Lawrence Berkeley National Laboratory

Title
Short sequence motifs, overrepresented in mammalian conserved non-coding sequences

Permalink
https://escholarship.org/uc/item/7904m9kr

Authors
Minovitsky, Simon
Stegmaier, Philip
Kel, Alexander
et al.

Publication Date
2007-02-21

Peer reviewed

Short sequence motifs, overrepresented in mammalian conserved non-coding sequences

Simon Minovitsky[1], Philip Stegmaier[2], Alexander Kel[2], Alexey S. Kondrashov[3], Inna Dubchak[1,4,5]

1. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720,
2. BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany,
3. Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48103
4. DOE Joint Genome Institute, Walnut Creek, CA 94598
5. Corresponding author, ildubchak@lbl.gov

Abstract

Background: A substantial fraction of non-coding DNA sequences of multicellular eukaryotes is under selective constraint. In particular, ~5% of the human genome consists of conserved non-coding sequences (CNSs). CNSs differ from other genomic sequences in their nucleotide composition and must play important functional roles, which mostly remain obscure.

Results: We investigated relative abundances of short sequence motifs in all human CNSs present in the human/mouse whole-genome alignments *vs*. three background sets of sequences: (i) weakly conserved or unconserved non-coding sequences (non-CNSs); (ii) near-promoter sequences (located between nucleotides -500 and -1500, relative to a start of transcription); and (iii) random sequences with the same nucleotide composition as that of CNSs. When compared to non-CNSs and near-promoter sequences, CNSs possess an excess of AT-rich motifs, often containing runs of identical nucleotides. In contrast, when compared to random sequences, CNSs contain an excess of GC-rich motifs which, however, lack CpG dinucleotides. Thus, abundance of short sequence motifs in human CNSs, taken as a whole, is mostly determined by their overall compositional properties and not by overrepresentation of any specific short motifs. These properties are: (i) high AT-content of CNSs, (ii) a tendency, probably due to context-dependent mutation, of A's and T's to clump, (iii) presence of short GC-rich regions, and (iv) avoidance of CpG contexts, due to their hypermutability. Only a small number of short motifs, overrepresented in all human CNSs are similar to binding sites of transcription factors from the FOX family.

Conclusion: Human CNSs as a whole appear to be too broad a class of sequences to possess strong footprints of any short sequence-specific functions. Such footprints should be studied at the level of functional subclasses of CNSs, such as those which flank genes with a particular pattern of expression. Overall properties of CNSs are affected by patterns in mutation, suggesting that selection which causes their conservation is not always very strong.

Background

Genomes of multicellular eukaryotes mostly consist of DNA segments which do not encode proteins. Still, a sizeable fraction of such non-coding DNA is subject to selective constraint and, thus, is conserved between species. Typically, a long intergenic region consists of alternating segments with

high and low rates of evolution [1]. A variety of terms have been used to refer to slowly-evolving segments [2, 3], here we will call them CNSs (conservative non-coding sequences).

A majority of mutations in segments which evolve at high rates are presumably selectively neutral or nearly-neutral. In contrast, a large fraction of mutations within CNSs must be deleterious enough to be removed by negative selection. Indeed, data on within-population genetic variability indicate that slow evolution of CNSs is due to negative selection, and not to locally reduced mutation rate [4]. In multicellular eukaryotes with compact genomes, such as *Drosophila melanogaster*, a majority of mutations affecting non-coding sequences may be removed by selection [5, 6]. For large-genome organisms, such as mammals, the fraction of selectively constrained non-coding sequences is probably between 3% [7] and ~10% [8].

Obviously, CNSs must perform important biological functions, but the whole range and nature of these functions remains unknown [9]. Still, many CNSs are certainly involved in regulation of transcription, and harbor binding sites of a variety of transcription factors [10]. Thus, we can expect some short sequence motifs to be overrepresented in at least some kinds of CNSs, as this is the case for proximal promoters [11]. Indeed, analyses of samples from human CNSs demonstrated overrepresentation of some short sequence motifs [12, 13].

New, powerful methods of detecting overrepresented motifs [*e. g.*, [14, 15]], make it possible to undertake the analysis of small-scale composition of mammalian CNSs at the genomic level. Such analysis has a potential to reveal short sequence-specific function(s) common for all human CNSs. Here, we report the results of application of discriminating matrix enumerator (DME) [14] to all strong human CNSs.

Results

We studied representation of short sequence motifs in all human CNSs against three backgrounds: unconserved or only weakly conserved segments of intergenic regions (non-CNSs), near-promoter non-coding sequences, and randomized sequences with the same nucleotide composition as that of CNSs. CNSs are relatively AT-rich [9]: frequencies of nucleotides A, T, G, and C are 30.7%, 30.7%, 19.3%, and 19.3% in CNSs, 26.3%, 26.4%, 23.6%, and 23.7% in non-CNSs, and 23.7%, 23.7%, 26.3%, and 26.3% in near-promoter sequences. Dinucleotide compositions of sequences of different classes were also substantially different (Fig. 1).
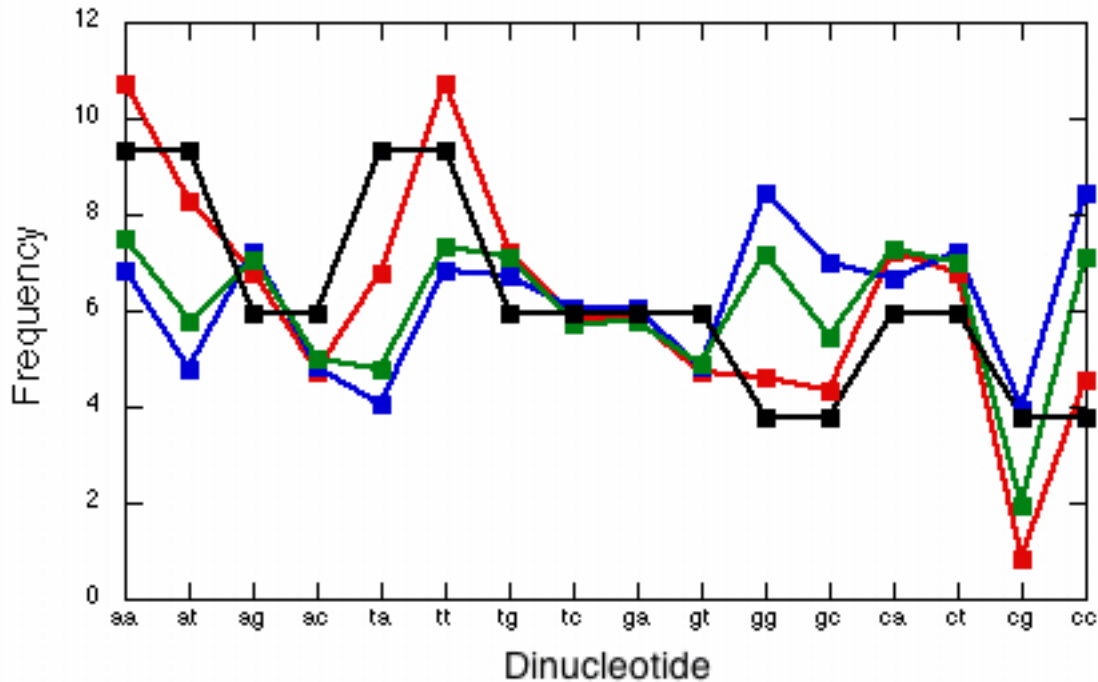
Fig. 1. Percentages of dinucleotide frequencies, in CNSs (red), non-CNSs (green), near-promoters (blue), and random sequences (black).

CNSs from human chromosomes with odd and even numbers were analyzed separately, to check the results for consistency. The overall lengths of CNSs were 27,112,333 on odd chromosomes and 24,962,379 on even chromosomes. Tables 1, 2, and 3 list top 30 motifs, overrepresented within CNSs over these three backgrounds. Overrepresentation was calculated as the ratio of the number of occurrences of a motif within CNSs, normalized to their overall length, over normalized number of occurrences of the motif within the background sequences.

Table 1 Motifs that are overrepresented in CNSs, over non-CNSs.

Odd Chromosomes			Even Chromosomes		
Motif	Number of occurrences	Overrepre-sentation	Motif	Number of occurrences	Overrepresen-tation
SYTAATTA	10620	3.45	TTAATTAV	12637	3.72
CTRATTAS	6152	3.14	TAATTRCW	12019	3.43
WGYAATTA	12596	3.09	GYAATTAS	6142	3.39
TTAATTAV	13141	3.08	TTTAATBA	15060	3.14
STAATTGV	8267	2.89	ATTAATBA	10910	3.07
VWGCTAAT	10503	2.84	TAATTWGM	10885	3.04
TTTAATBA	15800	2.77	GMWTAATT	9941	2.97
GMWTAATT	10290	2.72	CWTAATKA	10028	2.94
TAATTATV	10100	2.72	ATTAAWTT	11570	2.85
STTAATKG	5905	2.71	TTAATBAT	10115	2.79
ATTVAATT	12177	2.68	CWKTAATT	13079	2.75
ATTAATBA	11006	2.61	VWGCTAAT	9823	2.71
CWKTAATT	13577	2.59	CMATWAAT	10129	2.65
ATAATTAV	10536	2.58	ATTTVATT	15715	2.64

Motif	Occurrences	Over	Motif	Occurrences	Over
SMAATTAA	12754	2.57	CAATTRCH	8188	2.61
SBTAATGA	8828	2.56	MCWAATTA	9605	2.61
VATTWGCA	14265	2.53	ATTWWGCA	9959	2.61
TWAATCAR	10639	2.52	GKTAATTW	9019	2.59
AATTAVTT	12668	2.51	AATTAMCW	10053	2.58
GTAATTMM	7484	2.49	MATTDGCA	13694	2.58
GSABTAAT	7037	2.47	AATKCAWT	13437	2.58
AATTAMCW	10556	2.44	AATTGCWV	10857	2.55
YTSAATTA	10187	2.41	TAATGMAW	11617	2.55
WGVCTAAT	7960	2.40	VTAATTTA	10419	2.51
AATBAAAT	16556	2.40	VTAATTAT	9233	2.51
MCWTTAAT	9861	2.40	TTAATTBA	10974	2.49
AGMTTWAT	9378	2.39	ATTWARCT	8601	2.49
VAATTAAT	11645	2.39	CCAATTWV	8890	2.49
TCYAATTW	11410	2.37	AAAKCAWT	15678	2.46
ATTWWGCA	10301	2.37	AAATTRCW	13888	2.45

Table 2 Motifs that are overrepresented in CNSs, over near-promoter sequences.

Odd Chromosomes			Even Chromosomes		
Motif	Number of occurrences	Overrepre-sentation	Motif	Number of occurrences	Overrepresen-tation
STAATTAS	7576	4.55	SYTAATTA	9852	4.26
TTAATKAR	17516	4.33	TTAATTAD	14561	4.07
GBTAATKA	12299	3.96	CTRATTAS	5744	3.90
VTAATTGM	10174	3.91	ATTAATGN	9762	3.74
TTTMATKA	19449	3.86	TAATTATD	11760	3.73
MTTMATTA	13688	3.82	TTTAATDA	16633	3.66
AATKYAAT	15204	3.73	ATAATTAB	9233	3.62
TTAATKGV	12925	3.72	TAATKSAA	10418	3.59
RTAATKAA	13613	3.68	STAATTGV	7823	3.55
MMTAATTA	12518	3.68	GYAATWAA	10608	3.55
TSTAATTW	14964	3.49	TGYAATTW	13322	3.51
AATKMATT	18824	3.48	AATGMWTT	15412	3.49
TGATWAAW	12898	3.46	AGYAATTW	12585	3.41
KATAATKA	10739	3.46	AATTDATT	14693	3.39
CATTAAKV	10838	3.42	AATTATAD	10379	3.36
CATWAWTT	14599	3.39	TWAATTGR	8896	3.35
CATTWAAW	19325	3.37	AWTARCAT	9601	3.35
CAATTAKV	9515	3.33	TAATTHAT	12789	3.34
ATRATTYA	13356	3.30	CWTTAATR	9114	3.32
ATTTYMAT	20983	3.29	ATTSMATT	11547	3.27
CTBAATTR	11382	3.29	TAAATTAH	12840	3.27
MTKTAATT	17971	3.28	CWKTAATT	13079	3.21
TTCAAWTW	15188	3.27	ATGYAAWT	13544	3.21
CAAWTTWA	13854	3.25	AATTTHAT	14806	3.19
AAWTWGAT	10997	3.25	SAATTMAT	8653	3.18
GMATVATT	12027	3.23	KAAATCAW	10721	3.18

ATTAAYMT	10759	3.22	TAATTTGN	10830	3.17
AAAKTMAT	15757	3.22	MTTCATWA	9825	3.17
AATTRMTA	9186	3.22	CAATTAYW	8391	3.16
ATTTRAYA	11851	3.17	ATAATKSA	9081	3.16

Table 3 Motifs that are overrepresented in CNSs, over randomized sequences.

Odd Chromosomes			Even Chromosomes		
Motif	Number of occurrences	Overrepre- sentation	Motif	Number of occurrences	Overrepre- sentation
CWGSCWGS	32472	7.50	CWGSCWGV	38927	5.78
SCCHGSCH	42207	5.68	SCCWGGSN	33122	5.63
GGSWGGSN	39555	5.55	CYCWSCCH	33976	5.50
CWGSCCWS	24103	5.52	RGCWGSCH	30738	4.95
RGTCCTBY	22100	5.45	GGSDGRGV	34873	4.93
GRGSWGRG	25293	5.36	CWGSCYCH	29902	4.78
CCYYYCCH	40727	5.22	CWSCWGGV	31840	4.73
SCCWGGRV	33839	5.20	SCWGCWGV	30968	4.71
CWGSCYCH	36409	5.04	CWGGGRRV	31866	4.64
SCWGGGSN	36038	5.03	CWGRGSCH	28886	4.61
SCHGSCCH	36013	4.91	CCWGGRRV	31578	4.61
CWGRGSCH	35318	4.77	SCHGGSCH	28689	4.50
SCYCWGCH	34141	4.56	GGRARGRR	29240	4.47
NCAGCTGN	32928	4.52	RRGGCWGV	30772	4.44
CAGCTGNN	32867	4.51	RGGGRARR	29828	4.41
TWACWGAA	14781	4.48	GVWGGGRR	31019	4.37
RGGGRRAR	32929	4.42	CYCYVSCC	19097	4.37
CWGSAGSY	24140	4.37	KCCWSCCH	26417	4.33
SCWGGRAR	32065	4.37	CAGCYSNG	16617	4.28
GGARRGRR	33390	4.37	KKGGCWGV	28051	4.13
SWGGGGRV	27662	4.32	AARGRAAA	26189	4.10
DGCWGCCN	31093	4.32	CWSWGGGV	26455	4.04
SSSCKGGS	19245	4.28	WSCWGGGV	26019	4.00
SWGGGARR	31541	4.24	CWGGSCWV	26144	3.98
CWGSAGRR	29155	4.12	SCMGCKGS	9832	3.97
BCTSCAGV	28062	4.11	CCCWSWGV	26134	3.95
SDGGGMGS	18924	4.11	CCWSCCSS	8978	3.90
ARAGRAAA	28006	4.00	CYMSCYCC	12072	3.86
NCCASCCH	29182	3.93	CCYVSCCH	27206	3.86
CHCCCHCH	31310	3.92	CWGCTBCH	27372	3.84

In order to study a possible similarity of the overrepresented CNS motifs with known binding sites for transcription factors (TF), we applied our recently developed method m2transfac [16], and compareed all the motifs found at the previous step with the TRANSFAC library of positional weight matrices (PWMs). Relatively few matches between the motifs and the TF matrices were found. Out of 12000 motifs reported at the previous step as being overrepresented in CNS versus the three different backgrounds, we have identified just 20 motifs that match TF matrices with E-values lower than 0.001

and satisfy factor class-specific cut-offs (Table 4). The majority of these matches involved matrices for the factors of "Forkhead DNA-binding domain", especially of the FOX family, which were repeatedly found over two rather different backgrounds: of non-CNSs and randomized sequences. Among the motifs found over the background of near-promoter sequences, there was only one that matched a PWM.

Table 4. Motifs found matching transcription factor PWMs from TRANSFAC.

Accession	Consensus/ID	Factor class	Taxon	Binding factors
acns even				
DME280	ATAAACAN	Forkhead DNA-binding domain	Vertebrate	FOXI1a,FOXF1,FOXL1,FOXO4
DME424	WGTAAAYA	Forkhead DNA-binding domain	Vertebrate	FOXC1,FOXA4a,HNF-3beta
DME768	WTGTCATV	Basic region + leucine zipper (bZIP)	Nematode	Skn-1
DME1427	WGTCATSM	Basic region + leucine zipper (bZIP)	Nematode	Skn-1
acns odd				
DME27	VATTWGCA	POU	Vertebrate	POU2F1
DME349	ATAAACAN	Forkhead DNA-binding domain	Vertebrate	FOXI1a,FOXF1,FOXL1,FOXO4
DME1014	GTMAACAD	Forkhead DNA-binding domain	Vertebrate	FOXD1,HNF-3beta,FOXO1a
DME1700	CCAATMAB	DNA-binding domain with Histone fold	Fungal	HAP2,HAP3,HAP4
promoters even				
promoters odd				
DME1268	STGASTYA	Basic region + leucine zipper (bZIP)	Vertebrate	NF-E2,AP-1
random even				
DME90	VCAGATGN	Basic region + helix-loop-helix motif	Vertebrate	ITF-2,Tal-1beta
DME94	CATCTGBN	Basic region + helix-loop-helix motif	Vertebrate	ITF-2,Tal-1beta,E47
DME765	RTGWSTCA	Basic region + leucine zipper (bZIP)	Vertebrate	NF-E2,AP-1,Fos,Jun,Fra
DME1106	TGTTBACW	Forkhead DNA-binding domain	Vertebrate	HNF-3beta
DME1111	ATAAACAH	Forkhead DNA-binding domain	Vertebrate	FOXI1a,FOXF1,FOXL1,FOXO4
DME1920	CCACGTGG	Basic region + helix-loop-helix motif	Plant,Vertebrate	PIF3,c-Myc:Max
random odd				
DME11	CAGCTGNN	Basic region + helix-loop-helix motif	Vertebrate	AP-4
DME456	MAYAAACA	Forkhead DNA-binding domain	Vertebrate	FOXF1
DME790	TATGVAAA	POU	Vertebrate	POU2F1
DME930	ATAAAYAT	Forkhead DNA-binding domain	Vertebrate,Insect	FOXI1a,Croc
DME1145	TGTTBACW	Forkhead DNA-binding domain	Vertebrate	HNF-3beta

Discussion

We treated all human CNSs as a single class of sequences. Comparison of this class against three different backgrounds demonstrates that many short sequence motifs are substantially

overrepresented within CNSs (Tables 1-3). CNSs from odd- and from even-numbered human chromosomes show very similar patterns, which is consistent with the lack of any large-scale heterogeneity within CNSs. At a first glance, these results may seem to suggest that CNSs as a whole possess some complex sequence pattern(s), with possible implications for their functioning. However, this is probably not the case. Instead, the results can be explained by simple, generic properties of CNSs.

Indeed, when CNSs are analyzed against a background of non-CNSs (Table 1) or of near-promoter sequences (Table 2), almost all overrepresented motifs possess two common features: (i) they are AT-rich (consist of 75% or more of A and/or T) and (ii) they contain runs of A's and/or T's. Feature (i) simply reflects a well-known, although poorly understood, fact that CNSs are more AT-rich than the genome as a whole [9, 17] or that these two classes of background sequences. Feature (ii) appears to be due to general excess of AA and TT dinucleotides in CNSs, relatively to corresponding random sequences (Fig. 2). This tendency of A's and T'e to clump is probably due to patterns in mutation, and not to any functional constraint. Indeed, context-dependence of spontaneous mutation in mammals tends to produce runs of A's and T's, because at a site preceded and followed by A's (T's) T>A (A>T) transversions are ~2 times more common than A>T transversions [[18, 19]; Table 2].

Obviously, it is neccessary to consider CNSs against a background of the same nucleotide composition, as otherwise the impact of different compositions is the leading factor causing overrepresentation of some motifs. When CNSs are analyzed against a background of random sequences of the same, AT-rich, nucleotide composition, the results are very different (Table 3), and overrepresented motifs can be naturally subdivided into two classes. The first, larger class contains a variety of GC-rich motifs which, however, are devoid of CpG dinucleotides and are correspondingly enriched with CpA and CpT dinucleotides and with CWG short motif. The second, smaller class contains several motifs which are either purine- or pyrimidine-rich. Overrepresentation of motifs from the first class appear to be due to two simple factors: i) the presence, within CNSs, of short GC-rich segments and ii) hypermutability of CpG dinucleotides [18]. Indeed, CNSs are depleted of CpG's more than the other two classes of genomic sequences (Fig. 1), which might reflect strong methilation of CNSs. Overrepresentation of motifs of the second class simply reflects a well-known [20], although poorly understood, abundance of short segments with strong purine/pyrimidine imbalance between the two DNA stands within the human genome.

The analysis of all human CNSs does not reveal clear patterns consistent with overrepresentation of specific, functional motifs. A small number of the observed overrepresented motifs are similar to Position Weight Matrices (PWMs) from TRANSFAC database [21] (Table 4). Among them, the strongest similarity was to the PWMs of FOX family of factors which are characterized by a specific AT-rich pattern. The FOX factors are involved in many cellular processes and often control very first steps of organism development as well as cell cycle and differentiation; e. g. FOXF1 is highly expressed in mouse embryonic extraembryonic and lateral mesoderm [22] and control murine gut development [23]; FOXD1 is predominantly expressed in embryonic forebrain neuroepithelium, head mesenchyme and adrenal cortex [24] and controls normal brain and kidney morphogenesis and cellularity in the renal capsule [25]; FOXO1 governs cell growth in the heart [26]. Factors of other families, such as POU and bZIP are often involved in regulation of basic cell cycle machinery; e.g. POU2F1 is an ubiquitous factor involved in stimulation of replication [27] and also participates in early mouse embryogenesis [28]. In summary, it might be tempting to speculate that at

least some motifs overrepresented in all CNSs may play crucial role in organizing the process of development of the vertebrate organisms. However, the number of such motifs is not high., More specific classes of CNSs, such as those adjacent to genes with a particular pattern in expression [11, 12] should be considered in order to find a larger number of functional motifs.

In contrast, small-scale composition of human CNSs, considered as a whole, is strongly affected by patterns in mutation - hypermutability of CpG's and the tendency for A's and T's to form runs. This is unexpected because CNSs must be under negative selection which can overcome any impact of mutation [4]. Apparently, selective constraint on the evolution of individual nucleoitide site can be quite weak even within strongly conserved CNSs.

Conclusions

Abundance of short sequence motifs in all human CNSs is mostly dictated by their general features: overall AT-richness of CNSs, runs of A's and T's, GC-rich regions, avoidance of CpG's, and local purine/pyrimidine imbalance of the DNA strands. Apparently, CNSs as a whole are too broad a class to display strong overrepresentation of specific motifs. Instead, such motifs must be sought within subclasses of CNSs. In particular, tissue-specificity of expression of the genes adjacent to a CNS must be taken into account.

Methods

We used the VISTA pipeline infrastructure [29] with Shuffle-LAGAN glocal chaining algorithm [30] applied to local alignments produced by translated BLAT [31] for the construction of genome-wide pairwise human/mouse alignment. The level of conservation in the alignment was evaluated with the Gumby program [32]. Intervals with P-value threshold of 0.01 produced a set of 144,165 highly conserved sequences that totaled 49 Mb in length. We eliminated all conserved regions that coincide with the coding evidence provided by the UCSC data sets of mRNA, human spliced EST and human EST. We excluded CNSs located within (-1000, +1000) from the start and end of transcription.

Non-CNSs were defined as regions that have human/mouse alignment, conserved below 50% in a 100 bp window and not containing repeats and coding evidences. Random sequences were generated using standard C library pseudo-random generator. Overrepresentation of motifs in different sequences was calculated using DME [14]. DME identifies motifs, represented as position weight matrices that are overrepresented in one set of sequences relative to another set. The ability to directly optimize relative overrepresentation is a unique feature of DME, making DME an ideal tool for comparing two sets. In all of studies we compared 8-mers (parameter w=8) and bits/column bound was set to 1.6 (parameter i=1.6)

Authors' contribution.

SM designed and carried out the computational experiments; PS developed the program and analyzed TransFac PWMs, A. Kel provided biological insight and actively participated in discussion of the project and writing the paper, A. Kondrashov and ID designed and led the project.

Acknowledgements

References

Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437 (7062), 1149-1152.

Bernat, J.A., Crawford, G.E., Ogurtsov, A.Y., Collins, F.S., Ginsburg, D., Kondrashov, A.S. (2006) Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Hum Mol Genet* 15 (13), 2098-2105.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. Glocal alignment: finding rearrangements during alignment. Bioinformatics. 2003;19 Suppl 1:i54-62.

Dermitzakis E.T., Reymond, A., and Antonarakis, S.E. (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nature Rev Genetics* 6 (2), 151-157

Dermitzakis, E.T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., Antonarakis, S.E. (2003) Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (CNGs). *Science* 302 (5647), 1033-1035

Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T., and Hirschhorn, J.N. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics* 38 (2), 223-227

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004, 32(Web Server issue):W273-9.

Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res* 16 (7), 875-884 JUL 2006

Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* 16 (9), 369-372

Hatini V., Huh S. O., Herzlinger D., Soares V. C., Lai E. Essential role of stromal mesenchyme in kidney morphogenesis revealed by targeted disruption of Winged Helix transcription factor BF-2 Genes Dev. 10:1467-1478 (1996).

Hwang, D.G., Green, P. (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* 101 (39), 13994-14001.

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.

Kondrashov, F.A., Ogurtsov, A.Y., Kondrashov, A.S. (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J Theor Biol* 240 (4), 616-626.

Kondrashov, A.S, Shabalina, S.A. (2002) Classification of common conserved sequences in mammalian intergenic regions *Hum Mol Genet* 11 (6), 669-674

Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921.

Levinson R. S., Batourina E., Choi C., Vorontchikhina M., Kitajewski J., Mendelsohn C. L. Foxd1-dependent signals control cellularity in the renal capsule, a structure required for normal renal development. Development 132:529-539 (2005).

Liu, Y.Y., Wei, L.P., Batzoglou, S., Brutlag, D.L., Liu, J.S., Liu, X.S. (2004) A suite of web-based programs to search for transcriptional regulatory motifs. *Nucl Acids Res* 32 (Suppl 2), W204-W207.

Margulies, E.H., M. Blanchette, D. Haussler, and E.D. Green, *Identification and characterization of multi-species conserved sequences.* Genome Res, 2003. **13**(12): p. 2507-18.

Matys, V., et al. (2006) TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. Nucl Acids Res 34 (Sp. Iss. SI), D108-D110.

Ni Y. G., Berenji K., Wang N., Oh M., Sachan N., Dey A., Cheng J., Lu G., Morris D. J., Castrillon D. H., Gerard R. D., Rothermel B. A., Hill J. A. Foxo transcription factors blunt cardiac hypertrophy by inhibiting calcineurin signaling. Circulation 114:1159-1168 (2006).

Peterson R. S., Linn L., Ye H., Zhou H., Overdier D. G., Costa R. H.
The winged helix transcriptional activator HFH-8 is expressed in the mesoderm of the primitive streak stage of mouse embryos and its cellular derivatives
Mech. Dev. 69:53-69 (1997).

Prabhakar, S., F. Poulin, M. Shoukry, V. Afzal, E.M. Rubin, O. Couronne, and L.A. Pennacchio, *Close sequence comparisons are sufficient to identify human cis-regulatory elements.* Genome Res, 2006. **16**(7): p. 855-63.

Shabalina, S.A., and Kondrashov A.S. (1999) Pattern of selective constraint in C. elegans and C. briggsae genomes. *Genetical Res* 74 (1), 23-30

Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A., and Kondrashov, A.S. (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* 17 (7), 373-376

Schreiber E., Matthias P., Mueller M. M., Schaffner W. Identification of a novel lymphoid specific octamer binding protein (OTF-2B) by proteolytic clipping bandshift assay (PCBA) EMBO J. 7:4221-4229 (1988).

Smith, A.D., Sumazin, P., Xuan, Z.Y., Zhang, M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression *Proc Natl Acad Sci USA* 103 (16), 6275-6280.

Smith, A.D., Sumazin, P., Zhang, M.Q. (2005) Identifying tissue-selective transcription factor bindings sites in vertebrate promoters *Proc Natl Acad Sci USA* 102 (5), 1560-1565.

Schoeler H. R., Balling R., Hatzopoulos A. K., Suzuki N., Gruss P. Octamer binding proteins confer transcriptional activity in early mouse embryogenesis EMBO J. 8:2551-2557 (1989).

Stegmaier, P., Kel, A., Wingender, E. (2007) m2transfac – a PWM-PWM alignment interface to the TRANSFAC database. *Nucl Acids Res. Web issue. In press.*

Walter, K., Irina Abnizova, I., Elgara, G., and Walter R. Gilks, W.R. (2005) Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends in Genetics* 21 (8), 436-440.

Wang T., Miura N., Carlsson P. Foxf1 and Foxf2 control murine gut development by limiting mesenchymal Wnt signaling and promoting extracellular matrix production. Development 133:833-843 (2006).

Waterston, R.H., et al. (2002) Initial sequencing and analysis of the mouse genome. *Nature* 420 (6915), 520-562.

1.   Shabalina, S.A. and A.S. Kondrashov, *Pattern of selective constraint in C. elegans and C. briggsae genomes.* Genet Res, 1999. **74**(1): p. 23-30.

2.   Dermitzakis, E.T., A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S.E. Antonarakis, *Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs).* Science, 2003. **302**(5647): p. 1033-5.

3.   Margulies, E.H., M. Blanchette, D. Haussler, and E.D. Green, *Identification and characterization of multi-species conserved sequences.* Genome Res, 2003. **13**(12): p. 2507-18.

4.   Drake, J.A., C. Bird, J. Nemesh, D.J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S.E. Antonarakis, E.T. Dermitzakis, and J.N. Hirschhorn, *Conserved noncoding sequences are selectively constrained and not mutation cold spots.* Nat Genet, 2006. **38**(2): p. 223-7.

5.   Halligan, D.L. and P.D. Keightley, *Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison.* Genome Res, 2006. **16**(7): p. 875-84.

6.   Andolfatto, P., *Adaptive evolution of non-coding DNA in Drosophila.* Nature, 2005. **437**(7062): p. 1149-52.

7.   Waterston, R.H., K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J. Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R. Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J.P. Vinson, A.C. Von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody and E.S. Lander, *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.

8.   Shabalina, S.A., A.Y. Ogurtsov, V.A. Kondrashov, and A.S. Kondrashov, *Selective constraint in intergenic regions of human and mouse genomes.* Trends Genet, 2001. **17**(7): p. 373-6.

9.   Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis, *Conserved non-genic sequences - an unexpected feature of mammalian genomes.* Nat Rev Genet, 2005. **6**(2): p. 151-7.

10.   Hardison, R.C., *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends Genet, 2000. **16**(9): p. 369-72.

11.   Smith, A.D., P. Sumazin, Z. Xuan, and M.Q. Zhang, *DNA motifs in human and mouse proximal promoters predict tissue-specific expression.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6275-80.

12.   Bernat, J.A., G.E. Crawford, A.Y. Ogurtsov, F.S. Collins, D. Ginsburg, and A.S. Kondrashov, *Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs.* Hum Mol Genet, 2006. **15**(13): p. 2098-105.

13.   Kondrashov, A.S. and S.A. Shabalina, *Classification of common conserved sequences in mammalian intergenic regions.* Hum Mol Genet, 2002. **11**(6): p. 669-74.

14.   Smith, A.D., P. Sumazin, and M.Q. Zhang, *Identifying tissue-selective transcription factor binding sites in vertebrate promoters.* Proc Natl Acad Sci U S A, 2005. **102**(5): p. 1560-5.

15.   Liu, Y., L. Wei, S. Batzoglou, D.L. Brutlag, J.S. Liu, and X.S. Liu, *A suite of web-based programs to search for transcriptional regulatory motifs.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W204-7.

16.   Stegmaier, P., A. Kel, and E. Wingender, *m2transfac - a PWM-PWM alignment interface to the TRANSFAC database.* Nucleic Acids Res., 2007(Web issue): p. In press.

17.   Walter, K., I. Abnizova, G. Elgar, and W.R. Gilks, *Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences.* Trends Genet, 2005. **21**(8): p. 436-40.

18.   Hwang, D.G. and P. Green, *Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 13994-4001.

19.   Kondrashov, F.A., A.Y. Ogurtsov, and A.S. Kondrashov, *Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites.* J Theor Biol, 2006. **240**(4): p. 616-26.

20.   Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T.

Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y.J. Chen, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

21.   Matys, V., O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender, *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.

22.   Peterson, R.S., L. Lim, H. Ye, H. Zhou, D.G. Overdier, and R.H. Costa, *The winged helix transcriptional activator HFH-8 is expressed in the mesoderm of the primitive streak stage of mouse embryos and its cellular derivatives.* Mech Dev, 1997. **69**(1-2): p. 53-69.

23.   Ormestad, M., J. Astorga, H. Landgren, T. Wang, B.R. Johansson, N. Miura, and P. Carlsson, *Foxf1 and Foxf2 control murine gut development by limiting mesenchymal Wnt signaling and promoting extracellular matrix production.* Development, 2006. **133**(5): p. 833-43.

24.   Hatini, V., S.O. Huh, D. Herzlinger, V.C. Soares, and E. Lai, *Essential role of stromal mesenchyme in kidney morphogenesis revealed by targeted disruption of Winged Helix transcription factor BF-2.* Genes Dev, 1996. **10**(12): p. 1467-78.

25.   Levinson, R.S., E. Batourina, C. Choi, M. Vorontchikhina, J. Kitajewski, and C.L. Mendelsohn, *Foxd1-dependent signals control cellularity in the renal capsule, a structure required for normal renal development.* Development, 2005. **132**(3): p. 529-39.

26.   Ni, Y.G., K. Berenji, N. Wang, M. Oh, N. Sachan, A. Dey, J. Cheng, G. Lu, D.J. Morris, D.H. Castrillon, R.D. Gerard, B.A. Rothermel, and J.A. Hill, *Foxo transcription factors blunt cardiac hypertrophy by inhibiting calcineurin signaling.* Circulation, 2006. **114**(11): p. 1159-68.

27.   Schreiber, E., P. Matthias, M.M. Muller, and W. Schaffner, *Identification of a novel lymphoid specific octamer binding protein (OTF-2B) by proteolytic clipping bandshift assay (PCBA).* Embo J, 1988. **7**(13): p. 4221-9.

28.   Scholer, H.R., R. Balling, A.K. Hatzopoulos, N. Suzuki, and P. Gruss, *Octamer binding proteins confer transcriptional activity in early mouse embryogenesis.* Embo J, 1989. **8**(9): p. 2551-7.

29.   Frazer, K.A., L. Pachter, A. Poliakov, E.M. Rubin, and I. Dubchak, *VISTA: computational tools for comparative genomics.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W273-9.

30.   Brudno, M., S. Malde, A. Poliakov, C.B. Do, O. Couronne, I. Dubchak, and S. Batzoglou, *Glocal alignment: finding rearrangements during alignment.* Bioinformatics, 2003. **19 Suppl 1**: p. i54-62.

31.   Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

32.   Prabhakar, S., F. Poulin, M. Shoukry, V. Afzal, E.M. Rubin, O. Couronne, and L.A. Pennacchio, *Close sequence comparisons are sufficient to identify human cis-regulatory elements.* Genome Res, 2006. **16**(7): p. 855-63.