D**)). anb: anterior neural boundary, asv: anterior sensory vesicle, psv: posterior sensory vesicle, vg: visceral ganglion, nt: neural tube, ae: anterior epidermis (or epineural epidermis), p: palps (precursors), m: mesenchyme, n: notochord. Lateral views, anterior to the left

Tissue	**Genes in this category**	**Genes flanked by 2xGATTA/125 bp in a conserved non-coding alignment**	**Percentage**
anterior nervous system (specific)	100	26	0.26 *
posterior nervous system (specific)	58	7	0.12 *
notochord	346	18	0.05
epidermis	523	26	0.04
muscle	143	4	0.02

Table 1. Antero-posterior distribution of enhancers with 2xGATTA tags

* The percentage of positive anterior nervous system and positive posterior nervous system genes flanked by two GATTAs are significantly different (P=0.043, Fisher Exact two-tailed test.)

Supporting protocol S1

Mutations in D1bcde

For the mutational analysis of the enhancer D1bcde (**Fig. S1**), we omitted the first 16 bp (AAACGCGACGACCTCC) of D1abcde that were not conserved between *Ciona intestinalis* and *savignyi*. Each of the mutations was designed to perturb DNA-binding of the candidate *trans*-acting factors following various reports in the literature. Mutations were performed using the Stratagene QuickChange Kit. Seven new constructs called m0, m1, m2, m3, m4, m5/6, m7/8/9 were generated. After each electroporation, we observed LacZ expression in the tissues of the anterior neural boundary, epineural epidermis, ventro-anterior sensory vesicle and mesenchyme. We obtained a semi-quantitative estimation of the promoter activity by calculating the percentage of positive embryos (**Fig. S1**).

Supporting protocol S2: *In silico* protocols.

Annotating the expression pattern of genes in the nervous system.

We used a December 2007 version (gift of Fabrice Daian and Patrick Lemaire) of the database Aniseed (http://crfb.univ-mrs.fr/aniseed/) which is based to a large extent on images obtained from several large-scale whole-mount *in-situ* hybridization screens (Satou et al. 2001; Mochizuki et al. 2003; Imai et al. 2004; Satou et al. 2005). We only selected genes with a JGI Version 1 gene identifier that are expressed at early- or mid-tailbud stages (2396 genes) and removed all genes with the expression annotation "whole embryo" or "not expressed" at any of these two stages. This resulted in 1518 genes.

As control gene sets, we selected genes expressed in the territories "primary muscle", "epidermis" and "notochord", using the annotation in Aniseed (see Table 1 for the total number of genes in these classes).

In situ images are more difficult to annotate for nervous system sub-structures than for larger territories like muscle tissue or the epidermis. To improve the existing annotation in Aniseed, we copied the images from the Aniseed website and reannotated them manually using a simple web interface (http://genome.ciona.cnrs-gif.fr/scripts/insituFlash/ insituFlashInit.cgi). The result of the manual annotation is a list that assigns each gene to one or several of the classes "palps", "stomodeum", "tip of anterior sensory vesicle", "anterior sensory vesicle", "posterior sensory vesicle", "visceral ganglion" and "nerve cord", again dropping images of embryos with a semi-ubiquitous or weak expression. We also added 31 genes based on data from (Ikuta and

Saiga 2007) and two genes from (Auger et al. 2009). The resulting gene-annotation list is available as Supplementary Table S2.

A comparison between different territories can be confused by genes that are expressed in several domains at the same time. We for example noted that many genes in the anterior nervous system are also coexpressed in the posterior part, mostly in the visceral ganglion (see Table S2), and therefore removed from the "anterior nervous system" class genes which are expressed both in the posterior nervous system (visceral ganglion, nerve cord) and the anterior nervous system (stomodeum, sensory vesicle). Table 1 summarizes the number of genes in these two categories and the respective share of genes flanked by 2xGATTA.

Searching for genes flanked by a duplicated pentamer contained in a conserved noncoding region.

To find genes whose conserved flanking elements contain a combination of pentamers, we aligned the two repeatmasked genomes of *Ciona intestinalis* 2.0 and *Ciona savignyi* 2.0 using the UCSC BlastZ/Chain/Net/MultiZ pipeline as described in (Kent et al. 2003). Whereas BlastZ might be less sensitive than the Vista-Pipeline (Visel et al. 2007), we chose BlastZ because we are mainly interested in well-conserved sequences and because its integration with the UCSC alignment pipeline. Moreover, the Vista alignment of the genome in its latest version is completely lacking some loci, noteably the scaffold where PITX is located. The BlastZ alignment process is documented in (Auger et al. 2009) and also on http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto. The resulting alignments can be explored and downloaded from http://genome.ciona.cnrs-gif.fr. To retain only non-coding sequences for further analysis, we removed all basepairs that overlap exons or UTRs of Ensembl Release 44 gene models. These non-coding sequence alignments were subsequently annotated with their closest flanking JGI Version1 gene model ID to link them to *in-situ* annotations. The resulting alignment blocks cover 25.5 Mbp of the genome, organized in 168306 blocks that contain at least one stretch of five conserved nucleotides. The average length of these alignments on the *Ciona intestinalis* genome is 143 bp, the average number of completely conserved basepairs in them is 102bp.

As for the motifs to search in these alignments, we chose an exhaustive non-degenerate search of pentamers. This simple motif model has the advantage of reducing the overfitting problem inherent in all motif discovery approaches (MacIsaac and Fraenkel 2006). There are 512 pentamers corresponding to all possible combinations of five nucleotides. We searched the annotated consensus sequences for all possible 512 pentamers and kept only those where two identical pentamers occur within a certain distance and output the genes closest to these

alignments. We set the default maximum distance from the first to the last motif to 125 bp as this corresponds to the experimental results.

Our program (which is also accessible as an interactive web interface at http://genome. ciona.cnrs-gif.fr/scripts/cionator2/wordSearchForm.cgi) can search the non-coding alignment blocks for a given number and combinations of pentamers within a certain distance. The program outputs genes that are flanked by these matches. To assign a P-Value to a given enrichment, we calculate the binomial probability as in (Xie et al. 2005) but adapted it to a gene-based annotation. We call the logarithm of this p-Value the "motif-tissue score"; it reflects how well matches of a motif flank genes expressed in a tissue.

Obtaining a motif - tissue score from the overlap between predicted and true positive genes and influence of different parameters.

To illustrate the motif-tissue score, we give an example of a search performed with the motif GATTA, against the tissue "anterior nervous system". The foreground in this case is the population of genes specifically expressed in the "anterior nervous system" (100 genes), while all genes annotated as "anterior nervous system", "posterior nervous system", "muscle", "epidermis" or "notochord" represent the background (904 genes).

Of the 904 genes, 100 genes are expressed in the anterior nervous system. The probability to obtain the a gene expressed in the "anterior nervous system" without the knowledge of any motif is 11%. There are 68 matches to 2xGATTA/125bp in the genome; 26 of these (38%) are located in the anterior nervous system. Thus, using the GATTA-motif and testing many enhancers, one should obtain a four-fold enrichment of anterior nervous system enhancers. Using the binomial probability, we can calculate the p-Value, how probable it is to obtain 26 or more genes with anterior expression if the motif 2xGATTA had no influence on the result. The binomial probability to obtain 26 white balls or more when drawing 68 balls from an urn with 11% white balls is 5.42e-09. Taking the -log10 of this, we obtain the motif-tissue-score 8.26. This is very similar to the group specificity score of (Hughes et al. 2000; Yoseph Barash 2001; MacIsaac and Fraenkel 2006) but using the binomial probability instead of the hypergeometric one to simplify calculations, as in (Xie et al. 2005).

As can be seen from the example, the score is calculated as follows: Given a motif **m**, a list of **f** genes in the foreground and a list of **b** genes in the background, with **t** genes that are flanked (predicted) by a motif **m** and **x** genes that are flanked by motif m and are also in the foreground, we calculate the binomial probability to obtain **x** or more foreground-genes when

randomly drawing **t** times, given that the probability to draw a foreground gene is the ratio of **f/b**.

We know from the D1bcde-mutations and artificial enhancer experiments that the essential motifs for activity in the anterior neurectoderm need to be present in two copies. Pentamers fit well the binding properties of *bicoid*-like proteins. Therefore, we limit our search to all duplicated pentamers. There are 1024 different pentamers AAAAA, AAAAT...etc... to TTTTT, removing reverse complements leaves 512 different pentamers. As we are testing 512 different hypotheses at the same time, the minimal p-Value is not 0.01 but 0.01 divided by the number of hypotheses (Bonferroni-correction), leading to a minimal motif score of 4.7 to be significant. Calculating this score on different tissues and on 10%-shuffled gene sets leads to the different diagrams that are part of Figure 3. To illustrate that the window size parameters does not have a large influence on the results and that the most important parameter is indeed the list of genes expressed in a given tissue, we have plotted motif-tissues scores for the motif 2x-GATTA, different window sizes and tissues in Figure S4.

Other motifs with high motif-tissue scores

Our ranking also identified other motifs that are associated with anterior expression (See Supporting Table S2): The first is AAAAC which is also found twice in the minimal *Ci-pitx* enhancer but its mutation in the artificial enhancer does not lead to a complete reduction of expression (D1(ab)(ab-T^{mut}) of Fig. 1D). The second motif is AATTG, which is found in the 3'-part of D1 enhancer and conserved with *Ciona savignyi*. However, it seems to be dispensable for expression in the anterior neural boundary because it is present in a part of D1 that can be removed, absent from D1abbce. It represents a putative binding site for the transcription factor Hmx1/2/3. This factor is conserved in many bilaterians and its expression pattern suggests an ancestral function in rostral development (Wang et al. 2004; Wang and Lufkin 2005).

Annotation of in-situ patterns is a manual process and errors or omissions might influence the result. We tried to illustrate the impact of changes in it by a randomization experiment. We replaced 10% of the annotated genes by random genes out of the 1518 ones for which images are available, calculate the motif-tissue score for 2xGATTA and repeat this procedure 100 times. The results are shown in Fig. 3 and illustrate that the absolute value of the top motifs might be indeed sensitive to annotation error but that a lower enrichment in the posterior parts than in the anterior parts of the nervous system is found in all trials.

References

Auger H, Lamy C, Haeussler M, Khoueiry P, Lemaire P et al. (2009) Similar regulatory logic in Ciona intestinalis for two Wnt pathway modulators, ROR and SFRP-1/5. Developmental biology.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of molecular biology 296(5): 1205-1214.

Ikuta T, Saiga H (2007) Dynamic change in the expression of developmental genes in the ascidian central nervous system: revisit to the tripartite model and the origin of the midbrain-hindbrain boundary region. Dev Biol 312(2): 631-643.

Imai KS, Hino K, Yagi K, Satoh N, Satou Y (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Development 131(16): 4047-4058.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences of the United States of America 100(20): 11484-11489.

MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2(4): e36.

Mochizuki Y, Satou Y, Satoh N (2003) Large-scale characterization of genes specific to the larval nervous system in the ascidian Ciona intestinalis. Genesis 36(1): 62-71.

Satou Y, Kawashima T, Shoguchi E, Nakayama A, Satoh N (2005) An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. Zoological science 22(8): 837-843.

Satou Y, Takatori N, Yamada L, Mochizuki Y, Hamaguchi M et al. (2001) Gene expression profiles in Ciona intestinalis tailbud embryos. Development 128(15): 2893-2904.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic acids research 35(Database issue): D88-92.

Wang W, Lufkin T (2005) Hmx homeobox gene function in inner ear and nervous system cell-type specification and development. Experimental cell research 306(2): 373-379.

Wang W, Grimmer JF, Van De Water TR, Lufkin T (2004) Hmx2 and Hmx3 homeobox genes direct development of the murine inner ear and hypothalamus and can be functionally replaced by Drosophila Hmx. Developmental cell 7(3): 439-453.

Xie X, Lu J, Kulbokas EJ, Golub T, Mootha V et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. Nature aop(current).

Yoseph Barash GB, Nir Friedman. Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 28-31, 2001, Proceedings. In: Moret OGaBME, editor. Lecture Notes in Computer Science; 2001 2001; Aarhus, Denmark. Springer. pp. 278-293.

2.2 Automatic extraction of cis-regulatory sequences from the literature

At the Regcreative Meeting in 2006, Jeanette Hirschman, who is the central figure of the biological textmining community, explained during her talk that it was difficult to predict the model organism from the fulltext of papers. As the workshop also served to annotate papers and many participants like me had spend some hours during the afternoon typing primers into genome browsers to map cis-regulatory elements from publications, I asked whether it was not possible to do this automatically and use the BLAST scores to obtain the model organism for which the primers were designed. Casey Bergman was fascinated by the idea. He asked Stein Aerts if we could not integrate his pipeline that is predicting if a scientific article is focusing on cis-regulatory analysis based on keywords in the abstract. In a very short but efficient collaboration, I wrote software which downloaded thousands of PDF-files, extracted the primers and mapped them to genomes. I benchmarked the results against the database Oreganno (Griffith2008), for which I previously had written an script to import all sequences an annotation from the Mouse Enhancer Browser (Visel2007#337). Stein Aerts and Casey Bergman then described the whole pipeline in an article which was published in Genome Biology.

Research

Open Access

Text-mining assisted regulatory annotation

Stein Aerts[*†], Maximilian Haeussler[‡], Steven van Vooren[§], Obi L Griffith[¶], Paco Hulpiau[¥], Steven JM Jones[¶], Stephen B Montgomery[#], Casey M Bergman[**] and The Open Regulatory Annotation Consortium

Addresses: [*]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven, B-3000, Belgium. [†]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine, Herestraat, Leuven, B-3000, Belgium. [‡]Institut de Neurosciences A Fessard, Centre National de la Rechere Scientifique, Gif-sur-Yvette, 91 198, France. [§]Department of Electrical Engineering, Katholieke Universiteit Leuven, Heverlee, B-3001, Belgium. [¶]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, V5Z 4E6, Canada. [¥]VIB Department for Molecular Biomedical Research, Ghent University, Ghent, 9052, Belgium. [#]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK. [**]Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PT, UK.

Correspondence: Stein Aerts. Email: stein.aerts@med.kuleuven.be. Casey M Bergman. Email: casey.bergman@manchester.ac.uk

Published: 13 February 2008

Genome Biology 2008, 9:R31 (doi:10.1186/gb-2008-9-2-r31)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/2/R31

Received: 2 October 2007
Revised: 21 December 2007
Accepted: 13 February 2008

Abstract

Background: Decoding transcriptional regulatory networks and the genomic *cis*-regulatory logic implemented in their control nodes is a fundamental challenge in genome biology. High-throughput computational and experimental analyses of regulatory networks and sequences rely heavily on positive control data from prior small-scale experiments, but the vast majority of previously discovered regulatory data remains locked in the biomedical literature.

Results: We develop text-mining strategies to identify relevant publications and extract sequence information to assist the regulatory annotation process. Using a vector space model to identify Medline abstracts from papers likely to have high *cis*-regulatory content, we demonstrate that document relevance ranking can assist the curation of transcriptional regulatory networks and estimate that, minimally, 30,000 papers harbor unannotated *cis*-regulatory data. In addition, we show that DNA sequences can be extracted from primary text with high *cis*-regulatory content and mapped to genome sequences as a means of identifying the location, organism and target gene information that is critical to the *cis*-regulatory annotation process.

Conclusion: Our results demonstrate that text-mining technologies can be successfully integrated with genome annotation systems, thereby increasing the availability of annotated *cis*-regulatory data needed to catalyze advances in the field of gene regulation.

Background

The process of annotation is an essential first step in attributing biological information to genome sequences. Traditionally, the main focus of genome annotation has been the identification and annotation of well-studied biological entities, such as protein-coding genes, RNA genes and repetitive DNA. Efforts to annotate these genomic features typically adopt one of several established annotation paradigms - the 'museum,' 'jamboree,' 'cottage industry,' or 'factory' models of genome annotation (reviewed in [1,2]). Other important

functional regions of genomes that are more difficult to predict by *ab initio* or homology methods are often omitted from the standard genome annotation process, in particular the *cis*-regulatory sequences that control transcription. Instead, *cis*-regulatory sequences are typically annotated by manual curation from the literature either under the museum model in the private domain [3] or under a 'boutique' model [4] in the public domain, whereby small teams curate organism- or process-specific datasets from the primary literature for short-term research purposes. Such decentralized resources are disseminated and maintained in *ad hoc* ways that are often not integrated with the major genome database resources, and can present a bewildering array of choices to the computational or experimental end-user.

Recently, two efforts have been launched to develop integrated portals for *cis*-regulatory annotation - ORegAnno [5] and PAZAR [4] - that aim to support research in *cis*-regulatory sequence and network analysis. Both ORegAnno and PAZAR provide principled, standardized technologies for the long-term, community-driven, open-access annotation of *cis*-regulatory data in the context of the major genome database resources (for example, National Center for Biotechnology Information (NCBI), Ensembl, University of California Santa Cruz (UCSC)) and, as such, represent a new generation of resources for the annotation of *cis*-regulatory data. Despite these advances in infrastructure, many challenges still remain for the comprehensive community-based annotation of *cis*-regulatory data. First, as with all decentralized annotation efforts, community annotation of regulatory data from the literature requires systems to track the curation process, including 'triaging' relevant and irrelevant articles and monitoring the curation status of papers. Second, the scale of the *cis*-regulatory annotation challenge remains unknown, and thus it is critical to identify and prioritize the set of documents with high *cis*-regulatory potential for curation. Third, with curation times currently on the order of approximately one to two hours per paper, a major bottleneck remains in how to efficiently extract *cis*-regulatory data from primary text. Recently, rule-based information extraction systems have been developed to extract regulatory relations among pairs of genes and proteins [6-8]; however, many other types of data are necessary for comprehensive *cis*-regulatory annotation, such as the organism under investigation and, perhaps most importantly, the sequence and genomic location of *cis*-regulatory elements.

We have attempted to solve some of these challenges through the use of text-mining techniques to retrieve and extract relevant documents and data for the annotation of *cis*-regulatory networks and sequences. These efforts were inspired by (and conducted in part through) the RegCreative Jamboree [9], a workshop that was held in late 2006 that attempted to explore the interface between regulatory bioinformatics and text-mining communities. Elsewhere [10], we detail the development of a literature management system for the regulatory annotation community, which warehouses the set of papers that are likely to contain *cis*-regulatory data and maintains information on their current curation status. Here we develop a vector space model to identify Medline abstracts of papers that are likely to have high *cis*-regulatory content, and use this model to demonstrate that document relevance ranking can assist the annotation of transcriptional regulatory networks and be used to estimate the scale of the regulatory curation challenge. In addition, we show that DNA sequences can be extracted from full-text articles and mapped to genome sequences as a means to identify the location, organism and target gene information that is critical to the *cis*-regulatory annotation process. Collectively, our results demonstrate the utility (and the necessity) of employing text-mining approaches to accelerate the community-driven annotation of *cis*-regulatory sequences and networks that control transcription.

Results

A literature management system for community annotation and text mining

Assembling the set of documents that are relevant for annotation and tracking the curatorial status of papers are major challenges in community annotation. To help overcome these issues, we have developed a literature management 'queue' for the ORegAnno database, which allows registered users to input papers with known or suspected *cis*-regulatory content as targets for curation using their PubMed identifiers (PMIDs). A full description of the ORegAnno Publication Queue and its features is detailed elsewhere [10]; here, we briefly describe its contents to aid interpretation of our text-mining results. The ORegAnno Publication Queue was initially populated with expert entries obtained from the set of papers in ORegAnno plus existing sources of curated publications, including the *Drosophila* DNase I Footprint Database [11], REDfly [12], a catalog of regulatory elements for muscle-specific regulation of transcription [13,14], ABS [15], TRED [16], ooTFD [17] and DBTGR [18]. Additionally, a large number of papers were added manually by individual ORegAnno users from literature searches and review articles. Together, these PMIDs form the 'expert entry' component of the ORegAnno Publication Queue. In the current work, we show how, in addition to offering a powerful literature management system for community annotation, the ORegAnno Publication Queue offers a rich source of PMIDs for assessing information retrieval and information extraction techniques applied to biomedical text in the *cis*-regulatory domain.

A vector space model identifies Medline abstracts with high *cis*-regulatory content

As a first step in employing text-mining to aid *cis*-regulatory annotation, we attempted to identify a set of full-text papers that could enter the curation process by using information retrieval technology. To do this, we implemented a vector space model [19] that scores the approximately 16 million

scientific abstracts from Medline, each represented as a vector of index terms, against a model trained on a corpus of abstracts that *a priori* are known to have high *cis*-regulatory content. For initial model training purposes, 3,626 abstracts retrieved with the Pubmed query 'transcription and regulation and 'binding site' and (promoter or enhancer)' (see Materials and methods for details) were first split into two equal parts that form a training set (*POS1*) and a validation set (*POS2*). *POS1* contains 3,344 terms after stemming and stopword removal, representing vocabulary *VOC1*. We compared ten different relevancy rankings with *POS1* as query and either the complete *VOC1* or different subsets of *VOC1* as vocabulary. A vocabulary consisting of the 1,000 terms with the highest frequency in the full corpus yielded the highest performance when applied to *POS2* (results not shown). Similar results were obtained using a training set of 6,306 abstracts from papers previously curated in ORegAnno [5], TRANSFAC [3], or FlyReg [11]. Thus, we chose to develop our relevance ranking based on our '*cis*-regulatory' PubMed query to avoid biases towards data type, species, or other unknown factors. This approach has the additional advantage that existing sets of curated papers can legitimately be used later as validation sets. To generate the final relevancy ranking of Medline used in further analyses we used a model based on the 1,000 terms (from the 3,626 training abstracts) with the highest corpus frequency as vocabulary. Figure 1 shows the distribution of the final similarity scores for all approximately 16 million abstracts in Medline, with an indication of the top 10,000, top 50,000 and top 100,000 highest scoring abstracts in the distribution (these lists are called top10k, top50k, top100k and so on throughout the following text).

Using a similarity-based ranking rather than a classification procedure is particularly useful for our task because it does not require a negative training set, and because a similarity score allows a prioritization of documents for curation rather than a binary decision. To evaluate whether our similarity-based ranking agrees with other information retrieval technologies, we classified the entire 16 million Medline abstracts using a support vector machine (SVM) [20,21] trained on the same set of papers from our initial PubMed query as positives, and an equivalent number of randomly selected Medline abstracts as negatives. Using a radial basis function kernel, we find that 169,402 (1.07%) Medline abstracts are classified as positive and 95.6% of the top100k abstracts identified by our cosine similarity method are called positive by the SVM approach. Cosine similarity values and SVM decision function values are, furthermore, highly correlated (Pearson correlation coefficient is 0.88); 78.4% of abstracts are shared by the top100k when ranked by their cosine or SVM scores. Therefore, the cosine similarity and SVM methods both point to a very large but similar set of abstracts in Medline as having high *cis*-regulatory potential.

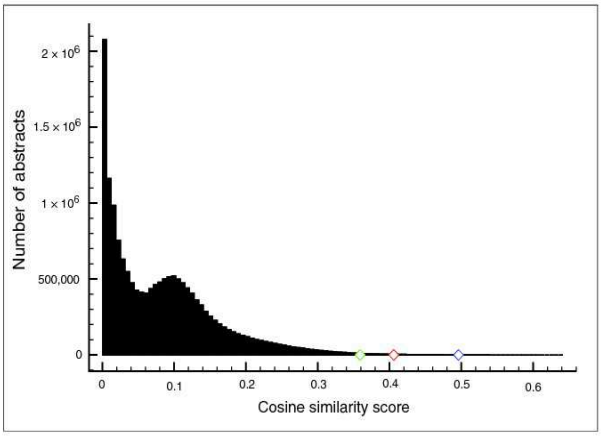


Figure 1
Distribution of cosine similarity scores between the query vector and each of the Medline abstract vectors, indicating the 10,000th (blue diamond) 50,000th (red diamond) and 100,000th (green diamond) ranked abstract.

The coverage of several validation sets within the final ranking is shown in Table 1. Before calculating the sensitivity (recall) for each validation set, we removed all Medline abstracts from these sets that were also part of the training set. As a first validation set we used TRANSFAC [3], a commercial database of manually curated transcription factor binding sites (TFBSs). We collected all 5,719 PMIDs from TRANSFAC (v10.4) that are linked to a curated TFBS. Of the set of 5,183 independent TRANSFAC PMIDs (536 were part of the training set), 75.4% are found within the top50k and 88.2% within the top100k abstracts. This shows that our model is able to generalize and recover many true positive abstracts with high *cis*-regulatory content. In fact, the vector space model realizes an increase in the proportion of TRANSFAC PMIDs from 14.7% in the 3,626 papers based on the initial PubMed query to 18.8% in the top 3,626 publications after relevancy ranking. Likewise, using a second validation set of 186 independent positive PMIDs from the FlyReg database of curated TFBSs in *Drosophila*, we find high sensitivities of 78.5% and 89.2% of FlyReg PMIDs in the top50k and top100k scoring abstracts in Medline, respectively.

Next, we investigated the coverage of true positive abstracts using curated papers from the ORegAnno database [5], including those curated as a part of the RegCreative Jamboree [9]. Prior to the Publication Queue, ORegAnno contained 376 curated papers, of which 340 are not part of the training set in the vector space model. Of these, 88.5% (n = 301) are covered in the top100k. Since the creation of the Publication Queue, curated papers are flagged with 'failure' or 'success,' depending on whether they contained enough data to allow the creation of a full ORegAnno record (that is, either a regulatory region or a TFBS with all required fields; see above). Surprisingly, in a set of 478 papers from the ORegAnno Publication Queue (see above) that were known *a priori* to have a high likelihood of containing curatable *cis*-regulatory data, only 54.4% (n = 260) were confirmed as 'success' papers during the RegCreative Jamboree. The remaining 218 'failure' papers contained either no regulatory data, or one or more

Table 1

Coverage of validation sets (excluding PMIDs in the training set) within the top10k, top50k, and top100k ranked abstracts for the vector space model relevancy ranking

	TRANSFAC	FlyReg	ORegAnno Queue	ORegAnno prior to RegCreative	RegCreative success	RegCreative failure
Number of PMIDs	5,719	200	4,145	376	260	218
Number of PMIDs (no training data)	5,183	186	3,687	340	228	212
Number in top10k	1,390	38	1,035	89	59	18
Percent in top10k	26.8%	20.4%	28.1%	26.2%	25.9%	8.5%
Number in top50k	3,908	146	2,753	260	165	79
Percent in top50k	75.4%	78.5%	74.7%	76.5%	72.4%	37.3%
Number in top100k	4,572	166	3,208	301	199	110
Percent in top100k	88.2%	89.2%	87.0%	88.5%	87.3%	51.9%

critical data fields were missing (for example, the regulatory sequence could not be identified or unambiguously mapped to a target gene or species). Excluding training abstracts, 87.3% (n = 199) of the success papers are found in the top100k but only 51.9% (n = 110) of the failure papers are found in the top100k, indicating that our relevance ranking increases the likelihood that a paper has curatable *cis*-regulatory data. Collectively, these experiments show that our vector space model successfully identifies and ranks papers with enriched *cis*-regulatory content based on Medline abstracts, and that information retrieval techniques can be used to populate a larger ORegAnno Publication Queue to assist the community annotation of *cis*-regulatory data.

Estimating the size of the *cis*-regulatory corpus

Although the sensitivities of our vector space model on evaluation sets are high, the calculations were performed on large sets of PMIDs (10k, 50k or 100k), meaning that the majority of candidate papers do not fall into any of the existing sets of curated papers. To investigate the degree to which the additional predictions show high true positive rates, we conducted a validation experiment that also gives us an indication of the scale of the *cis*-regulatory annotation challenge. We constructed a sample of 200 PMIDs evenly spaced every 500 abstracts across the top100k abstracts. Full-text papers for these 200 samples were subjected to a 'pseudo-curation' procedure in which the paper was read by an expert and, instead of being fully curated, was only scored with respect to its 'curatability' for containing a TFBS (see Materials and methods). This experiment allowed us to estimate how the proportion of true positives and false positives vary as a function of position in the ranked list of the top100k scoring Medline abstracts. Figure 2 shows the positive predictive value (PPV) for each threshold of the top100k. The first 10 samples were all success papers, indicating that the top scoring 4,501 papers are extremely likely to contain curatable *cis*-regulatory data. From then onwards, the PPV starts to decrease but still remains above 30% for the entire top100k scoring abstracts. This curve can be used to determine an optimal threshold for including papers in the ranked Medline list into the ORegAnno Publication Queue. As noted above, the proportion of success papers from the expert-entry ORegAnno Publication Queue was 54.4% during the RegCreative Jamboree. To achieve a similar curation success rate in the set of papers identified by the vector space model (namely PPV approximately 50%), we would include the top 58,000 scoring abstracts. Therefore, we estimate that the scale of the full corpus with curatable *cis*-regulatory data in Medline is on the order of approximately 30,000 papers. We note that this is a conservative measure because the success criteria are strict. Indeed, among the failure papers are many that contain regulatory data or references to other potential success papers (Figure 3). Based on these results, we added PMIDs and ranks for the top 58,000 scoring papers in Medline as 'text-mining entries' to the ORegAnno Publication Queue.

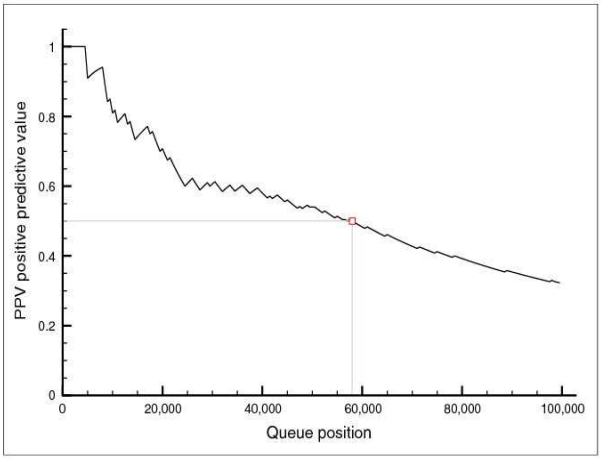


Figure 2
PPV calculated for each threshold in the top100k of the final relevancy ranking, using the pseudo-curation results of 200 evenly distributed samples. The length of the final 'text-mining entry' component of the ORegAnno Publication Queue was chosen at 58,000, which yields a PPV of 50%.

Figure 3
Results of the pseudo-curation procedure on 200 evenly distributed samples across the top100k.

Abstract relevance ranking aids the construction of regulatory networks

To illustrate the utility of identifying papers with high *cis*-regulatory content, we queried the top58k scoring abstracts for a particular transcription factor (TF), namely the *Drosophila* homeodomain-containing gene *even-skipped* (*eve*). Our goal was to use the set of papers enriched for *cis*-regulatory content to construct a literature-based transcriptional regulatory network focused on the upstream regulating factors and downstream target genes (TGs) of *eve*, based on high-quality published TFBS data. For this experiment we started with the entire list of 664 references associated with *eve* in FlyBase [22], which also includes papers not related to *cis*-regulatory data (for example, genetic interactions). We cross-referenced this list of all papers on *eve* with the top58k list to filter for papers on *eve* that are likely to contain *cis*-regulatory data. Of the 664 *eve* papers, 88 are found in the top58k list (147 are in the top100k), and for 85 of those (144 for the top100k) we retrieved the full PDF paper. We conducted a pseudo-curation analysis on these 85 papers to identify those that reported binary TF→TG relationships. We classified 35 out of these 85 candidates as 'success' papers, which revealed 43 unique binary TF→TG relationships (there were 47 relationships in total, including 4 relationships that occurred twice), 20 of which involved *eve* either as TF or as TG. A summary of the identified regulatory interactions is presented in Figure 4 as a network constructed using Cytoscape [23]. By comparison with previously curated binary TF→TG relationships for *eve* in the FlyReg database [11], our automated document retrieval process recovered 100% (12 of 12) of known upstream activating TFs, and 85% (6 of 7) of known downstream TGs. The only downstream TG curated in FlyReg that was missing in this analysis was *Abdominal-A* (*Abd-A*), which was omitted because it was not present in the original list of *eve*-related papers curated by FlyBase. These results show that cross-referencing general PMID lists for a given gene against our vector space model can enrich for papers that report direct *cis*-regulatory interactions for that gene, that transcriptional regulatory networks can be assembled from text-extracted binary TF→TG relationships [6-8,24], and that TF→TG interactions may be extracted from text even when full curation of *cis*-regulatory sequences may not be possible.

Full-text articles contain *cis*-regulatory sequences that can be automatically mapped to genomes

We also evaluated the possibility of automatically annotating *cis*-regulatory sequences from publications with high *cis*-regulatory content by extracting DNA-like strings from text and mapping these putative DNA sequences to genomes. Previously, it has been shown that short protein and nucleic acid sequence strings can be extracted from text with high precision, and that many extracted DNA sequences correspond to regulatory sequences or motifs [25]. Using automated downloads of full-text articles based on the NCBI eutils, followed by HTML-scanning for links that end with 'pdf,' we obtained PDFs for 86.9% (n = 9,940) of 11,437 papers with high *cis*-regulatory content. This recovery rate of PDFs from PMID lists is slightly higher than a rate of 79.6% reported for papers on bacterial gene regulation [8]. We converted 95.0% (9,440/9,940) of full-text PDFs into plain text files of greater than 2,000 bytes, a cutoff that represented the lower size of converted files with *cis*-regulatory content based on manual inspection. We extracted DNA-like strings from 85.4% (8,066/9,440) of these text files using a rule-based approach involving regular expressions and word size cutoffs (see Materials and methods). In total, we obtained nearly 2.8 Mb of DNA-like text from these 8,066 papers. We obtained BLAST hits of 10e-5 or greater to at least one of the five genomes under investigation for DNA sequences from 36.9% (2,975/8,066) of the PMIDs with extractable fasta sequence. Numbers of documents obtained at each stage of the process for the different source PMID lists are shown in Table 2. Overall, the proportion of papers with sequences that can be mapped to one of the five genomes is 26.0% (2,975/11,437), with the lowest efficiency step being the mapping of short sequence elements to genomes. Similar results were obtained using a previously reported Markov chain method [25] to extract DNA sequences from full-text (data not shown), with differences mainly attributable to the inclusion of lowercase DNA characters by the method of Wren *et al.* [25].

To provide biologically meaningful *cis*-regulatory annotations, automatic text-based sequence extraction must identify genomic regions that match true *cis*-regulatory elements but not a large number of other irrelevant features. To test this we used a set of 3,208 regulatory elements with known genomic location from a list of 850 'evaluation' papers with manually curated entries in ORegAnno. Three papers (PMIDs 12566409 [26], 17086198 [27] and 17558387 [28]) with 947 ORegAnno records from high-throughput experiments in humans that were imported in bulk into ORegAnno were omitted from this analysis. The numbers of regulatory elements annotated in ORegAnno, regions mapped with extracted text, and their overlap are shown in Table 3. Overall, the PPV of our approach is reasonably high (64.8%), typically with lower PPV in large mammalian genomes (42.2-70.6%) and higher PPV in small invertebrate genomes (79.3-81.3%). At the *cis*-regulatory element level, sequences overlapping approximately 33% of known ORegAnno annotations

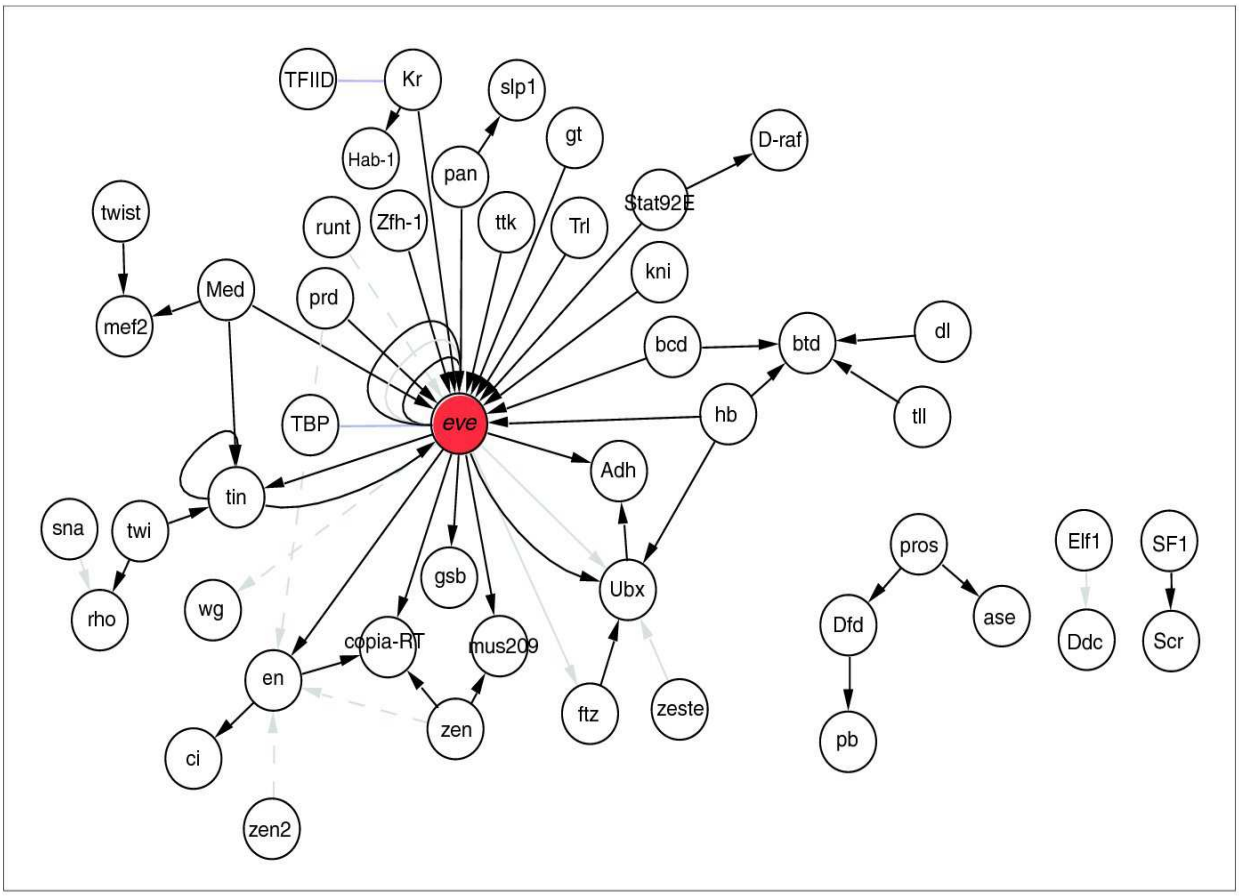


Figure 4
Transcriptional regulatory sub-network around the *Drosophila* transcription factor *even-skipped* (*eve*). All nodes and edges were retrieved from *eve*-related publications in the top100k abstract list. Black edges are success papers (that is, fully curatable publications); grey edges are failure papers that report regulatory data (for example, consensus sites) but are not the primary reference; grey dashed edges are failure papers that contain regulatory data that are not complete enough to allow full curation; blue edges are failures that report protein-protein interactions.

Table 2

Efficiency of document recovery, sequence extraction and genome mapping for the source lists of PMIDs with high *cis*-regulatory content

	TRANSFAC	FlyReg	ORegAnno	Queue	top4,501	All
Number of PMIDs	5,719	202	914	4,145	4,491	11,437
Number of PMIDs with PDF	5,302	187	835	3,710	3,677	9,940
Percent PMIDs with PDF	92.7%	92.6%	91.4%	89.5%	81.9%	86.9%
Number of PMIDs with text >2 Kbytes	5,051	175	793	3,517	3,498	9,440
Percent PMIDs with text >2 Kbytes	88.3%	86.6%	86.8%	84.8%	77.9%	82.5%
Efficiency of text conversion	95.3%	93.6%	95.0%	94.8%	95.1%	95.0%
Number of PMIDs with fasta sequence	4,357	155	660	3,044	3,080	8,066
Percent PMIDs with fasta sequence	76.2%	76.7%	72.2%	73.4%	68.6%	70.5%
Efficiency of sequence extraction	86.3%	88.6%	83.2%	86.6%	88.1%	85.4%
Number of PMIDs with fasta sequence mapped to genome	1,518	75	303	1,279	1,260	2,975
Percent PMIDs with fasta sequence mapped to genome	26.5%	37.1%	33.2%	30.9%	28.1%	26.0%
Efficiency of genome mapping	34.8%	48.4%	45.9%	42.0%	40.9%	36.9%

Note that totals are less than the sum of the sets since many PMIDs are found in more than one source list.

Table 3

Performance of text-based sequence extraction for *cis*-regulatory annotation

	dm2	hg18	mm8	ce2	rn4	All
Number of ORegAnno annotations	2,079	589	255	178	107	3,208
Number of PMIDs with ORegAnno annotation	389	283	113	30	48	850
Number of PMIDs with Ensembl target gene name(s)	388	253	107	29	42	819
Number of text hits from PMIDs with ORegAnno annotation	188	128	51	16	32	415
Number of text hits that overlap ORegAnno annotation	149	54	36	13	17	269
Percent text hits that overlap ORegAnno annotation (PPV)	79.3%	42.2%	70.6%	81.3%	53.1%	64.8%
Number of ORegAnno annotations overlapped by a text hits	681	133	149	22	64	1,049
Percent ORegAnno annotations overlapped by a text hits (SN)	32.8%	22.6%	58.4%	12.4%	59.8%	32.7%
Number of PMIDs with text hits	124	91	44	12	24	295
Percent PMIDs with text hits (coverage)	31.9%	32.2%	38.9%	40.0%	50.0%	32.2%
Number of PMIDs with text hits to correct species	123	84	37	12	18	274
Percent PMIDs with text hits to correct species (PPV)	99.2%	92.3%	84.1%	100.0%	75.0%	92.9%
Number of PMIDs with text hits and Ensembl target gene name(s)	122	77	33	11	16	259
Number of PMIDs with text hits and perfect match to correct target gene name(s)	67	57	24	4	10	162
Number of PMIDs with text hits and partial match to correct target gene name(s)	16	12	5	3	4	40
Percent PMIDs with text hits and match to correct target gene name (PPV)	68.0%	89.6%	87.9%	63.6%	87.5%	78.0%
Number of PMIDs without ORegAnno annotation with text hits	76	1,291	841	13	459	2,680
Number of text hits from PMIDs without ORegAnno annotation	126	2,602	2,131	14	1,002	5,875
Number of text hits from PMIDs without ORegAnno annotation that overlap ORegAnno annotation	59	202	58	1	18	338
Number of ORegAnno annotations overlapped by text hits from PMIDs without ORegAnno annotation	200	347	139	3	33	722

overall can be obtained directly from primary text and mapped to genomes. For *Drosophila melanogaster*, we find that text-based regulatory sequence extraction can yield annotations that have a higher PPV but lower sensitivity than the best *de novo* regulatory element prediction methods [29]. Higher sensitivities for text-based regulatory sequence prediction are observed in mouse and rat (58.4-59.8%) relative to human, worms and flies (12.4-32.8%), which can be explained by the fact that these latter species have been the subject of dedicated annotation efforts in ORegAnno and are likely to contain a deeper level of human inference in their annotation. Since only 54.4% of papers were deemed 'success' papers in the RegCreative Jamboree (see above), these relatively low sensitivities are perhaps not surprising and indicate that, in some species, we may be achieving sensitivities approaching the upper bound of what is possible automatically. An example of the accuracy and utility of text-based regulatory sequence extraction is shown in Figure 5. The *Hsp70* promoter region is duplicated seven times in the *D. melanogaster* genome, with only one locus currently annotated in FlyReg (*Hsp70Ab*). Our method cleanly extracts and correctly maps several *Hsp70* regulatory elements from full-text to genome coordinates, both from previously annotated ('evaluation') papers plus other ('prediction') papers not currently annotated in ORegAnno (Figure 5a). In addition, the unbiased nature of our method improves the current annotation of *Hsp70* regulatory sequences in *Drosophila*, with text hits mapping to all six copies of the *Hsp70* gene as well as the promoter region of the *α-γ-element* noncoding RNA gene that is expressed in response to heat shock [30,31] (Figure 5a,b).

DNA sequences extracted from text identify organisms and target genes

The organism referred to in a paper critically affects systems that attempt to recognize gene names in biomedical text and cross-reference them to external database identifiers [32]. Species identifiers are also a mandatory field in the ORegAnno curation process. Thus, we investigated if our sequence extraction and genome mapping process may provide a novel solution to the species identification problem in text mining. Of the 850 unique PMIDs with ORegAnno annotations in one or more of five species studied here (11 PMIDs have ORegAnno records for 2 species, and 1 PMID has ORegAnno records for 3 species), 295 had best genome hits obtained from extracted sequences. The correct species was identified using the genome with highest scoring BLAST hit for 92.9% (274/295) of PMIDs with hits extracted from text and ORegAnno annotations. We manually inspected the best genome hits that were incorrectly assigned to the wrong species and found that the vast majority were for hits among the three closely related mammalian species studied here (rat, mouse and human). Most of these incorrect assignments result from the requirement of a single best genome match, which can cause the wrong species identification for two rea-

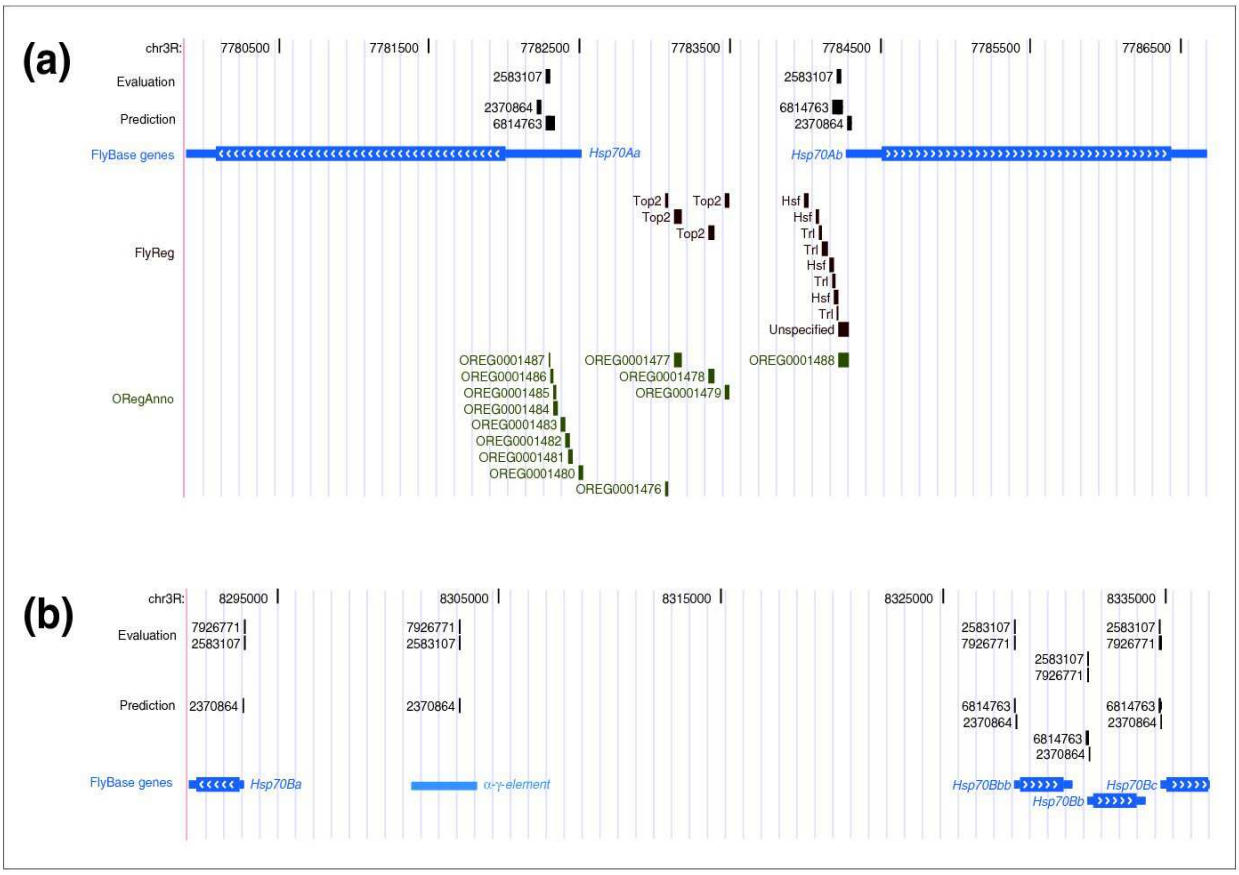


Figure 5

Comparison of automatically extracted text-based annotation and manual annotation of the *D. melanogaster Hsp70* gene regions. **(a)** The *Hsp70Aa-Ab* region. **(b)** The *Hsp70Ba-Bc* region. The 'evaluation' track refers to text-based hits extracted from papers with curated regulatory data in ORegAnno; the 'prediction' track refers to text hits extracted from papers not currently curated in ORegAnno, but with high predicted *cis*-regulatory content. Annotations in both text-based tracks are labeled with their corresponding PMIDs. Also shown are the original manual annotation in the FlyReg database, the automated mapping of these curated data in ORegAnno, and FlyBase genes, including the *α-γ-element* noncoding RNA gene that is expressed in response to heat shock. Differences in the FlyReg and ORegAnno mappings in (a) arise because the sequences for these regions are duplicated in the genome and alternative unique mappings are chosen in the two databases.

sons: first, a single PMID may report sequences (and therefore have ORegAnno records) for multiple species but only a single species gets chosen; second, only a single species is reported in the paper and annotated in ORegAnno, but the wrong species is assigned because the sequences (and BLAST scores) in another species are identical. In addition, a small number of 'incorrect' species assignments are because the species was actually incorrectly curated in the current ORegAnno annotation (for example, OREG0000115). These incorrect annotations have been deprecated and replaced by correct annotations in ORegAnno (for example, OREG0004685). These results demonstrate that primary text contains valuable information about the species under investigation encoded in extractable DNA sequences, but that mistaken species assignments may occur among closely related species or when sequences from multiple species are reported in a single paper.

Gene name recognition and normalization to database identifiers is an essential step in many text mining applications, but is a challenging task because of ambiguity and variation in how genes are named and used [33]. The identity of the target gene regulated by a *cis*-regulatory sequence is a key piece of information in regulatory bioinformatics and is a required field in an ORegAnno annotation. Thus, we investigated whether it is possible to automatically identify the target gene of putative *cis*-regulatory sequences extracted from text and mapped to genomes. To do this we simply identified the closest Ensembl gene to each text hit that was mapped to one of the five genomes. In the case of text hits found in introns, the closest gene was predicted to be the gene containing the intron, even if additional genes were present within the intron that were closer to the text hit. Each hit for PMIDs that generated multiple genomic hits was assigned its own putative target gene and evaluated for whether any of the PMID-target gene relationships were found in ORegAnno. For this analysis, we used a set of 259 PMIDs with ORegAnno annotations

that provided a best hit to one of the five genomes and for which one or more predicted target gene names were found in the set of Ensembl normalized GeneIds in ORegAnno. For 162 PMIDs, the list of closest genes matched the list of correct target genes perfectly, and for an additional 40 PMIDs there was a partial match between the list of putative target genes and the true list of target GeneIDs in ORegAnno. Overall, 78.0% of PMIDs generated at least one text hit whose closest gene was the correct target gene. In general, extracting sequences from text yields a higher proportion of correct target genes (87.5-89.6%) in the larger mammalian genomes where gene density is relatively low. In contrast, in the compact genomes of *D. melanogaster* and *Caenorhabditis elegans*, a lower proportion of target genes is correctly identified (63.6-68.0%) since a text hit can have a higher probability of being closer to a neighboring gene than its true target in a compact genome. Remarkably, our simple DNA sequence-based gene name recognition method achieves levels of PPV (precision) that are higher than the median performance in BioCreAtIvE Task 1B [32] of advanced gene name recognition systems for flies (65.9%) and mice (76.5%). Additionally, since each PMID with a text extracted hit leads to at least one predicted target gene, our sequence extraction method identifies gene names from full-text articles at a rate (26.0%) comparable to dictionary-based gene name recognition in Medline abstracts (19.4%) [34].

A draft annotation of more than 2,000 papers with high *cis*-regulatory content

Among the 10,587 papers not currently curated in ORegAnno in our set of 11,437 PMIDs with high *cis*-regulatory content, we obtained hits to 5,875 genomic regions from 2,680 PMIDs. If we assume that approximately 65% of text hits from these 'prediction' papers are true positives (based on the overall PPV estimates above), we expect that approximately 3,800 of these text hits correspond to *cis*-regulatory sequences. The addition of these records would increase the number of annotations curated from small-scale experiments in ORegAnno by approximately 120%. Indeed, many of these are likely to be *bona fide* regulatory sequences, as shown by the fact that 338 text hits from papers not currently curated overlap 722 pre-existing ORegAnno annotations. For example, PMIDs 6814763 [35] and 2370864 [36] (which were both identified as having high *cis*-regulatory content by our vector space model) each provided an extractable sequence that mapped to previously annotated *cis*-regulatory elements in the *Hsp70* promoter (Figure 5a). This result suggests even the most highly curated genomes have yet to achieve 'saturation annotation' and that a high level of redundant publication may exist for some regulatory elements, which can be used to support or extend current ORegAnno annotations. These predictions are not sufficient to stand as full ORegAnno records on their own, but should substantially decrease the time needed for the community annotation of these papers. In addition, these regions may be of sufficient resolution to be used by other workers in regulatory bioinformatics, and for these reasons we provide browser extensible data (BED) files for text-extracted sequences from both evaluation and prediction papers for the *D. melanogaster* (Additional data file 1), human (Additional data file 2), mouse (Additional data file 3), *C. elegans* (Additional data file 4), and rat (Additional data file 5) genomes.

Discussion

A principle aim of genome biology is to decode complete transcriptional networks, so as to better understand how the activation of specific subnetworks affect developmental processes or responses to the environment, and how variation in transcriptional networks can lead to functional diversity over evolutionary time. As with all grand challenges in interpreting genome sequences, solving this ultimate aim will require combining both computational and experimental approaches. As the reliability of predictive regulatory sequence bioinformatics is relatively low [37], high-throughput experimental techniques currently prove to be the most efficient means of identifying regulatory sequences and assembling regulatory networks [38,39]. The gold standard for evaluating both computational and high-throughput experimental techniques continues to be the sizable body of prior knowledge contained in small-scale experimental studies on *cis*-regulatory sequences, much of which remains locked in the biomedical literature. Here we have shown that application of text-mining technologies, including literature management, information retrieval and information extraction systems, can accelerate the community annotation of *cis*-regulatory networks and sequences. These advances should help generate the necessary training and test sets to improve the reliability of computational and high-throughput experimental methods in regulatory biology.

Previously, it has been shown that manually curated and automatically extracted binary TF→TG interactions can be assembled into transcriptional regulatory networks [6-8,24]. Here we show that abstract relevance ranking using a vector space model can be used to enhance the manual annotation of binary TF→TG interactions, and should likewise further improve the automated extraction of binary TF→TG interactions to construct regulatory networks. We have also shown that the binary TF→TG interactions that are central to the construction of transcriptional regulatory networks can be extracted from text even when a full curation of the *cis*-regulatory sequence responsible for this interaction may not be possible. Our vector space model also has allowed us to generate an enhanced 'queue' of papers for annotation, and to gain a deeper insight into the size of the corpus of papers that may contain curatable *cis*-regulatory sequences, which we estimate is on the order of 30,000 papers or more. At the rate of approximately 1-2 hours curation time per paper, it would take a single person approximately 15-30 years to curate and annotate this corpus manually. This estimate demonstrates the need for distributed community annotation systems and

for computational tools that can assist the extraction of relevant *cis*-regulatory information.

We have also investigated the potential of exploiting information contained in the DNA sequences reported in papers with high *cis*-regulatory content to assist regulatory annotation. Given the large number of DNA, RNA and peptide sequences reported in the biomedical literature, and the fact that sequences important enough to deserve mention in publication are likely to be of high biological significance, surprisingly little work has been conducted on extracting sequences from primary text [25,40]. The pioneering work of Wren *et al.* [25] showed that Markov models trained on English text, proteins and/or genomic DNA can be used to extract both DNA and peptide sequences from abstracts and full text with high precision. Wren *et al.* [25] also demonstrated that the extraction of DNA is more precise than peptides, and that the terminological context of the majority of extracted DNA sequences revealed that the sequence was likely to be a 'regulatory site' or 'motif' [25]. Our results directly support the claim that primary text contains a large number of DNA strings that are *cis*-regulatory sequences, which we also show can be automatically mapped to genome sequences to accelerate and enhance regulatory annotation. In addition to validating our approach, overlaps between ORegAnno annotations and text-based hits can be used as an automatic procedure to authenticate ORegAnno annotations, which can be indicated in the 'Score' profile for each ORegAnno record. As identifying and annotating *cis*-regulatory sequences in genomes currently remain among the most challenging branches of bioinformatics, ironically it may now be easier and more productive to identify functional *cis*-regulatory sequences in biomedical text rather than in DNA itself.

Our rule-based system for extracting and mapping DNA sequences could potentially be improved in several ways. One area to explore would be to implement more sophisticated sequence recognition techniques such as Markov models [25], although our initial comparisons suggest very similar overall performance. Inclusion of lowercase letters or degeneracy in the DNA alphabet of our rule-based method may allow many more *cis*-regulatory motifs to be extracted, but may also allow many more DNA-like English words to be extracted. Aside from variation in formatting [25], DNA strings in text should be easily discernable from English words and, therefore, identifiable by many alternative methods, since the upper limit of English words that can be spelled entirely in the DNA alphabet is small. For example, in a dictionary of approximately 355,000 English words [41], only 47 can be spelled entirely in DNA letters [ACGT], with an upper length of 7 characters for the word 'attacca,' a directive used at the end of a piece of music that is unlikely to be found in biomedical text. Inclusion of the entire set of ambiguity codes for DNA [ACGTMRWSYKVHDBXN] leads to a maximal English word size of only 13 characters for 'dharmashastra,' an ancient form of Indian jurisprudence. Thus, the vast majority of DNA-like strings of sufficient length to be mapped unambiguously to genomes are almost certainly *bona fide* DNA sequences. The main challenge for extracting DNA from text will be inaccuracies in the text encoding in older PDF documents, and the fact that many DNA sequences are embedded in tables, figures and supplementary materials. Although some figures have corresponding text encoded in the PDF, the use of text-recognition algorithms that operate on images would almost certainly improve the predictive power of our approach, and preliminary experiments have shown that this is the case (results not shown).

The area with the largest scope for improvement in using DNA in text to annotate genomes is the mapping of sequences to genomes (Table 2), in part because of the short length of many *cis*-regulatory sequences. One way to solve this problem would be to combine sequence extraction with term recognition [25] to identify species or target gene names that could be used to reduce the search space for mapping extracted sequences to genomes. Another improvement would be to accept mappings to multiple species, which is also a more realistic solution than the requirement for a single 'best' species since the biological function of a reported sequence is likely to be the same closely related species. Improvements may also come from more lenient BLAST thresholds or the use of non-RepeatMasked versions of genomes, although these would almost certainly lead to higher false positive rates. Mapping regulatory sequences to repetitive genomic regions is a general problem, not only for text-extracted sequences, but also for manually curated data (Figure 5a). However, since many *cis*-regulatory elements may arise from transposable element sequences [42] or be located in segmental duplications (Figure 5), it will be necessary to solve the problem of representing and storing repetitive *cis*-regulatory elements for comprehensive regulatory annotation.

As presaged by Lincoln Stein [1], our results demonstrate that it is indeed possible to leverage text-mining technologies to accelerate genome annotation. Our proof of principle in the field of regulatory annotation is only one potential application of text-based genome sequence annotation. The general combining of information retrieval systems (for example, [19]) with sequence extraction techniques (for example, [25]) should allow researchers to enrich for any specific subdomain of biomedical research and use sequence data reported in these corpora to directly annotate genomic regions of interest in a highly automated fashion. For example, the false positive mappings that correspond to coding sequences in our set of documents with high *cis*-regulatory content (see above) are likely to be mainly for proteins that bind to *cis*-regulatory sequences, and thus strategies similar to ours could accelerate the labor intensive identification of sequence specific TFs [43,44]. Clearly, it is preferable that researchers deposit and store their sequences and annotations in databases as a condition for publication and thereby

preclude the need for post-publication extraction of such valuable biological data. With established databases for general sequence submission (for example, [45]) and specialized *cis*-regulatory annotation [4,5], researchers now have the necessary tools to deposit and archive their *cis*-regulatory data. In the absence of direct database submission, we recommend that researchers report certain minimum information (that is, absolute coordinates with genome build, sequence with sufficient flank, standard gene identifiers, official species name or identifiers) to assist the regulatory annotation (both human and automated) that is needed to help catalyze advances in the field of gene regulation.

Materials and methods

Implementation of a vector space model to identify Medline abstracts with high *cis*-regulatory content

To identify papers with potential *cis*-regulatory data for community annotation, we used a vector space model [19] that represents each of the approximately 16 million scientific abstracts in Medline as a vector of index terms. Each vector element is a weight that is proportional to the relative importance of the term in the abstract (using the inverse document frequency or IDF). Relevancy ranking of the corpus is then achieved by calculating the similarity between each abstract and a query. This query can be represented by the same kind of vector as the documents, so that the similarities can be calculated by the cosine similarity measure between individual abstract vectors and the composite query vector. In practice, a good query vector can be constructed from the average properties of a training set of true positive abstracts. In this study, we used a '*cis*-regulatory' PubMed query that yielded a very high amount of true positives to generate our training set, namely: 'transcription and regulation and 'binding site' and (promoter or enhancer)'.

Pseudocuration of full-text articles

To evaluate the ability of our model to predict papers with high *cis*-regulatory content, we selected 344 papers from the top 100,000 scoring abstracts, of which 200 are uniformly distributed and 144 are related to the *Drosophila* transcription factor *eve*. Because the full curation of all 344 papers would require the organization of a second annotation jamboree, we opted for a distributed 'pseudocuration' procedure. Particularly, nine experienced curators examined whether these papers describe experimentally verified regulatory data and, if so, whether they also contain all the required data to allow genome annotation (that is, at a minimum the species, the sequence and its genomic location, the TF, and the TG). A web application was created where the curators could open a pending PMID and score the full-text paper as success or failure. Failures could be of four types: the publication describes binding site or promoter but there is insufficient information to annotate it; the publication describes transcription factor (complex) but not a binding site or promoter; the publication describes consensus binding sites or a reference to a primary publication but is itself not the correct source for annotation; and the publication does not describe a regulatory element. Regulatory interactions in the form of TF→TG were recorded as free text.

Extraction of DNA sequences from full-text and mapping to genome sequences

A unique list of 11,437 PMIDs was compiled from papers previously curated in FlyReg [11], ORegAnno [5] TRANSFAC v10.4 [3], plus unannotated papers in the ORegAnno Publication Queue, and the top 4,501 scoring abstracts identified by the vector space model that are extremely likely to contain *cis*-regulatory data (see above). To allow access to information in both older and more recent articles, full-text was downloaded automatically as PDFs where available using a custom script employing NCBI eutils [46]. PDFs were converted to plain text using pdftotext (v3.0) with option '-nopgbrk' [47]. Text was split into words and words greater than 10 characters in length with greater than 40% of characters from the capitalized DNA alphabet [ACGT] were extracted using regular expressions to isolate putative DNA sequences. All putative DNA sequences extracted from each paper were concatenated in the order they appeared in the text into a single fasta sequence and labeled with the corresponding PMID. Concatenation of sequences was performed to merge sequences split by line breaks in the text conversion, and because we reasoned that inappropriate joins would be reconciled at the genome level by local alignment procedures. Extracted, concatenated sequences were used as queries to BLAST RepeatMasked versions of genome sequences downloaded from the UCSC genome database [48] for the five species with greater than 100 ORegAnno database annotations: *D. melanogaster* (dm2), human (hg18), mouse (mm8), *C. elegans* (ce2) and rat (rn4). We note that these five genomes represent approximately 99% of the records currently in ORegAnno. NCBI-BLASTN v2.2.10 [49] was used to map extracted sequences to genome coordinates with an E-value cutoff of 10e-5. BLAST output was parsed into BED format using Jim Kent's source tree utilities, blastToPsl and pslToBed [50]. BLAST results for all five species were concurrently searched to find the genome that provided the best sum of BLAST scores to each fasta sequence, and this list of PMID-best genome matches was used to filter BED files to minimize spurious cross-species mapping. We then joined fragmented hits in the same genomic interval by clustering BED annotations for the same PMID within 1.0 KB on the same chromosome. Filtered, clustered BED annotations were assessed for their overlap with the 20-JUL-2007 mapping of ORegAnno annotations [51] using the Kent source tree utilities overlapSelect and bedIntersect. Finally, we identified a single putative target gene for each hit as the Ensembl [52] GeneId closest to each filtered, clustered BED annotation.

Abbreviations

BED, browser extensible data; NCBI, National Center for Biotechnology Information; PMID, PubMed Identifier; PPV, positive predictive value; SVM, support vector machine; TF, transcription factor; TFBS, transcription factor binding site; TG, target gene; UCSC, University of California Santa Cruz.

Authors' contributions

SA, MH, SvV and CMB conceived of the study and conducted the text mining experiments and analysis. SA, OLG, SJMJ, SBM and CMB designed and implemented the ORegAnno Publication Queue. SA, MH, OLG, PH, SJMJ, SBM, CMB and The Open Regulatory Annotation Consortium contributed to the curation activities of the RegCreative Jamboree. SA and CMB drafted the manuscript and all authors read and contributed to the final manuscript.

Additional data files

The following additional data files are available. Each additional data file is a UCSC genome BED formatted file that lists the chromosome, start coordinate, stop coordinate and PubMed identifier of text-extracted sequences on UCSC genome browser assemblies. Additional data file 1 provides genomic coordinates of text hits to the dm2 version of the *D. melanogaster* genome. Additional data file 2 provides genomic coordinates of text hits to the hg18 version of the human genome. Additional data file 3 provides genomic coordinates of text hits to the mm8 version of the mouse genome. Additional data file 4 provides genomic coordinates of text hits to the ce2 version of the *C. elegans* genome. Additional data file 5 provides genomic coordinates of text hits to the rn4 version of the rat genome.

Acknowledgements

We thank Jonathan Wren for help running his Markov sequence extraction method as well as all of the participants of the RegCreative Jamboree for many fruitful discussions before, during and after the Jamboree. We are especially grateful to Martin Krallinger, Lynette Hirschman, Alfonso Valencia and Ewan Birney for encouraging links between the regulatory informatics and text-mining communities. SA is Postdoctoral Research Fellow of the FWO-Vlaanderen; MH is supported by a Marie Curie Early Stage Research Training Fellowship (MEST-CT-2004-504854) and the Plurigenes STREP project (LSHG-CT-2005-018673); OLG is supported by the Canadian Institutes of Health Research and the Michael Smith Foundation for Health Research; SBM is supported by the European Molecular Biology Organization and the Natural Sciences and Engineering Research Council of Canada. We also thank ENFIN, the BioSapiens Network, the Research Foundation - Flanders (FWO-Vlaanderen), Genome Canada and Genome British Columbia for financial support of the RegCreative Jamboree. This work is conducted as part of the NESCent *cis*-regulatory evolution working group supported by the NSF National Evolutionary Synthesis Center (NSF #EF-0423641).

References

1. Stein L: **Genome annotation: from sequence to biology.** *Nat Rev Genet* 2001, **2**:493-503.
2. Elsik CG, Worley KC, Zhang L, Milshina NV, Jiang H, Reese JT, Childs KL, Venkatraman A, Dickens CM, Weinstock GM, Gibbs RA: **Community annotation: procedures, protocols, and supporting tools.** *Genome Res* 2006, **16**:1329-1333.
3. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34(Database issue)**:D108-D110.
4. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: **PAZAR: a framework for collection and dissemination of *cis*-regulatory sequence annotation.** *Genome Biol* 2007, **8**:R207.
5. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ: **ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.** *Bioinformatics* 2006, **22**:637-640.
6. Saric J, Jensen LJ, Rojas I: **Large-scale extraction of gene regulation for model organisms in an ontological context.** In *Silico Biol* 2005, **5**:21-32.
7. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2006, **22**:645-650.
8. Rodriguez-Penagos C, Salgado H, Martinez-Flores I, Collado-Vides J: **Automatic reconstruction of a bacterial regulatory network using Natural Language Processing.** *BMC Bioinformatics* 2007, **8**:293.
9. **The RegCreative Jamboree** [http://www.dmbr.ugent.be/bioit/contents/regcreative/]
10. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ, The Open Regulatory Annotation Consortium: **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Res* 2008, **36(Database issue)**:D107-D113.
11. Bergman CM, Carlson JW, Celniker SE: ***Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*.** *Bioinformatics* 2005, **21**:1747-1749.
12. Gallo SM, Li L, Hu Z, Halfon MS: **REDfly: a Regulatory Element Database for *Drosophila*.** *Bioinformatics* 2006, **22**:381-383.
13. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278**:167-181.
14. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic Acids Res* 2005, **33**:3154-3164.
15. Blanco E, Farré D, Albà MM, Messeguer X, Guigó R: **ABS: a database of annotated regulatory binding sites from orthologous promoters.** *Nucleic Acids Res* 2006, **34(Database issue)**:D63-D67.
16. Zhao F, Xuan Z, Liu L, Zhang MQ: **TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies.** *Nucleic Acids Res* 2005, **33(Database issue)**:D103-D107.
17. Ghosh D: **Object-oriented Transcription Factors Database (ooTFD).** *Nucleic Acids Res* 2000, **28**:308-310.
18. Sierro N, Kusakabe T, Park K-J, Yamashita R, Kinoshita K, Nakai K: **DBTGR: a database of tunicate promoters and their regulatory elements.** *Nucleic Acids Res* 2006, **34(Database issue)**:D552-D555.
19. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information.** *Genome Biol* 2004, **5**:R43.
20. Joachims T: **Making large-scale support vector machine learning practical.** In *Advances in Kernel Methods: Support Vector Learning*. Edited by Schölkopf B, Burges C, Smola A. MIT Press; 1999:169-184.
21. **SVM Light** [http://svmlight.joachims.org/]
22. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM, The FlyBase Consortium: **FlyBase: genomes by the dozen.** *Nucleic Acids Res* 2007, **35(Database issue)**:D486-D491.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**:2498-2504.

24. Ashburner M, Bergman CM: ***Drosophila melanogaster*: a case study of a model genomic sequence and its consequences.** *Genome Res* 2005, **15:**1661-1667.
25. Wren JD, Hildebrand WH, Chandrasekaran S, Melcher U: **Markov model recognition and classification of DNA/protein sequences within large text databases.** *Bioinformatics* 2005, **21:**4046-4053.
26. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome Res* 2003, **13:**308-312.
27. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: ***In vivo* enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:**499-502.
28. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4:**651-657.
29. Pierstorff N, Bergman CM, Wiehe T: **Identifying *cis*-regulatory modules by combining comparative and compositional analysis of DNA.** *Bioinformatics* 2006, **22:**2858-2864.
30. Lis JT, Prestidge L, Hogness DS: **A novel arrangement of tandemly repeated genes at a major heat shock site in *D. melanogaster*.** *Cell* 1978, **14:**901-919.
31. Livak KJ, Freund R, Schweber M, Wensink PC, Meselson M: **Sequence organization and transcription at two heat shock loci in *Drosophila*.** *Proc Natl Acad Sci USA* 1978, **75:**5613-5617.
32. Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S11.
33. Leser U, Hakenberg J: **What makes a gene name? Named entity recognition in the biomedical literature.** *Brief Bioinform* 2005, **6:**357-369.
34. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28:**21-28.
35. Pelham HR: **A regulatory upstream promoter element in the *Drosophila hsp 70* heat-shock gene.** *Cell* 1982, **30:**517-528.
36. Gilmour DS, Dietz TJ, Elgin SC: **UV cross-linking identifies four polypeptides that require the TATA box to bind to the *Drosophila hsp70* promoter.** *Mol Cell Biol* 1990, **10:**4233-4238.
37. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-144.
38. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431:**99-104.
39. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.
40. Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R: **PepBank - a database of peptides based on sequence text mining and public peptide data sources.** *BMC Bioinformatics* 2007, **8:**280.
41. **The GNU Collaborative International Dictionary of English** [http://www.ibiblio.org/webster/]
42. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet* 2003, **19:**68-72.
43. Adryan B, Teichmann SA: **FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*.** *Bioinformatics* 2006, **22:**1532-1533.
44. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJ: **A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks.** *Genome Biol* 2005, **6:**R110.
45. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2007, **35(Database issue):**D21-D25.
46. **Entrez Programming Utilities** [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html]
47. **pdftotext** [http://www.foolabs.com/xpdf/]
48. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ: **The UCSC genome browser database: update 2007.** *Nucleic Acids Res* 2007, **35(Database issue):**D668-D673.
49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
50. **Kent Source Tree** [http://genome.ucsc.edu/google/admin/cvs.html]
51. **ORegAnno Wiki** [http://www.bcgsc.ca/wiki/display/oreganno/DataFiles]
52. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, *et al.*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35(Database issue):**D610-D617.

2.3 Finding homologous cis-regulatory sequences by using genes as anchors

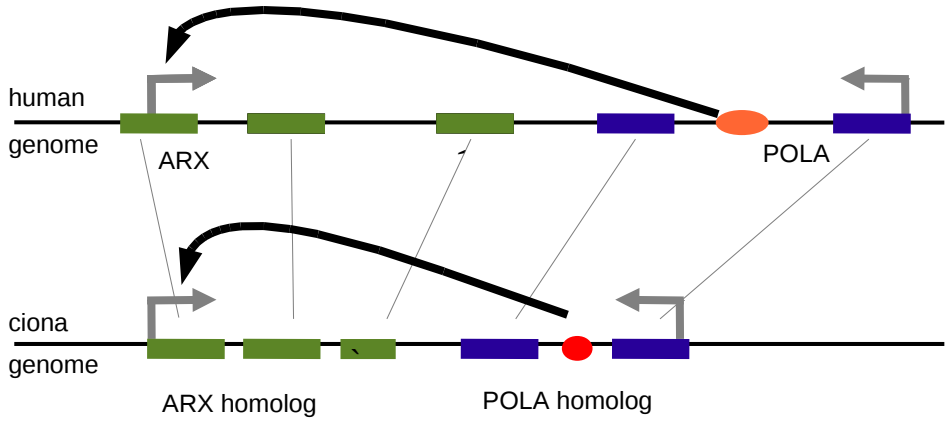


Figure 10: An example for two genes which are neighboring in both the human and the ciona genome: Marked by grey lines, the exons are alignable. Tom Becker (SARS Insitute, Norway) presented this example in our lab and hypothesized that non-alignable cis-regulatory sequences are located in the intron of one gene (not drawn to scale, distances would have to be reduced by factor 30 in the *C. intestinalis* genome)

As described in chapter 1, none of the thousands of CNEs in vertebrates are alignable to ascidians (Bejerano2004). Independently, some vertebrate enhancers have been shown to be located within an intron of a neighboring gene, probably leading to evolutionary pressure to keep this structure intact (see chapter 1). If this is the case in invertebrate species as well, then one should be able to find conserved neighboring orthologs (See Figure 10). One can then test the ascidian introns individually. The results would limit the vertebrate tests to only these introns.

It is not obvious to identify the essential sites in a vertebrate element of which many span more than 200bp. But two non-alignable ascidian/vertebrate enhancer sequences with a similar expression pattern could be more informative than mammalian/fish alignments, as the invertebrate sequences are expected to share just a handful of motifs. Currently the only example where such an enhancer comparison has been applied, to my knowledge, is described by (Yoshida2008). The authors identified the Pitx intronic left-sided enhancer in *C. intestinalis* and found that it contains FoxH and Nkx binding sites that reduce the activity of the element when mutated or when

the transcription factors are knocked-out. This corresponds to results in vertebrates (Shiratori2006).

Nevertheless, other reasons can keep genes very close together. For example, the two proteins that synthesize and transport acetylcholine (choline acetyltransferase, ChAT, and the vesicular acetylcholine transporter, VACht) share their first exon and are located on the same strand. The VACht gene is completely contained within the first intron of ChAT in flies, nematodes and in mammals and the two coding sequences are separated mostly at the splicing stage. One explanation for this tight linkage is that synchronized coexpression of both genes is necessary for acetylcholine usage and therefore the structure is conserved in all animals that were studied (Schutz2004). Obviously, this does not account for the fact that other neurotransmitters are produced by enzymes with a different configuration.

How many genes have kept flanking or overlapping orthologs in human/non-vertebrate comparisons? The *C. intestinalis* genome paper (Dehal2002) reported only exemplary cases, 16 genes located on the same chromsome, as in vertebrates, but many at mega-base pair distances. (Danchin2003) found one syntenic region between flies and human, but separated by other intervening genes and (Wang2007#349) detect a co-linear stretch between amphioxus and human in the PAX1/9 locus. Genomes of amphioxus (Putnam2008) and hydra (Putnam2007) have been analysed to find regions of synteny. But the aim of this work was rather the elucidation of the original ancestral gene order, not the reason for the unusual linkage. With the exception of (Wang2007#349) (who propose embedded enhancers), the functional reasons of conserved gene order were rarely evaluated.

As it was impossible to find a list of genes that were kept syntenic during chordate evolution and long-range synteny could be due to random rearrangements, I concentrated on pairs of genes that are located very close in the vertebrate, *C. intestinalis* and Amphioxus genomes. To determine human/vertebrate orthology relationships, data from Ensembl Compara Release 45 was used. For human/ciona and human/amphioxus, the best four BLAST matches were kept as “preliminary orthologs”, with the synteny information deciding on the real ortholog. The number four is the most common number of human homologs of a given ascidian gene (Dehal2002), probably a result of two rounds of whole genome duplication in the verte-

brate lineage (Dehal2005). The program is simply iterating over all human genes, checking for each if they are directly adjacent in the *C. intestinalis* genome. If this is the case, the gene pair is output, together with all other Ensembl genomes where they are adjacent. This results in a list of 117 pairs of genes. As some of them are part of two pairs, they sum up to only 208 individual genes. This number is much lower than the respective count in a human/amphioxus comparison (data not shown), again confirming the relatively stable genome architecture of cephalochordates compared to ascidians, as described previously (Putnam2008).

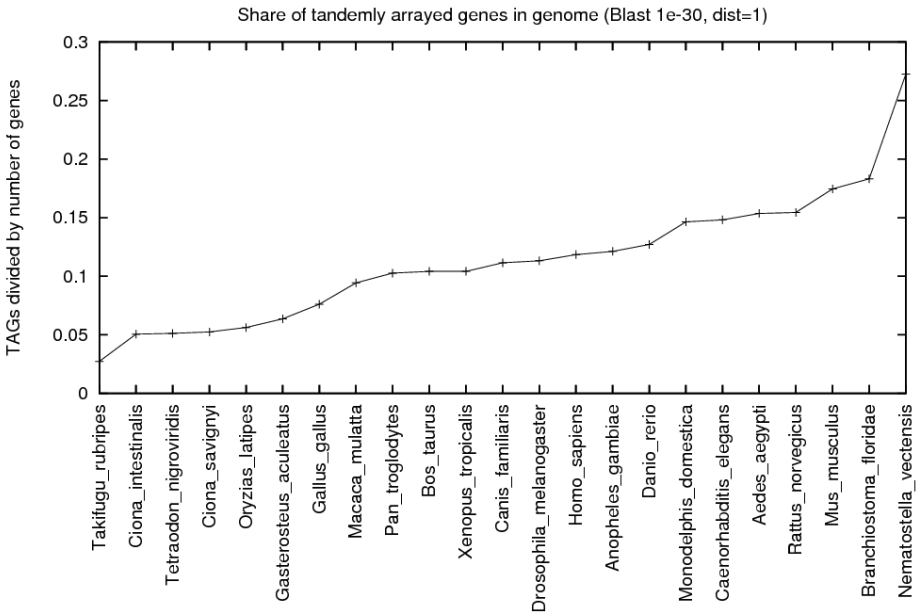


Figure 11 The relative share of tandem arrayed genes in 23 genomes (The *N. vectensis* genomes has been added only for comparative purposes and is not part of other analyses. Its high rate of tandem duplicates might be an artefact: Either due to split genes (N. vectensis is a genome with a very low number of ESTs, around 100k) or to duplication problems in the assembly process (*N. vectensis* contains many ancient repeated regions not found in other animals)

The most striking feature of this list of pairs conserved in human/*C. intestinalis* is that many of them, around 50%, are tandem duplicates. As the percentage of tandemly arranged genes in all our genomes is between 7 and 28 % (Figure 11), there are two possible reasons for the over-representation of tandemly arrayed genes: either tandem duplicates are less likely to break apart than other genes or they are easily re-created. The second possibility is much more likely, as tandem duplicates are among the most common DNA changes: a human gene has a probability of around 0.001 to be duplicated in

the primate lineage per million years (Pan2007), with an average life-span of 4 million years (Lynch2000). Moreover, most recent human genomic changes are short duplications <200bp (Messer2007) and re-sequencing studies suggest that a substantial fraction of genetic differences in the human population consists of tandem duplicates (Bailey2008). We therefore assume that syntenic tandemly arrayed genes will often be due to independent duplications. As it is very difficult to determine which of them are recent and which ancestral, we remove all tandemly arrayed genes from the following analyses (examples include paralogs of ACOT, CYP26A, HOXA and DLX. The exact orthology of the two ascidian DLX paralogs could indeed not be determined with different phylogenetic algorithms in (Irvine2007)).

This leaves 50 pairs composed of 98 genes. Of this list, we remove all genes that include at least one gene that has not an assigned official gene symbol, as it will be difficult to find information on these with a literature search, which reduced the list to 36 pairs (72 genes), shown in Figure 12. They are very likely functional and have not been re-created in the ascidian lineage, as all of them are arranged in the same manner in at least 8 other genomes. For instance, ADAMTS20 and PUS7L are adjacent in the genomes of the lancelet, dog, *C. intestinalis*, chicken, macaque, mouse, rat and chimp.

Surprisingly, I could not find any obvious common feature of genes in this list or any over-represented gene ontology category, except “sequence-specific DNA binding”, but with a relatively high p-Value of 0.001. As DNA-binding proteins are already enriched in human/ascidian homologs and the gene set rather small, I cannot derive any hypothesis from this.

Gene1	Gene2	Conserved in non-vertebrates	Gene1	Gene2	Conserved non-vertebrate
ADAMTS20	PUS7L	Amphioxus	KIAA0367	PRUNE2	Amphioxus
ARMC2	SESN1	No	**LRBA**	**MAB21L2**	**No**
BRF1	BTBD6	Insects	MAFA	ZC3H3	No
C16orf80	CSNK2A2	No	**NBEA**	**MAB21L1**	**Amphioxus**
C4orf22	*BMP3*	*Sea urchin*	PHF21B	NUP50	Amphioxus
CD226	RTTN	Insects	*POU4F2*	*TTC29*	*No*
CHAT	**SLC18A3**	**Insects**	PSMD1	HTR2B	No
CYB561D2	TMEM115	No	SAR1B	SEC24A	Amphioxus
CYR61	C1orf181	No	SLC9A3R2	NTHL1	No
DCTN6	RBPMS	Sea urchin	STYXL1	MDH2	Sea urchin
EFHA1	*FGF9*	Sea urchin	TAF4	LSM14B	No
EHBP1	*OTX1*	Sea urchin	**TMEM142A**	**MORN3**	**No**
FAM38B	C18orf30	Amhioxus	TMUB1	CENTG3	No
FBXW4	**FGF8**	**No**	WDR34	SET	No
FGF20	*EFHA2*	*Sea urchin*	WFS1	PPP2R2C	No
HHAT	KCNH1	No	*WNT5A*	*ERC2*	*Insects*
ISL2	**ZNF291**	**Insects**	**DUOX2**	**DUOXA2**	**Insects**

Figure 12: List of genes neighboring in the genome of Homo sapiens and *Ciona intestinalis* (Genome assemblies as in Ensembl 45). Pairs with experimental data or literature that propose a functional relationship between both genes are highlighted in bold. Pairs containing developmental regulators, the best candidates for further study, are highlighted in italic.

The list indeed includes the genes mentioned above ChAT/VACht (SLC18A3). It also shows up a the recently described “eukarotic operon equivalent” (Grasberger2006) DUOX2/DUOXA2, consisting of an enzyme iodinizing thyroid hormone and its maturation- or possibly more general cofactor (Morand2009).

One gene pair that stood out is ORAI1 (CRACM1) and MORN3 due to their conserved divergent orientation. It seemed that they could be co-regulated by a shared promoter. ORAI1/CRACM1 is the pore of a calcium release-activated calcium channel, regulating the flux of calcium across the endoplasmatic reticulum which is triggered during the immune response of T-cells. The role of the ORAI1 in calcium influx has been described recently (Luik2008), but the identification of additional related genes remains an active field of research (Vig2007).

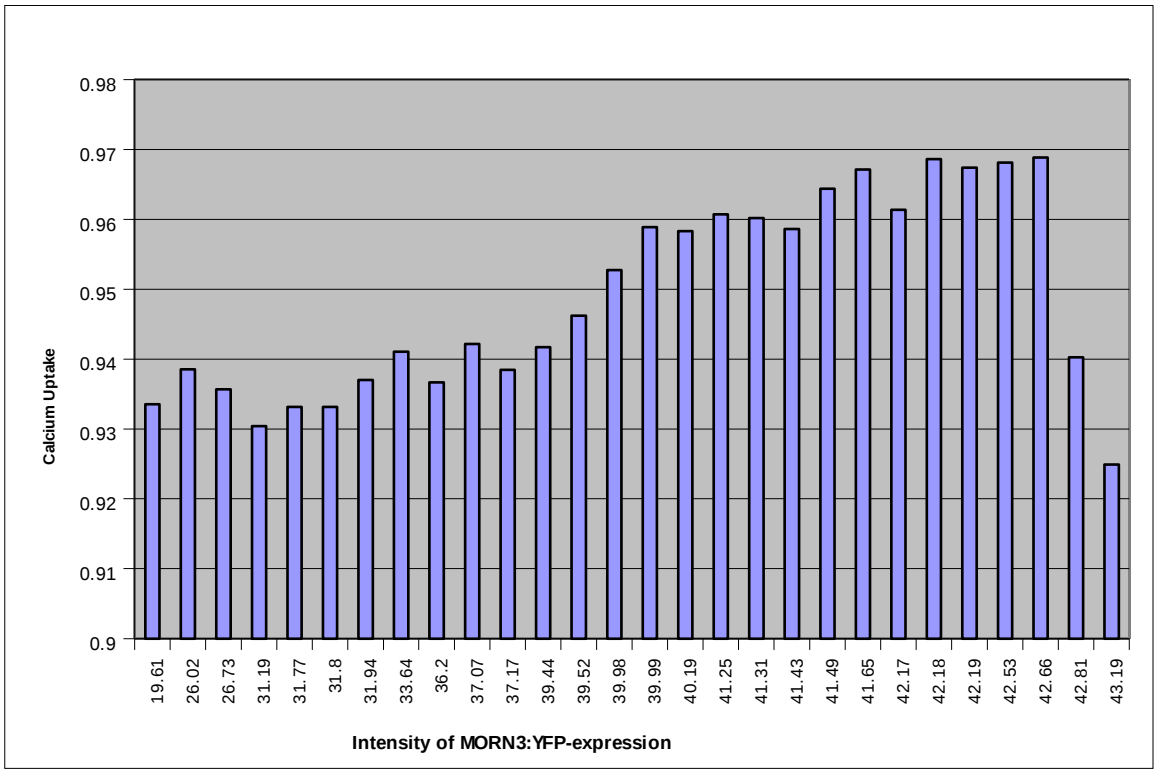


Figure 13: Preliminary data on the effect of MORN3 overexpression, most likely mediated by ORAI1, on calcium uptake, by Jun Liou, Stanford University: Experimental procedures are as in (Liou2005): HeLa-Cell cultures were treated with Fura-2 and calcium-content measured by fluorescence (310/340nm). Different quantities of MORN3-YFP plasmids were transfected and store-operated calcium entry was induced by thapsigargin: YFP intensity and calcium content were then measured. As can be seen, calcium uptake of cells is increasing when over-expressing Morn3. The drop for very high doses of Morn3 is expected and due to the toxicity of extreme quantities of YFP.

The neighbor of ORAI1 is MORN3. It is a protein without annotated function but contains the MORN-domain. This domain can be found in 16 human proteins, most of them without any functional annotation as well. Four of these have been named *Junctophilins* and play an essential role in junctional complexes between the plasma membrane and the endoplasmatic reticulum. (Wu2006) showed with electron microscopy that the calcium-activated calcium channel contacts the plasma membrane but could not explain this, as ORAI1 did not seem to establish this contact. Based on the divergent orientation and the strong conservation of both genes, one could hypothesize that MORN3 might play a role in calcium influx by establishing contacts with the endoplasmatic reticulum. I contacted the author who had previously identified the calcium sensor (Liou2005) with this idea. The prelimi-

nary results are very encouraging (Figure 13). She will continue to evaluate the role of MORN3 by confocal microscopy.

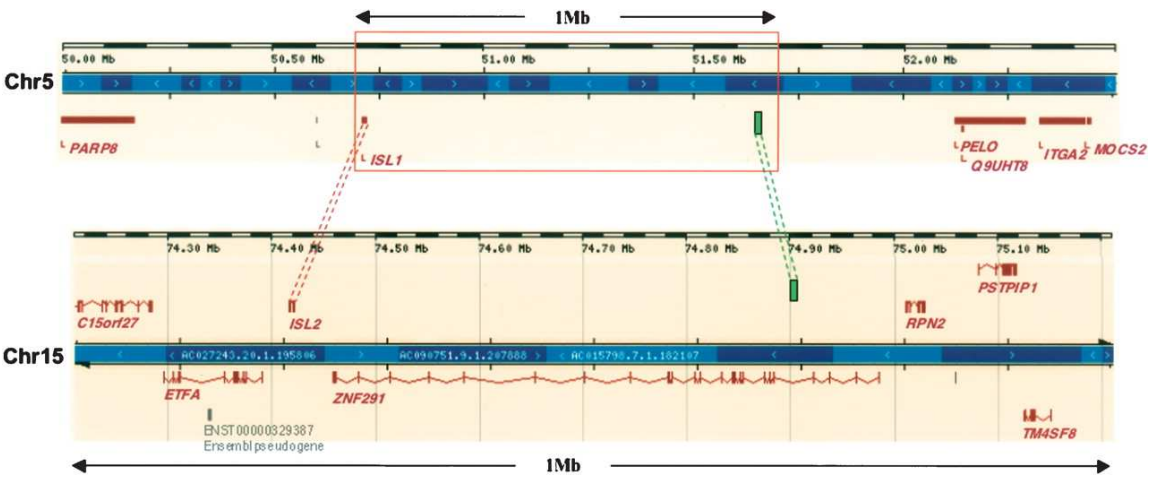


Figure 14: Location of a duplicated CNE in the vicinity of homeobox paralogs ISL1 (Chr5) and ISL2 (Chr15). The conserved non-coding block on the right, one copy residing within ZNF291 and the second one upstream of ISL1, is very likely a developmental enhancer. ISL1/2 and ZNF291 are also neighboring in the C. intestinalis genome. Ensembl screenshot, copied from (Vavouri2006)

Apart from these interactions due to co-expression of interacting enzymes, I also found pairs that had been previously described to be linked by cis-regulatory elements or are part of ongoing research. FGF8/FBXW4 have been described in the context of an enhancer trap in zebrafish (Kikuta2007). A GFP insertion within FBXW4 reflects the expression pattern of FGF8 and not FBXW4. Work in the laboratory of Francois Spitz (EMBL Heidelberg) showed that enhancer elements for FGF8 reside in introns of FBXW4 (personal communication). The list also contains the gene POLA, which harbors elements that clearly regulate ARX, as demonstrated by experiments on mice in the laboratory of Len Pennacchio, LBL Berkeley (personal communication). We also find the pair ZNF291/ISL1 which has been predicted to be related by long-range regulatory regions but has not been tested yet (see Figure 13). I have cloned this element from medaka DNA but was not able to inject it, as the transgenesis technique on Medaka is not completely set up yet in our lab.

Another example of overlapping genes are MAB21L1-NBEA and MAB21L2-LRBA. It is a duplicated ancestral gene pair, with MAB21L1/2 located in the first intron of NBEA/LRBA. Their conserved relative location has been described before ((Nikolaidis2007). Luckily, just a few days before finishing this text, duplicated non-coding conserved elements were pub-

lished (dCNEs, see page 43) that drive a reporter gene reminiscent of the MAB21-expression pattern (Tsang2009). These enhancers are spread over the introns of LRBA/NBEA. This shows that even nesting of genes can be explained by embedded enhancers

From the distribution of non-coding conserved sequences, one would rather expect developmental regulators like transcription factors or signaling molecules to be flanked by long-range enhancer embedded in introns of neighboring genes. Intriguingly, several other well-known developmental regulators are contained in the list: WNT5, BMP3, FGF8, FGF20, OTX1, POU4F2. Their neighbors are obvious candidates for future cross-species enhancer screens that are not limited to ascidians, but can be extended to flies and mammals.

Chapter 3: Discussion

> *For a biologist it is tempting to compare the evolution of ideas with the evolution of living nature. [...] ideas have kept some of the properties of organisms. Like them they want to propagate and multiply their structure, like them they can mix, recombine and reseparate their content, like them they have an evolution, and in this evolution, selection undoubtedly plays a big role.*
>
> *Jacques Monod, Chance and Necessity, 1971*

In the preceding chapters, I have approached the annotation of cis-regulatory sequences from three different angles, notably their prediction, annotation and cross-species homology detection. Here I want to highlight would could be improved and how recently published results could be incorporated.

3.1 Enhancer Prediction based on short sequence motifs

The widest field for additional work is the first part of the thesis, with almost endless room for changes and parameter modifications. One of them is the starting set of motifs that are scored: Whereas neurectoderm enhancers seem to require duplicated GATTA motifs, the experimental data also show that this motif is clearly not sufficient alone. Developing nervous system expression seems to be less “simple” than final differentiation into dopaminergic neuron or muscle cells where in both cases one single motif as short as 16bp directs specific GFP expression (Kusakabe2004) (Flames2009).

This corresponds to the observation that conserved non-coding sequences cluster around developmental regulators. It is easy to imagine that the expression of these genes follows stricter, more complex and redundant rules than terminal differentiation or environment-dependent reactions (e.g. heat-shock, infection): Stricter, as the mis-expression of regulators entails grave developmental defects. Complex, because various conditions (signaling molecule gradients, a certain developmental time point and cell lineage) have to be fulfilled to trigger the expression. Redundant, because in the presence of multiple signals, mutations can select randomly some for each gene and the overall redundancy of the motifs (not necessarily the whole different cis-regulatory blocks, as suggested by the “shadow enhancer” concept) increases the stability of essential developmental regulators.

In our case, we miss probably not one motif but rather different possible alternatives. At least, the current motif combination ranking algorithm suggests this, as any addition of a pentamer to 2xGATTA only lowers the motif-tissue scores. But what if the missing motif is not a pentamer? This is likely, as many transcription factors do not recognize pentamers but other types of sequences. One could easily extend the list of tested motif combinations to all tetramers and degenerate hexamers and also add all entries from Uniprobe or Transfac, leading to several thousand motifs. However, the running time is increasing exponentially with the number of motif combinations (millions) and the number of combinations of combinations (at least billions). The current implementation of the motif search has to move a sliding window over all non-coding positions to find matches, which is possible for 512

motifs but not sufficient for billions of possible models to test. A better indexing method is therefore required to speed up the search.

One possible strategy could involve breaking non-coding sequences into different numbered blocks that are assumed to represent putative enhancers. Then, an index could point for each motif to all numbers of blocks where the motif is conserved. A combination of motifs would then only require two lookups in this table, followed by the intersection of the resulting lists of numbers. As the two operations “lookup” and “intersection” are among the fastest of a microprocessor, a considerable gain in speed should be attainable. Research into this direction would better fit our model of the cell's transcription machinery than current motif discovery software. Since the early 90s, they still are searching today for the longest, best conserved and statistically rarest sets of motifs (Marschall2009), instead of – potentially very weak – binding sites that specifically match in combinations around the target gene set and not elsewhere.

It would be clearly preferable to integrate a thermodynamical model into the score, such that a fragment ranks higher with increasing the total number and affinity of the motifs in it, unlike a presence-absence decision as in our current pipeline. More and more implementation based on this concept are available but even the latest one (Roider2009) does not take into account cross-species conservation, distal enhancers or combinations of motifs. Therefore, significant improvements have still to be made before thermodynamical models can be used for an exhaustive search like ours.

An increased sensitivity in the motif finding step might find signals for other structures that are not related to the anterior neurectoderm. One of the deceptions of our approach is that it did not find any motif combination in muscle or notochord tissue. This might be due to a type of motif (degenerate hexamers) that did not fit our model or to a set of alternative motifs, which escaped our ranking based on individual motif duplicates.

Another critical point of our and other predictions is that we limit the search to conserved non-coding sequences. It is becoming increasingly evident that these are not distributed equally around all genes. Genes with longer upstream regions necessarily contain more conserved elements than others and upstream length distribution has been found to be biased in the

human genome, with categories like 'cell adhesion', 'nervous system' and 'transcription factors', among others, showing longer lengths (Taher2009). A better analysis of this phenomenon with in-situ databases as well as a ranking scheme that disentangles the mix and shows the contribution of this factor to the final score would be preferable.

On the wet-lab side, the bottleneck of cis-regulatory tests in ascidians are currently the cloning and maxipreparation steps. Replacing the Gateway system with a more standard cloning procedure, like restriction enzymes or ligation-independent cloning, can save a couple of days. If pure PCR-products instead of cloned plasmids can be injected in chicken (Hen2006), *C. elegans* (Boulin2006) and fish (Goode2005), they might also work in ascidians, requiring 10-15 PCR tubes per electroporation.

Looking beyond the ascidians, the most important extension of my work is less these algorithmic aspect as the currently weak link to vertebrates. The significance of the GATTA motif in previous forebrain enhancer predictions (Pennacchio2006) demonstrated that there is at least something detectable. Being able to obtain similar results from the mammalian genome alignments would result in a much higher impact of these results. While I have prepared vertebrate non-coding alignments, the difficulty was the annotation of in-situ data in the MGI (mouse) and Zfin (zebrafish) databases. One has to select a single stage and harmonize annotations to a common level of detail first. Currently, the databases contain some very few genes with hundreds of detailed annotations from publications (e.g. PAX6) and several thousand genes with rather low-quality images from large-scale screens, grouped into relatively large tissue classes,. The Eurexpress (http://eurexpress .org) database with ~14000 whole-mount in-situ mouse images, annotated with a standard ontology and data in BioMart and UCSC format, promises to make this much easier.

3.2 Enhancer annotation with text mining approaches

Any further algorithmic work on the prediction of tissue-specific sequences needs unbiased benchmark data sets. The simple sequence extraction scripts should be made available as a website. Then, annotators could paste a Pubmed-ID and only choose from a set of sequences instead of searching them in the paper, typing them into the BLAST form and waiting for the results. For this, it would be very helpful to see the text around the primer matches. Annotators of model databases like Zfin or MGI should be interested in such a system. This might ultimately help to improve the availability of benchmark data, at least on the sequence side.

It is tempting to extend such a system and try to infer automatically the annotation of the tissue of expression. One possible way to tackle this is the recognition of tissue names from the full text of the literature. The complex vocabulary to describe embryonic territories makes this challenging. One solution is the unpublished first version of the ontology MIAA ("Minimum Information about Anatomy")[3], which tries to harmonize and group the various terms into 400 tissues across several model tissues. Another ontology, Uberon[4], tries to be more specific (4000 terms) and links directly to tissue annotations from model databases like Zfin and homologous mouse tissue identifiers from MGI. These efforts open the way to a recognition of tissue terms in English text and their visualization with exemplary in-situ images automatically extracted from several model organism databases.

A simpler system might extract only the images itself from the publication. This is inspired by the habit of most readers to first look at the images of an article to decide if it is interesting for them. In addition, a software could display only images that actually show a section of an embryo, a relative standard classification task, e.g. based on color-space histograms (Faloutsos94). As such, coherent sets of scientific images can be constructed, like the protein blots in blotBase (Schlamp2008). One search engine that is based on an annotated selection of fulltext images from open-access articles is the Yale Image Finder (Xu2008): It outputs expression patterns for PITX2 at a mouse click, without having to open any fulltext article.

3 http://www.compbio.ox.ac.uk/data.shtml (Computational biology group, Oxford)
4 http://obofoundry.org/wiki/index.php/UBERON:Main_Page

A combination of such an image database with sequence extraction could lead to a genome browser that highlights non-coding sequence matches around genes and annotates them with the images from the article. This would save time when searching for a cis-regulatory sequence around a gene of interest.

3.3 Enhancer annotation in syntenic genes

The interpretation of my list of syntenic genes is relying on a literature search to show the importance of these individual examples. The data illustrates how the gene order itself can help to delineate the range of cis-regulatory action. The result could have not been found with a trivial data mining approach based on Gene ontology terms, published microarrays or by looking at hundreds of in-situ expression images. Finding support for a hypothesis in such a way is sometimes called “cherry picking” in bioinformatics, is not considered an important part of bioinformatics and neither of biological research which often has to be focused on a certain tissue of interest. The literature search nevertheless brought up interesting results matching very recent work in collaborating laboratories. To show a functional link for the closer gene pairs, contacting specialists will be the only way to validate them, as in the case of ORAI1/MORN3. The remaining cis-regulatory-related examples will only be verifiable by further cis-regulatory tests in the wet-lab.

My list has the the particular advantage that the functional link between the pairs in it can be researched in almost any animal model while still being mappable to the human genome. It contains a couple of well-known developmental regulators. Their further analysis should lead to the identification of several cross-species enhancers, at a homologous position in ascidians and vertebrates. It might also help to identify long-range enhancers in Ciona, where all cis-regulatory information, to my knowledge, has been searched and found in the immediate upstream flanking regions of genes until now.

Unlike the genetic code, with its 64 triplets coding for 20 amino acids, the cis-regulatory code contains a lot more information: All conditions to express or repress genes, their timing and their quantity in the hundreds of different known tissues and all precursor cells. Being able to “read” it would result in an immediate functional assignment for all proteins and non-coding RNAs in a genome. Given the low number of enhancers of which we have detailed knowledge (mostly interferon and beta-globin, since 20 years), this

goal is certainly very far and might be never completely achieved: Contrary to the genetic code which was decrypted in a cell-free system, regulatory instructions make only sense in the context of the development and evolution of a particular cell type. The complexity of the whole and our inability to directly observe transcription factors bound to the DNA leads to a lot of conflicting results but also contributes to its fascination.

Chapter 4: Appendix

This chapter gives an overview of the command line tools that were created during the last three years and a list of publications where the author participated

4.1 Command line tools

Like every computational biologist, I developed tools for the projects that I was working on. These are usually not appear in the text but still represent several hundred hours of work and might be useful for someone else. Following a UCSC browser convention, my tools start with the main filetype that they accept as input. They usually fulfill only one task per tool but can be chained with UNIX pipes as many support the keyword 'stdin' instead of a filename. Altogether, these sum up to roughly 20k lines of code (mostly Python but also Perl (Ensembl-API)and C (with the Jim Kent source library). Sources can be downloaded from:

http://genome.ciona.cnrs-gif.fr/max/max-tools.tar.gz

bedChain	join features into one if they are closer than x basepairs
bedChromDensity	plot feature density on chromosomes with UCSC genome graphs
bedcluster	output regions that contain at least x features within y basepairs (of a given type)
bedfastaextract	return the sequences of features from a fasta file
bedFilter	remove features based on their name or the number of their occurences
bedFindNeighbors	find the neighboring genes for genome features
bedGappedToUngapped	map feature positions from an unaligned sequence to the positions on the multiple alignment
bedLenDownstream	output the average length of upstream regions per gene type
bedLongestTranscript	keeps only the longest transcript for each gene (Ensembl genes)
bedNameRewrite	change the names of features according with regular expressions
bedNameTable	prepare feature names for import into R as tabsep-file
bedProject	map features coordinates from a fasta file to the genome
bedUpstream	given a genes (UCSC format), return their upstream region coordinates, (handles overlaps with various rules)
blastAndBed	run tblastx on two fasta files and convert results to UCSC format
blastBestMatches	keep only best BLAST hits from a set of files
blastOnAll	blast fasta against a directory and submit jobs to LSF or GridEngine clusters
blastPubmed	given a PubmedID, download the pdf of the article via CNRS/INRA fulltext accounts, extract the nucleotide text from them, blat it onto selected genomes using UCSC's blat servers, upload the resulting bed files into UCSC and show a link to the results
dataToTable	collect values from ini-style files (var1=5) and convert to a table for R
ensemblgff2bed	convert Ensembl's gff to UCSC, trying to add 'chr' if necessary
faFilterLongest	keep only the longest sequence for each fasta file (=longest transcript)
fasta2apollofa, fasta2gccontent, fasta2ic, fasta2mat, fasta2plot, fasta2vista	convert fasta files to the Apollo Editor format, get their GC content, information content, Transfac Matrix, nucleotide distribution and the old, original VISTA input (glass) format
fastaexplode	split fasta file into individual sequences
fastafilterseqname	keep only sequences whose ids are listed in a textfile
fastaFromUCSC	given genomic coordiates, get the DNA sequence from UCSC

fastaMotifOverrep	search a sequence for a motif and use 1000x shuffling to estimate its over-representation
featurePlotter	textmode prettyprinter: print multialignment and try to map all features from bed files onto it, even if coordinates are on the ungapped sequences
fastaSearch	search sequence for motifs and output in UCSC format
fastaWrap	format fasta for 80char width printing
hashBenchmark	Given a benchmark and prediction key-value file, calculate sensitivity, specificity, precision, recall and F1
hashesToTable	convert key-value files to R tables
hashFilter	filter a key-value file by keys, by key-counts or replace keys with others
hashIntersect	given to key-value files, display only their common keys and values
hashRankCompare	compare two ranks for values for two sorted key-value files
hashToArff	convert key-value files to ARFF (Weka input format)
jaspar2tf	convert jaspar to transfac format
logadd, logmenu, log	add the last command from the bash history to a file .log in the current directory, display a menu of these commands for execution, or display this logfile
lstGoAnalysis	given foreground and background gene lists, run a over-representation analysis using the GO_Func program, sort and filter output
lstIntersect	print the intersection of two textfiles
lstMyLoad	parse tabsep file, create a table for all column-headers with mysql and use LOAD INFILE to populate the table
lstOp	given two text files do: a) mass-replace given a key-value file (e.g. ids to gene names) or b) a mysql-like join on the two textfiles on selected columns or c) remove lines that match/don't match any line a second file
lstRandomizeAvg	read values from textfiles (one per line), calculate average, determine how probable >= average is if only a subset of values is used (jackknife)
maf2EvoPrinter, maf2faDir, maf2faFiles	UCSC multi alignment format: Imitate EvoPrinter display of multialignments, split multialignment over files of directories
mafScan	scan multiple alignment for conserved consensus motif matches, output a .word file
motifGenerator	generate a list of motifs with a given lengths a given number of degenerate positions
mudi	show motifs that are conserved in all of a certain number of alignments and appear in all of them
musca	(multi scan) search transfac matrix matches in aligment, report only conserved matches exceeding certain cutoff
oboAncester	try to find the parent nodes of an OBO-Ontology file for a given set of nodes
pmidToPdf.pl	similar to blastPubmed, but in PERL and without blatting, requires WWW:Mechanize
restrictUnique	find unique restriction sites in fasta files using Emboss
retrAniseed	download insitus from Aniseed (given insitu ids)
retrEnsembl	download any Ensembl table via Biomart (e.g. gene coordinates, homologies, protein alignments etc)
retrEnsemblGenomes	download all Ensembl genomes or genes from a certain version
retrEnsOrthoSeqs	API-based version of retrEnsembl for orthologous sequences, retrEnsembl is usually much faster
retrPubmed	given a list of term, show number of matching pubmed records by year or by term, download these abstract or download the associated nucleotide sequences for them
retrZfin	download insitus images from Zfin given gene ids
t2g_*	the different steps of the text2Genome pipeline, including xml parsing, nucleotide extraction, blasting on GridEngine Clusters, filtering of blast matches and displaying them via a DAS-server on the Ensembl genomes
word*	various tools to filter, sort, index .word files for the Cionator-pipeline

4.2 Other Publications

Auger H, Lamy C, Haeussler M, Khoueiry P, Lemaire P, Joly JS.
Similar regulatory logic in Ciona intestinalis for two Wnt pathway modulators, ROR and SFRP-1/5
Dev Biol. 2009 May 15;329(2):364-73. Epub 2009 Feb 25.

Contribution: Application of the motif-search software.

Result: The role of the factor FOXA in the determination of the a-lineage is reflected by a biased distribution of FOXA-motifs in cis-regulatory regions flanking genes expressed in this lineage at the 110-cell stage

Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, et al; Open Regulatory Annotation Consortium
ORegAnno: an open-access community-driven resource for regulatory annotation.
Nucl Acid Res 2008 Jan;36:D107-13.

Contribution: An import script for the VISTA enhancer browser

Result: ~500 enhancers and their sequences are automatically imported into the Oreganno database and updated with every new release

Jaszczyszyn Y, Haeussler M, Heuzé A, Debiais-Thibaud M, Casane D, Bourrat F, Joly JS. *Comparison of the expression of medaka (Oryzias latipes) pitx genes with other vertebrates shows high conservation and a case of functional shuffling in the pituitary*. Gene 2007 Dec 30;406(1-2):42-50

Contribution: A figure illustrating the flanking genes and homology relationships (synteny) in the PITX2 locus in different vertebrates

Result: There is not doubt about the phylogenetic relationships of the different PITX paralogs

References

Adryan, B., Woerfel, G., Birch-Machin, I. et al. (2007). Genomic mapping of suppressor of hairy-wing binding sites in drosophila. *Genome Biol.* **8**: R167.

Ahituv, N., Akiyama, J., Chapman-Helleboid, A. et al. (2007). In vivo characterization of human apoa5 haplotypes. *Genomics* **90**: 674-679.

Ahituv, N., Prabhakar, S., Poulin, F. et al. (2005). Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Hum. Mol. Genet.* **14**: 3057-3063.

Ahituv, N., Zhu, Y., Visel, A. et al. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**: e234.

Akbari, OS., Bae, E., Johnsen, H. et al. (2008). A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the drosophila embryo. *Development* **135**: 123-131.

Akbari, OS., Bousum, A., Bae, E. et al. (2006). Unraveling cis-regulatory mechanisms at the abdominal-a and abdominal-b genes in the drosophila bithorax complex. *Dev. Biol.* **293**: 294-304.

Amaral, PP. and Mattick, JS. (2008). Noncoding rna in development. *Mamm. Genome* **19**: 454-492.

Aparicio, S., Morrison, A., Gould, A. et al. (1995). Detecting conserved regulatory elements with the model genome of the japanese puffer fish, fugu rubripes. *Proc. Natl. Acad. Sci. U.S.A.* **92**: 1684-1688.

Arányi, T., Faucheux, BA., Khalfallah, O. et al. (2005). The tissue-specific methylation of the human tyrosine hydroxylase gene reveals new regulatory elements in the first exon. *J. Neurochem.* **94**: 129-139.

Armit, C. (2007). Developmental biology and databases: how to archive, find and query gene expression patterns using the world wide web. *Organogenesis* **3**: 70-73.

Arnosti, DN. (2003). Analysis and function of transcriptional regulatory elements: insights from drosophila. *Annu. Rev. Entomol.* **48**: 579-602.

Attanasio, C., Reymond, A., Humbert, R. et al. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol.* **9**: R168.

Azumi, K., Sabau, SV., Fujie, M. et al. (2007). Gene expression profile during the life cycle of the urochordate ciona intestinalis. *Dev. Biol.* **308**: 572-582.

Bailey, JA., Kidd, JM. and Eichler, EE. (2008). Human copy number polymorphic genes. *Cytogenet. Genome Res.* **123**: 234-243.

Banerji, J., Rusconi, S. and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote sv40 dna sequences. *Cell* **27**: 299-308.

Barthel, KKB. and Liu, X. (2008). A transcriptional enhancer from the coding region of adamts5. *PLoS ONE* **3**: e2184.

Barton, LM., Gottgens, B., Gering, M. et al. (2001). Regulation of the stem cell leukemia (scl) gene: a tale of two fishes. *Proc. Natl. Acad. Sci. U.S.A.* **98**: 6747-6752.

Bejerano, G., Lowe, CB., Ahituv, N. et al. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87-90.

Bejerano, G., Pheasant, M., Makunin, I. et al. (2004). Ultraconserved elements in the human genome. *Science* **304**: 1321-1325.

Berman, BP., Nibu, Y., Pfeiffer, BD. et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 757-762.

Berman, BP., Pfeiffer, BD., Laverty, TR. et al. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in drosophila melanogaster and drosophila pseudoobscura. *Genome Biol.* **5**: R61.

Birney, E., Stamatoyannopoulos, JA., Dutta, A. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447**: 799-816.

Blanchette, M., Bataille, AR., Chen, X. et al. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**: 656-668.

Boffelli, D., Weer, CV., Weng, L. et al. (2004). Intraspecies sequence comparisons for annotating genomes. *Genome Res.* **14**: 2406-2411.

Bouchard, M., Grote, D., Craven, SE. et al. (2005). Identification of pax2-regulated genes by expression profiling of the mid-hindbrain organizer region. *Development* **132**: 2633-2643.

Boulin, T., Etchberger, JF. and Hobert, O. (2006). Reporter gene fusions. *WormBook* : 1-23.

Bruce, AW., López-Contreras, AJ., Flicek, P. et al. (2009). Functional diversity for rest (nrsf) is defined by in vivo binding affinity hierarchies at the dna sequence level. *Genome Res.* : .

Brudno, M., Malde, S., Poliakov, A. et al. (2003). Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**: i54-62.

Butler, JE. and Kadonaga, JT. (2001). Enhancer-promoter specificity mediated by dpe or tata core promoter motifs. *Genes Dev.* **15**: 2515-2519.

Calhoun, VC., Stathopoulos, A. and Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the drosophila antennapedia complex. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 9243-9247.

Cameron, RA., Chow, SH., Berney, K. et al. (2005). An evolutionary constraint: strongly disfavored class of change in dna sequence during divergence of cis-regulatory modules. *Proc. Natl. Acad. Sci. U.S.A.* **102**: 11769-11774.

Cameron, RA., Oliveri, P., Wyllie, J. et al. (2004). Cis-regulatory activity of randomly chosen genomic fragments from the sea urchin. *Gene Expr. Patterns* **4**: 205-213.

Caputi, L., Andreakis, N., Mastrototaro, F. et al. (2007). Cryptic speciation in a model invertebrate chordate. *Proc. Natl. Acad. Sci. U.S.A.* **104**: 9364-9369.

Carter, D., Chakalova, L., Osborne, CS. et al. (2002). Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* **32**: 623-626.

Casillas, S., Barbadilla, A. and Bergman, CM. (2007). Purifying selection maintains highly conserved noncoding sequences in drosophila. *Mol. Biol. Evol.* **24**: 2222-2234.

Cerda, GA., Hargrave, M. and Lewis, KE. (2009). Rna profiling of fac-sorted neurons from the developing zebrafish spinal cord. *Dev. Dyn.* **238**: 150-161.

[Chabry1887] Chabry, L. Embryologie et teratologique des ascidies. La Faculte des Sciences de Paris. 1887.

Chen, H. and Blanchette, M. (2007). Detecting non-coding selective pressure in coding regions. *BMC Evol. Biol.* **7 Suppl 1**: S9.

Chen, J., Huisinga, KL., Viering, MM. et al. (2002). Enhancer action in trans is permitted throughout the drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 3723-3728.

Chen, JCJ. and Goldhamer, DJ. (2004). The core enhancer is essential for proper timing of myod activation in limb buds and branchial arches. *Dev. Biol.* **265**: 502-512.

Cheng, Y., King, DC., Dore, LC. et al. (2008). Transcriptional enhancement by gata1-occupied dna segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res.* **18**: 1896-1905.

Chi, X., Chatterjee, PK., Wilson, W3. et al. (2005). Complex cardiac nkx2-5 gene expression activated by noggin-sensitive enhancers followed by chamber-specific modules. *Proc. Natl. Acad. Sci. U.S.A.* **102**: 13490-13495.

Chiba, S., Sasaki, A., Nakayama, A. et al. (2004). Development of ciona intestinalis juveniles (through 2nd ascidian stage). *Zool. Sci.* **21**: 285-298.

Christiaen, L., Bourrat, F. and Joly, J. (2005). A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. *Dev. Biol.* **277**: 557-566.

Christiaen, L., Burighel, P., Smith, WC. et al. (2002). Pitx genes in tunicates provide new molecular insight into the evolutionary origin of pituitary. *Gene* **287**: 107-113.

Christiaen, L., Davidson, B., Kawashima, T. et al. (2008). The transcription/migration interface in heart precursors of ciona intestinalis. *Science* **320**: 1349-1352.

Christiaen, L., Stolfi, A., Davidson, B. et al. (2009). Spatio-temporal intersection of lhx3 and tbx6 defines the cardiac field through synergistic activation of mesp. *Dev. Biol.* **328**: 552-560.

Condie, BG. and Capecchi, MR. (1994). Mice with targeted disruptions in the paralogous genes hoxa-3 and hoxd-3 reveal synergistic interactions. *Nature* **370**: 304-307.

Conradt, B. and Horvitz, HR. (1999). The tra-1a sex determination protein of c. elegans regulates sexually dimorphic cell deaths by repressing the egl-1 cell death activator gene. *Cell* **98**: 317-327.

Conte, I. and Bovolenta, P. (2007). Comprehensive characterization of the cis-regulatory code responsible for the spatio-temporal expression of olsix3.2 in the developing medaka forebrain. *Genome Biol.* **8**: R137.

Corbo, JC., Di Gregorio, A. and Levine, M. (2001). The ascidian as a model organism in developmental and evolutionary biology. *Cell* **106**: 535-538.

Corbo, JC., Erives, A., Di Gregorio, A. et al. (1997). Dorsoventral patterning of the vertebrate neural tube is conserved in a protochordate. *Development* **124**: 2335-2344.

Dahl, JA. and Collas, P. (2008). Microchip--a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res.* **36**: e15.

Danchin, EGJ., Abi-Rached, L., Gilles, A. et al. (2003). Conservation of the mhc-like region throughout evolution. *Immunogenetics* **55**: 141-148.

Dathe, K., Kjaer, KW., Brehm, A. et al. (2009). Duplications involving a conserved regulatory element downstream of bmp2 are associated with brachydactyly type a2. *Am. J. Hum. Genet.* **84**: 483-492.

de la Calle-Mustienes, E., Feijóo, CG., Manzanares, M. et al. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate iroquois cluster gene deserts. *Genome Res.* **15**: 1061-1072.

De Val, S., Chi, NC., Meadows, SM. et al. (2008). Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell* **135**: 1053-1064.

Dehal, P. and Boore, JL. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**: e314.

Dehal, P., Satou, Y., Campbell, RK. et al. (2002). The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298**: 2157-2167.

Dekker, J., Rippe, K., Dekker, M. et al. (2002). Capturing chromosome conformation. *Science* **295**: 1306-1311.

Delsuc, F., Brinkmann, H., Chourrout, D. et al. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**: 965-968.

Dermitzakis, ET., Reymond, A., Scamuffa, N. et al. (2003). Evolutionary discrimination of mammalian conserved non-genic sequences (cngs). *Science* **302**: 1033-1035.

Dong, X., Fredman, D. and Lenhard, B. (2009). Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol.* **10**: R86.

Donmez, N., Bazykin, GA., Brudno, M. et al. (2009). Polymorphism due to multiple amino acid substitutions at a codon site within ciona savignyi. *Genetics* **181**: 685-690.

Dorsett, D. (1999). Distant liaisons: long-range enhancer-promoter interactions in drosophila. *Curr. Opin. Genet. Dev.* **9**: 505-514.

Dorsett, D. (1993). Distance-independent inactivation of an enhancer by the suppressor of hairy-wing dna-binding protein of drosophila. *Genetics* **134**: 1135-1144.

Dostie, J., Richmond, TA., Arnaout, RA. et al. (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**: 1299-1309.

Drake, JA., Bird, C., Nemesh, J. et al. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**: 223-227.

Dubchak, I., Brudno, M., Loots, GG. et al. (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**: 1304-1306.

Ejsmont, RK., Sarov, M., Winkler, S. et al. (2009). A toolkit for high-throughput, cross-species gene engineering in drosophila. *Nat. Methods* : .

Elnitski, L., Jin, VX., Farnham, PJ. et al. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.* **16**: 1455-1464.

Engström, PG., Ho Sui, SJ., Drivenes, O. et al. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**: 1898-1908.

Ertzer, R., Müller, F., Hadzhiev, Y. et al. (2007). Cooperation of sonic hedgehog enhancers in midline expression. *Dev. Biol.* **301**: 578-589.

Faloutsos, C., Equitz, W., Flickner, M. et al. (1994). Efficient and effective querying by image content. *Journal of Intelligent Information Systems* **3**: 231-262.

Farhadi, HF., Lepage, P., Forghani, R. et al. (2003). A combinatorial network of evolutionarily conserved myelin basic protein regulatory sequences confers distinct glial-specific phenotypes. *J. Neurosci.* **23**: 10214-10223.

Farré, D., Bellora, N., Mularoni, L. et al. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* **8**: R140.

Feng, J., Bi, C., Clark, BS. et al. (2006). The evf-2 noncoding rna is transcribed from the dlx-5/6 ultraconserved region and functions as a dlx-2 transcriptional coactivator. *Genes Dev.* **20**: 1470-1484.

Fisher, S., Grice, EA., Vinton, RM. et al. (2006). Conservation of ret regulatory function from human to zebrafish without sequence similarity. *Science* **312**: 276-279.

Flames, N. and Hobert, O. (2009). Gene regulatory logic of dopamine neuron differentiation. *Nature* **458**: 885-889.

Fondrat, C. and Kalogeropoulos, A. (1994). Approaching the function of new genes by detection of their potential upstream activation sequences in saccharomyces cerevisiae: application to chromosome iii. *Curr. Genet.* **25**: 396-406.

Frith, MC., Valen, E., Krogh, A. et al. (2008). A code for transcription initiation in mammalian genomes. *Genome Res.* **18**: 1-12.

Gaudet, J. and Mango, SE. (2002). Regulation of organogenesis by the caenorhabditis elegans foxa protein pha-4. *Science* **295**: 821-825.

Gertz, J., Siggia, ED. and Cohen, BA. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**: 215-218.

Geschwind, D. (2004). Gensat: a genomic resource for neuroscience research. *Lancet Neurol* **3**: 82.

Ghanem, N., Jarinova, O., Amores, A. et al. (2003). Regulatory roles of conserved intergenic domains in vertebrate dlx bigene clusters. *Genome Res.* **13**: 533-543.

Ghanem, N., Yu, M., Long, J. et al. (2007). Distinct cis-regulatory elements from the dlx1/dlx2 locus mark different progenitor cell populations in the ganglionic eminences and different subtypes of adult cortical interneurons. *J. Neurosci.* **27**: 5012-5022.

Goode, DK., Snell, P., Smith, SF. et al. (2005). Highly conserved regulatory elements around the shh gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86**: 172-181.

Gorman, J. and Greene, EC. (2008). Visualizing one-dimensional diffusion of proteins along dna. *Nat. Struct. Mol. Biol.* **15**: 768-774.

Göttgens, B., Barton, LM., Gilbert, JG. et al. (2000). Analysis of vertebrate scl loci identifies conserved enhancers. *Nat. Biotechnol.* **18**: 181-186.

Grasberger, H. and Refetoff, S. (2006). Identification of the maturation factor for dual oxidase. evolution of an eukaryotic operon equivalent. *J. Biol. Chem.* **281**: 18269-18272.

Griffith, OL., Montgomery, SB., Bernier, B. et al. (2008). Oreganno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**: D107-13.

Gross, DS. and Garrard, WT. (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**: 159-197.

Guo, G., Bauer, S., Hecht, J. et al. (2008). A short ultraconserved sequence drives transcription from an alternate fbn1 promoter. *Int. J. Biochem. Cell Biol.* **40**: 638-650.

Guyot, B., Valverde-Garduno, V., Porcher, C. et al. (2004). Deletion of the major gata1 enhancer hs 1 does not affect eosinophil gata1 expression and eosinophil differentiation. *Blood* **104**: 89-91.

Halfon, MS., Gallo, SM. and Bergman, CM. (2008). Redfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in drosophila. *Nucleic Acids Res.* **36**: D594-8.

Hallikas, O., Palin, K., Sinjushina, N. et al. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**: 47-59.

Harbison, CT., Gordon, DB., Lee, TI. et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**: 99-104.

Hardison, R. and Miller, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* **10**: 73-102.

Hardison, RC., Oeltjen, J. and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7**: 959-966.

Hare, EE., Peterson, BK. and Eisen, MB. (2008). A careful look at binding site reorganization in the even-skipped enhancers of drosophila and sepsids. *PLoS Genet.* **4**: e1000268.

Hare, EE., Peterson, BK., Iyer, VN. et al. (2008). Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet.* **4**: e1000106.

Harismendy, O., Ng, PC., Strausberg, RL. et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**: R32.

Heintzman, ND., Hon, GC., Hawkins, RD. et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* : .

Hen, G., Bor, A., Simchaev, V. et al. (2006). Expression of foreign genes in chicks by hydrodynamics-based naked plasmid transfer in vivo. *Domest. Anim. Endocrinol.* **30**: 135-143.

Hill, MM., Broman, KW., Stupka, E. et al. (2008). The c. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Res.* **18**: 1369-1379.

Hobert, O. (2002). Pcr fusion-based approach to create reporter gene constructs for expression analysis in transgenic c. elegans. *BioTechniques* **32**: 728-730.

Hobert, O. (2008). Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc. Natl. Acad. Sci. U.S.A.* **105**: 20067-20071.

Holland, LZ., Albalat, R., Azumi, K. et al. (2008). The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**: 1100-1111.

Hong, J., Hendrix, DA. and Levine, MS. (2008). Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.

Horan, GS., Kovàcs, EN., Behringer, RR. et al. (1995). Mutations in paralogous hox genes result in overlapping homeotic transformations of the axial skeleton: evidence for unique and redundant function. *Dev. Biol.* **169**: 359-372.

Hudson, C. and Lemaire, P. (2001). Induction of anterior neural fates in the ascidian ciona intestinalis. *Mech. Dev.* **100**: 189-203.

Hufton, AL., Mathia, S., Braun, H. et al. (2009). Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome Res.* : .

Hurst, LD., Pál, C. and Lercher, MJ. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**: 299-310.

Huxley-Jones, J., Robertson, DL. and Boot-Handford, RP. (2007). On the origins of the extracellular matrix in vertebrates. *Matrix Biol.* **26**: 2-11.

Iannelli, F., Pesole, G., Sordino, P. et al. (2007). Mitogenomics reveals two cryptic species in ciona intestinalis. *Trends Genet.* **23**: 419-422.

Irvine, SQ., Cangiano, MC., Millette, BJ. et al. (2007). Non-overlapping expression patterns of the clustered dll-a/b genes in the ascidian ciona intestinalis. *J. exp. zool. B. Mol. Dev. Evol.* **308**: 428-441.

Irvine, SQ., Fonseca, VC., Zompa, MA. et al. (2008). Cis-regulatory organization of the pax6 gene in the ascidian ciona intestinalis. *Dev. Biol.* **317**: 649-659.

Ishihara, T., Sato, S., Ikeda, K. et al. (2008). Multiple evolutionarily conserved enhancers control expression of eya1. *Dev. Dyn.* **237**: 3142-3156.

Iwahori, A., Fraidenraich, D. and Basilico, C. (2004). A conserved enhancer element that drives fgf4 gene expression in the embryonic myotomes is synergistically activated by gata and bhlh proteins. *Dev. Biol.* **270**: 525-537.

Izumi, K., Aramaki, M., Kimura, T. et al. (2007). Identification of a prosencephalic-specific enhancer of sall1: comparative genomic approach using the chick embryo. *Pediatr. Res.* **61**: 660-665.

Jack, J., Dorsett, D., Delotto, Y. et al. (1991). Expression of the cut locus in the drosophila wing margin is required for cell type specification and is regulated by a distant enhancer. *Development* **113**: 735-747.

Jaffe, DB., Butler, J., Gnerre, S. et al. (2003). Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Res.* **13**: 91-96.

Jiang, Y., Matevossian, A., Huang, H. et al. (2008). Isolation of neuronal chromatin from brain tissue. *BMC Neurosci* **9**: 42.

Johnson, DS., Davidson, B., Brown, CD. et al. (2004). Noncoding regulatory sequences of ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Res.* **14**: 2448-2456.

Joly, J., Kano, S., Matsuoka, T. et al. (2007). Culture of ciona intestinalis in closed systems. *Dev. Dyn.* **236**: 1832-1840.

Jones, EA. and Flavell, RA. (2005). Distal enhancer elements transcribe intergenic rna in the il-10 family gene cluster. *J. Immunol.* **175**: 7437-7446.

Juan, AH. and Ruddle, FH. (2003). Enhancer timing of hox gene expression: deletion of the endogenous hoxc8 early enhancer. *Development* **130**: 4823-4834.

Juven-Gershon, T., Cheng, S. and Kadonaga, JT. (2006). Rational design of a super core promoter that enhances gene expression. *Nat. Methods* **3**: 917-922.

Kent, WJ., Baertsch, R., Hinrichs, A. et al. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **100**: 11484-11489.

Keys, DN., Lee, B., Di Gregorio, A. et al. (2005). A saturation screen for cis-acting regulatory dna in the hox genes of ciona intestinalis. *Proc. Natl. Acad. Sci. U.S.A.* **102**: 679-683.

Khandekar, M., Suzuki, N., Lewton, J. et al. (2004). Multiple, distant gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system. *Mol. Cell. Biol.* **24**: 10263-10276.

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

Kim, JH., Waterman, MS. and Li, LM. (2007). Diploid genome reconstruction of ciona intestinalis and comparative analysis with ciona savignyi. *Genome Res.* **17**: 1101-1110.

Kim, TH., Abdullaev, ZK., Smith, AD. et al. (2007). Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome. *Cell* **128**: 1231-1245.

Kimbacher, S., Gerstl, I., Velimirov, B. et al. (2009). Drosophila p transposons of the urochordata ciona intestinalis. *Mol. Genet. Genomics* : .

Kimura-Yoshida, C., Kitajima, K., Oda-Ishii, I. et al. (2004). Characterization of the pufferfish otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131**: 57-71.

King, DC., Taylor, J., Zhang, Y. et al. (2007). Finding cis-regulatory elements using comparative genomics: some lessons from encode data. *Genome Res.* **17**: 775-786.

Kleinjan, DA., Seawright, A., Childs, AJ. et al. (2004). Conserved elements in pax6 intron 7 involved in (auto)regulation and alternative transcription. *Dev. Biol.* **265**: 462-477.

Kobayashi, A., Watanabe, Y., Akasaka, K. et al. (2007). Real-time monitoring of functional interactions between upstream and core promoter sequences in living cells of sea urchin embryos. *Nucleic Acids Res.* **35**: 4882-4894.

Kulkarni, MM. and Arnosti, DN. (2003). Information display by transcriptional enhancers. *Development* **130**: 6569-6575.

Kusakabe, T., Yoshida, R., Ikeda, Y. et al. (2004). Computational discovery of dna motifs associated with cell type-specific gene expression in ciona. *Dev. Biol.* **276**: 563-580.

Lampe, X., Samad, OA., Guiguen, A. et al. (2008). An ultraconserved hox-pbx responsive element resides in the coding sequence of hoxa2 and is active in rhombomere 4. *Nucleic Acids Res.* **36**: 3214-3225.

Lamy, C. and Lemaire, P. (2008). [ascidian embryos: from the birth of experimental embryology to the analysis of gene regulatory networks]. *Med Sci (Paris)* **24**: 263-269.

Lee, AM. and Wu, C. (2006). Enhancer-promoter communication at the yellow gene of drosophila melanogaster: diverse promoters participate in and regulate trans interactions. *Genetics* **174**: 1867-1880.

Lee, AP., Koh, EGL., Tay, A. et al. (2006). Highly conserved syntenic blocks at the vertebrate hox loci and conserved regulatory elements within and outside hox gene clusters. *Proc. Natl. Acad. Sci. U.S.A.* **103**: 6994-6999.

Lee, R. (2005). Web resources for c. elegans studies. *WormBook* : 1-16.

Lettice, LA., Heaney, SJH., Purdie, LA. et al. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**: 1725-1735.

Lettice, LA., Hill, AE., Devenney, PS. et al. (2008). Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum. Mol. Genet.* **17**: 978-985.

Li Song, D. and Joyner, AL. (2000). Two pax2/5/8-binding sites in engrailed2 are required for proper initiation of endogenous mid-hindbrain expression. *Mech. Dev.* **90**: 155-165.

Li, Q., Barkess, G. and Qian, H. (2006). Chromatin looping and the probability of transcription. *Trends Genet.* **22**: 197-202.

Li, X. and Noll, M. (1994). Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the drosophila embryo. *EMBO J.* **13**: 400-406.

Li, X., MacArthur, S., Bourgon, R. et al. (2008). Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol.* **6**: e27.

Li, X., Tan, L., Wang, L. et al. (2009). Isolation and characterization of conserved non-coding sequences among rice (oryza sativa l.) paralogous regions. *Mol. Genet. Genomics* **281**: 11-18.

Lifanov, AP., Makeev, VJ., Nazina, AG. et al. (2003). Homotypic regulatory clusters in drosophila. *Genome Res.* **13**: 579-588.

Lin, Q., Chen, Q., Lin, L. et al. (2004). The promoter targeting sequence mediates epigenetically heritable transcription memory. *Genes Dev.* **18**: 2639-2651.

Liou, J., Kim, ML., Heo, WD. et al. (2005). Stim is a ca2+ sensor essential for ca2+-store-depletion-triggered ca2+ influx. *Curr. Biol.* **15**: 1235-1241.

Liu, L., Xiang, J., Dong, B. et al. (2006). Ciona intestinalis as an emerging model organism: its regeneration under controlled conditions and methodology for egg dechorionation. *J Zhejiang Univ Sci B* **7**: 467-474.

Lomvardas, S., Barnea, G., Pisapia, DJ. et al. (2006). Interchromosomal interactions and olfactory receptor choice. *Cell* **126**: 403-413.

Long, Q., Meng, A., Wang, H. et al. (1997). Gata-1 expression pattern can be recapitulated in living transgenic zebrafish using gfp reporter gene. *Development* **124**: 4105-4111.

Long, X. and Miano, JM. (2007). Remote control of gene expression. *J. Biol. Chem.* **282**: 15941-15945.

Loots, GG. (2008). Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. *Adv. Genet.* **61**: 269-293.

Luik, RM., Wang, B., Prakriya, M. et al. (2008). Oligomerization of stim1 couples er calcium depletion to crac channel activation. *Nature* **454**: 538-542.

Lynch, M. and Conery, JS. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

MacIsaac, KD. and Fraenkel, E. (2006). Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput. Biol.* **2**: e36.

[DGRPWebsite] . http://service004.hpc.ncsu.edu/mackay/Good_Mackay_site/DBRP.html.

Maconochie, MK., Nonchev, S., Studer, M. et al. (1997). Cross-regulation in the mouse hoxb complex: the expression of hoxb2 in rhombomere 4 is regulated by hoxb1. *Genes Dev.* **11**: 1885-1895.

Maeda, RK. and Karch, F. (2007). Making connections: boundaries and insulators in drosophila. *Curr. Opin. Genet. Dev.* **17**: 394-399.

Makeev, VJ., Lifanov, AP., Nazina, AG. et al. (2003). Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.* **31**: 6016-6026.

Markstein, M., Markstein, P., Markstein, V. et al. (2002). Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the drosophila embryo. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 763-768.

Markstein, M., Pitsouli, C., Villalta, C. et al. (2008). Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.* **40**: 476-483.

Marlin, S., Blanchard, S., Slim, R. et al. (1999). Townes-brocks syndrome: detection of a sall1 mutation hot spot and evidence for a position effect in one patient. *Hum. Mutat.* **14**: 377-386.

Marschall, T. and Rahmann, S. (2009). Efficient exact motif discovery. *Bioinformatics* **25**: i356-64.

Maston, GA., Evans, SK. and Green, MR. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29-59.

Matthysse, AG., Deschet, K., Williams, M. et al. (2004). A functional cellulose synthase from ascidian epidermis. *Proc. Natl. Acad. Sci. U.S.A.* **101**: 986-991.

McEwen, GK., Woolfe, A., Goode, D. et al. (2006). Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome Res.* **16**: 451-465.

McGaughey, DM., Vinton, RM., Huynh, J. et al. (2008). Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res.* **18**: 252-260.

McLean, C. and Bejerano, G. (2008). Dispensability of mammalian dna. *Genome Res.* **18**: 1743-1751.

McLellan, AS., Kealey, T. and Langlands, K. (2006). An e box in the exon 1 promoter regulates insulin-like growth factor-i expression in differentiating muscle cells. *Am. J. Physiol., Cell Physiol.* **291**: C300-7.

Merli, C., Bergstrom, DE., Cygan, JA. et al. (1996). Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**: 1260-1270.

Messer, PW. and Arndt, PF. (2007). The majority of recent short dna insertions in the human genome are tandem duplications. *Mol. Biol. Evol.* **24**: 1190-1197.

Missal, K., Rose, D. and Stadler, PF. (2005). Non-coding rnas in ciona intestinalis. *Bioinformatics* **21 Suppl 2**: ii77-8.

Morand, S., Ueyama, T., Tsujibe, S. et al. (2009). Duox maturation factors form cell surface complexes with duox affecting the specificity of reactive oxygen species generation. *FASEB J.* **23**: 1205-1218.

Nakashima, K., Yamada, L., Satou, Y. et al. (2004). The evolutionary origin of animal cellulose synthase. *Dev. Genes Evol.* **214**: 81-88.

Navratilova, P., Fredman, D., Hawkins, TA. et al. (2009). Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**: 526-540.

Ng, YK., Wu, W. and Zhang, L. (2009). Positive correlation between gene coexpression and positional clustering in the zebrafish genome. *BMC Genomics* **10**: 42.

Nikolaidis, N., Chalkia, D., Watkins, DN. et al. (2007). Ancient origin of the new developmental superfamily danger. *PLoS ONE* **2**: e204.

Nishihara, H., Smit, AFA. and Okada, N. (2006). Functional noncoding sequences derived from sines in the mammalian genome. *Genome Res.* **16**: 864-874.

Nishiyama, A. and Fujiwara, S. (2008). Rna interference by expressing short hairpin rna in the ciona intestinalis embryo. *Dev. Growth Differ.* **50**: 521-529.

Nobrega, MA. and Pennacchio, LA. (2004). Comparative genomic analysis as a tool for biological discovery. *J. Physiol. (Lond.)* **554**: 31-39.

Nobrega, MA., Ovcharenko, I., Afzal, V. et al. (2003). Scanning human gene deserts for long-range enhancers. *Science* **302**: 413.

Nóbrega, MA., Zhu, Y., Plajzer-Frick, I. et al. (2004). Megabase deletions of gene deserts result in viable mice. *Nature* **431**: 988-993.

Norden-Krichmar, TM., Holtz, J., Pasquinelli, AE. et al. (2007). Computational prediction and experimental validation of ciona intestinalis microrna genes. *BMC Genomics* **8**: 445.

Nydam, M. and Harrison, R. (2007). Genealogical relationships within and among shallow-water ciona species (ascidiacea). *Marine Biology* **151**: 1839-1847.

Ohtsuki, S., Levine, M. and Cai, HN. (1998). Different core promoters possess distinct regulatory activities in the drosophila embryo. *Genes Dev.* **12**: 547-556.

Oliver, B., Parisi, M. and Clark, D. (2002). Gene expression neighborhoods. *J. Biol.* **1**: 4.

Ondov, BD., Varadarajan, A., Passalacqua, KD. et al. (2008). Efficient mapping of applied biosystems solid sequence data to a reference genome for functional genomic applications. *Bioinformatics* **24**: 2776-2777.

Ovcharenko, I., Loots, GG., Nobrega, MA. et al. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**: 137-145.

Palin, K., Taipale, J. and Ukkonen, E. (2006). Locating potential enhancer elements by comparative genomics using the eel software. *Nat Protoc* **1**: 368-374.

Pan, D. and Zhang, L. (2007). Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. *Genome Biol.* **8**: R158.

Panne, D., Maniatis, T. and Harrison, SC. (2007). An atomic model of the interferon-beta enhanceosome. *Cell* **129**: 1111-1123.

Papachatzopoulou, A., Kaimakis, P., Pourfarzad, F. et al. (2007). Increased gamma-globin gene expression in beta-thalassemia intermedia patients correlates with a mutation in 3'hs1. *Am. J. Hematol.* **82**: 1005-1009.

Papatsenko, D. (2007). Clusterdraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics* **23**: 1032-1034.

Papatsenko, D. and Levine, M. (2005). Computational identification of regulatory dnas underlying animal development. *Nat. Methods* **2**: 529-534.

Park, HC., Kim, CH., Bae, YK. et al. (2000). Analysis of upstream elements in the huc promoter leads to the establishment of transgenic zebrafish with fluorescent neurons. *Dev. Biol.* **227**: 279-293.

Patrinos, GP., de Krom, M., de Boer, E. et al. (2004). Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes Dev.* **18**: 1495-1509.

Pena, RN. and Whitelaw, CBA. (2005). Duplication of stat5-binding sites within the beta-lactoglobulin promoter compromises transcription in vivo. *Biochimie* **87**: 523-528.

Pennacchio, LA., Ahituv, N., Moses, AM. et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499-502.

Permanyer, J., Gonzàlez-Duarte, R. and Albalat, R. (2003). The non-ltr retrotransposons in ciona intestinalis: new insights into the evolution of chordate genomes. *Genome Biol.* **4**: R73.

Pierstorff, N., Bergman, CM. and Wiehe, T. (2006). Identifying cis-regulatory modules by combining comparative and compositional analysis of dna. *Bioinformatics* **22**: 2858-2864.

Potts, W., Tucker, D., Wood, H. et al. (2000). Chicken beta-globin 5'hs4 insulators function to reduce variability in transgenic founder mice. *Biochem. Biophys. Res. Commun.* **273**: 1015-1018.

Poulin, F., Nobrega, MA., Plajzer-Frick, I. et al. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**: 774-781.

Prabhakar, S., Poulin, F., Shoukry, M. et al. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.* **16**: 855-863.

Putnam, NH., Butts, T., Ferrier, DEK. et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064-1071.

Putnam, NH., Srivastava, M., Hellsten, U. et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**: 86-94.

Rabinovich, A., Jin, VX., Rabinovich, R. et al. (2008). E2f in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res.* **18**: 1763-1777.

Rahimov, F., Marazita, ML., Visel, A. et al. (2008). Disruption of an ap-2alpha binding site in an irf6 enhancer is associated with cleft lip. *Nat. Genet.* **40**: 1341-1347.

Rajewsky, N., Vergassola, M., Gaul, U. et al. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC Bioinformatics* **3**: 30.

Rebeiz, M., Reeves, NL. and Posakony, JW. (2002). Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 9888-9893.

Retelska, D., Beaudoing, E., Notredame, C. et al. (2007). Vertebrate conserved non coding dna regions have a high persistence length and a short persistence time. *BMC Genomics* **8**: 398.

Rinn, JL., Kertesz, M., Wang, JK. et al. (2007). Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell* **129**: 1311-1323.

Roider, HG., Kanhere, A., Manke, T. et al. (2007). Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics* **23**: 134-141.

Roider, HG., Manke, T., O'Keeffe, S. et al. (2009). Pastaa: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* **25**: 435-442.

Romano, LA. and 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.

Ronshaugen, M. and Levine, M. (2004). Visualization of trans-homolog enhancer-promoter interactions at the abd-b hox locus in the drosophila embryo. *Dev. Cell* **7**: 925-932.

Rumble, SM., Lacroute, P., Dalca, AV. et al. (2009). Shrimp: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**: e1000386.

Sabherwal, N., Bangs, F., Röth, R. et al. (2007). Long-range conserved non-coding shox sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Hum. Mol. Genet.* **16**: 210-222.

Sakuraba, Y., Kimura, T., Masuya, H. et al. (2008). Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm. Genome* **19**: 703-712.

Sanchez-Elsner, T., Gou, D., Kremmer, E. et al. (2006). Noncoding rnas of trithorax response elements recruit drosophila ash1 to ultrabithorax. *Science* **311**: 1118-1123.

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

Sandve, GK. and Drabløs, F. (2006). A survey of motif discovery methods in an integrated framework. *Biol. Direct* **1**: 11.

Santagati, F., Abe, K., Schmidt, V. et al. (2003). Identification of cis-regulatory elements in the mouse pax9/nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics* **165**: 235-242.

Satoh, N. (1994). Developmental biology of ascidians. : 4.

Satoh, N., Satou, Y., Davidson, B. et al. (2003). Ciona intestinalis: an emerging model for whole-genome analyses. *Trends Genet.* **19**: 376-381.

Satou, Y., Imai, KS. and Satoh, N. (2004). The ascidian mesp gene specifies heart precursor cells. *Development* **131**: 2533-2541.

Satou, Y., Kawashima, T., Shoguchi, E. et al. (2005). An integrated database of the ascidian, ciona intestinalis: towards functional genomics. *Zool. Sci.* **22**: 837-843.

Satou, Y., Mineta, K., Ogasawara, M. et al. (2008). Improved genome assembly and evidence-based global gene model set for the chordate ciona intestinalis: new insight into intron and operon populations. *Genome Biol.* **9**: R152.

Scheibye-Alsing, K., Hoffmann, S., Frankel, A. et al. (2009). Sequence assembly. *Comput Biol Chem* **33**: 121-136.

Schlamp, K., Weinmann, A., Krupp, M. et al. (2008). Blotbase: a northern blot database. *Gene* **427**: 47-50.

Schmidtke, J. and Engel, W. (1980). Gene diversity in tunicate populations. *Biochem. Genet.* **18**: 503-508.

Schroeder, MD., Pearce, M., Fak, J. et al. (2004). Transcriptional control in the segmentation gene network of drosophila. *PLoS Biol.* **2**: E271.

[Schutz2004] Schutz, B., Schafer, M., Weihe, E. et al. 19. transcriptional control of the cholinergic gene locus: a mosaic model for regulation of the cholinergic phenotype. Israel Silman, Hermona Soreq, Lili Anglister, Daniel M. Michaelson, Abraham Fisher. Taylor \& Francis, 2004.

Segal, E., Raveh-Sadka, T., Schroeder, M. et al. (2008). Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature* **451**: 535-540.

Sharan, SK., Thomason, LC., Kuznetsov, SG. et al. (2009). Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* **4**: 206-223.

Shi, W., Hendrix, D., Levine, M. et al. (2009). A distinct class of small rnas arises from pre-mirna-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.* **16**: 183-189.

Shi, W., Levine, M. and Davidson, B. (2005). Unraveling genomic regulatory networks in the simple chordate, ciona intestinalis. *Genome Res.* **15**: 1668-1674.

Shin, JT., Priest, JR., Ovcharenko, I. et al. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res.* **33**: 5437-5445.

Shiratori, H., Yashiro, K., Shen, MM. et al. (2006). Conserved regulation and role of pitx2 in situs-specific morphogenesis of visceral organs. *Development* **133**: 3015-3025.

Shoguchi, E., Kawashima, T., Nishida-Umehara, C. et al. (2005). Molecular cytogenetic characterization of ciona intestinalis chromosomes. *Zool. Sci.* **22**: 511-516.

Shoguchi, E., Kawashima, T., Satou, Y. et al. (2006). Chromosomal mapping of 170 bac clones in the ascidian ciona intestinalis. *Genome Res.* **16**: 297-303.

Siepel, A., Bejerano, G., Pedersen, JS. et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**: 1034-1050.

Sierro, N., Kusakabe, T., Park, K. et al. (2006). Dbtgr: a database of tunicate promoters and their regulatory elements. *Nucleic Acids Res.* **34**: D552-5.

Simonis, M., Klous, P., Splinter, E. et al. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet.* **38**: 1348-1354.

Small, KS., Brudno, M., Hill, MM. et al. (2007). Extreme genomic variation in a natural population. *Proc. Natl. Acad. Sci. U.S.A.* **104**: 5698-5703.

Smith, AD., Sumazin, P. and Zhang, MQ. (2007). Tissue-specific regulatory elements in mammalian promoters. *Mol. Syst. Biol.* **3**: 73.

Smith, J. (2008). A protocol describing the principles of cis-regulatory analysis in the sea urchin. *Nat Protoc* **3**: 710-718.

Spitz, F. and Duboule, D. (2008). Global control regions and regulatory landscapes in vertebrate development and evolution. *Adv. Genet.* **61**: 175-205.

Spradling, AC. and Rubin, GM. (1983). The effect of chromosomal position on the expression of the drosophila xanthine dehydrogenase gene. *Cell* **34**: 47-57.

Stephen, S., Pheasant, M., Makunin, IV. et al. (2008). Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.* **25**: 402-408.

Sun, H., Skogerbø, G. and Chen, R. (2006). Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet.* **15**: 2911-2922.

Suster, ML., Kania, A., Liao, M. et al. (2009). A novel conserved evx1 enhancer links spinal interneuron morphology and cis-regulation from fish to mammals. *Dev. Biol.* **325**: 422-433.

Suzuki, MM., Nishikawa, T. and Bird, A. (2005). Genomic approaches reveal unexpected genetic divergence within ciona intestinalis. *J. Mol. Evol.* **61**: 627-635.

Taher, L. and Ovcharenko, I. (2009). Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics* **25**: 578-584.

Tassy, O., Daian, F., Hudson, C. et al. (2006). A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. *Curr. Biol.* **16**: 345-358.

Taylor, J., Tyekucheva, S., King, DC. et al. (2006). Esperr: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.* **16**: 1596-1604.

Tompa, M., Li, N., Bailey, TL. et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**: 137-144.

Tronche, F., Ringeisen, F., Blumenfeld, M. et al. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**: 231-245.

Tsang, WH., Shek, KF., Lee, TY. et al. (2009). An evolutionarily conserved nested gene pair- mab21 and lrba/nbea in metazoan. *Genomics* : .

Tsuritani, K., Irie, T., Yamashita, R. et al. (2007). Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes. *Genome Res.* **17**: 1005-1014.

Tümpel, S., Cambronero, F., Sims, C. et al. (2008). A regulatory module embedded in the coding region of hoxa2 controls expression in rhombomere 2. *Proc. Natl. Acad. Sci. U.S.A.* **105**: 20077-20082.

Tursun, B., Cochella, L., Carrera, I. et al. (2009). A toolkit and robust pipeline for the generation of fosmid-based reporter genes in c. elegans. *PLoS ONE* **4**: e4625.

Uchikawa, M., Ishida, Y., Takemoto, T. et al. (2003). Functional analysis of chicken sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4**: 509-519.

Vakoc, CR., Letting, DL., Gheldof, N. et al. (2005). Proximity among distant regulatory elements at the beta-globin locus requires gata-1 and fog-1. *Mol. Cell* **17**: 453-462.

Vavouri, T. and Elgar, G. (2005). Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.* **15**: 395-402.

Vavouri, T., McEwen, GK., Woolfe, A. et al. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet.* **22**: 5-10.

Vavouri, T., Walter, K., Gilks, WR. et al. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol.* **8**: R15.

Venken, KJT., Carlson, JW., Schulze, KL. et al. (2009). Versatile p[acman] bac libraries for transgenesis studies in drosophila melanogaster. *Nat. Methods* : .

Vig, M. and Kinet, J. (2007). The long and arduous road to crac. *Cell Calcium* **42**: 157-162.

Vinson, JP., Jaffe, DB., O'Neill, K. et al. (2005). Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome Res.* **15**: 1127-1135.

Visel, A., Akiyama, JA., Shoukry, M. et al. (2009). Functional autonomy of distant-acting human enhancers. *Genomics* : .

Visel, A., Blow, MJ., Li, Z. et al. (2009). Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854-858.

Visel, A., Bristow, J. and Pennacchio, LA. (2007). Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.* **18**: 140-152.

Visel, A., Minovitsky, S., Dubchak, I. et al. (2007). Vista enhancer browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**: D88-92.

Voth, H., Oberthuer, A., Simon, T. et al. (2009). Co-regulated expression of hand2 and dein by a bidirectional promoter with asymmetrical activity in neuroblastoma. *BMC Mol. Biol.* **10**: 28.

Wada, S., Katsuyama, Y., Sato, Y. et al. (1996). Hroth an orthodenticle-related homeobox gene of the ascidian, halocynthia roretzi: its expression and putative roles in the axis formation during embryogenesis. *Mech. Dev.* **60**: 59-71.

Walter, K., Abnizova, I., Elgar, G. et al. (2005). Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.* **21**: 436-440.

Wang, H., Zhang, Y., Cheng, Y. et al. (2006). Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res.* **16**: 1480-1492.

Wang, Q., Prabhakar, S., Chanan, S. et al. (2007). Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol.* **8**: R1.

Wang, T., Chen, Y., Liu, C. et al. (2002). Functional analysis of the proximal promoter regions of fish rhodopsin and myf-5 genes using transgenesis. *Mar. Biotechnol.* **4**: 247-255.

Wang, W., Zhong, J., Su, B. et al. (2007). Comparison of pax1/9 locus reveals 500-myr-old syntenic block and evolutionary conserved noncoding regions. *Mol. Biol. Evol.* **24**: 784-791.

Washietl, S., Pedersen, JS., Korbel, JO. et al. (2007). Structured rnas in the encode selected regions of the human genome. *Genome Res.* **17**: 852-864.

Wasserman, WW. and Fickett, JW. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**: 167-181.

Wasserman, WW. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**: 276-287.

Waterston, RH., Lindblad-Toh, K., Birney, E. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Wederell, ED., Bilenky, M., Cullum, R. et al. (2008). Global analysis of in vivo foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* **36**: 4549-4564.

Wenick, AS. and Hobert, O. (2004). Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in c. elegans. *Dev. Cell* **6**: 757-770.

Werner, T., Hammer, A., Wahlbuhl, M. et al. (2007). Multiple conserved regulatory elements with overlapping functions determine sox10 expression in mouse embryogenesis. *Nucleic Acids Res.* **35**: 6526-6538.

West, AG. and Fraser, P. (2005). Remote control of gene transcription. *Hum. Mol. Genet.* **14 Spec No 1**: R101-11.

Woltering, JM. and Duboule, D. (2009). Conserved elements within open reading frames of mammalian hox genes. *J. Biol.* **8**: 17.

Woolfe, A. and Elgar, G. (2007). Comparative genomics using fugu reveals insights into regulatory subfunctionalization. *Genome Biol.* **8**: R53.

Woolfe, A. and Elgar, G. (2008). Organization of conserved elements near key developmental regulators in vertebrate genomes. *Adv. Genet.* **61**: 307-338.

Woolfe, A., Goodson, M., Goode, DK. et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**: e7.

Wratten, NS., McGregor, AP., Shaw, PJ. et al. (2006). Evolutionary and functional analysis of the tailless enhancer in musca domestica and drosophila melanogaster. *Evol. Dev.* **8**: 6-15.

Wu, MM., Buchanan, J., Luik, RM. et al. (2006). Ca2+ store depletion causes stim1 to accumulate in er regions closely associated with the plasma membrane. *J. Cell Biol.* **174**: 803-813.

Xie, X., Kamal, M. and Lander, ES. (2006). A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl. Acad. Sci. U.S.A.* **103**: 11659-11664.

Xiong, L., Catoire, H., Dion, P. et al. (2009). Meis1 intronic risk haplotype associated with restless legs syndrome affects its mrna and protein expression levels. *Hum. Mol. Genet.* **18**: 1065-1074.

Xiong, N., Kang, C. and Raulet, DH. (2002). Redundant and unique roles of two enhancer elements in the tcrgamma locus in gene regulation and gammadelta t cell development. *Immunity* **16**: 453-463.

Xu, S., McCusker, J. and Krauthammer, M. (2008). Yale image finder (yif): a new search engine for retrieving biomedical images. *Bioinformatics* **24**: 1968-1970.

Xu, X., Scott, MM. and Deneris, ES. (2006). Shared long-range regulatory elements coordinate expression of a gene cluster encoding nicotinic receptor heteromeric subtypes. *Mol. Cell. Biol.* **26**: 5636-5649.

Yagi, K., Takatori, N., Satou, Y. et al. (2005). Ci-tbx6b and ci-tbx6c are key mediators of the maternal effect gene ci-macho1 in muscle cell differentiation in ciona intestinalis embryos. *Dev. Biol.* **282**: 535-549.

Yanagisawa, H., Clouthier, DE., Richardson, JA. et al. (2003). Targeted deletion of a branchial arch-specific enhancer reveals a role of dhand in craniofacial development. *Development* **130**: 1069-1078.

Yang, GS., Banks, KG., Bonaguro, RJ. et al. (2009). Next generation tools for high-throughput promoter and expression analysis employing single-copy knock-ins at the hprt1 locus. *Genomics* **93**: 196-204.

Yoshida, K. and Saiga, H. (2008). Left-right asymmetric expression of pitx is regulated by the asymmetric nodal signaling through an intronic enhancer in ciona intestinalis. *Dev. Genes Evol.* **218**: 353-360.

Yoshikawa, S., Norcom, E., Nakamura, H. et al. (2007). Transgenic analysis of the anterior eye-specific enhancers of the zebrafish gelsolin-like 1 (gsnl1) gene. *Dev. Dyn.* **236**: 1929-1938.

Yuan, Y., Guo, L., Shen, L. et al. (2007). Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* **3**: e243.

Yuh, CH., Bolouri, H. and Davidson, EH. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**: 1896-1902.

Zhang, C., Xuan, Z., Otto, S. et al. (2006). A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.* **34**: 2238-2246.

Zhang, ZD., Paccanaro, A., Fu, Y. et al. (2007). Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions. *Genome Res.* **17**: 787-797.

Zheng, JB., Zhou, YH., Maity, T. et al. (2001). Activation of the human pax6 gene through the exon 1 enhancer by transcription factors sef and sp1. *Nucleic Acids Res.* **29**: 4070-4078.

Zhou, J. and Levine, M. (1999). A novel cis-regulatory element, the pts, mediates an anti-insulator activity in the drosophila embryo. *Cell* **99**: 567-575.

[NCICRF] . http://recombineering.ncifcrf.gov.