



# Pyknons



Aleksander Sudakov

Journal Club 12.01.07

# Source

Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., Platt, D. 2006. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. PNAS 103:6605-6610

Meynert, A., Birney, E. 2006 Picking pyknons out of human genome. Cell 125:836-838.

# Pyknons

Greek adj. serried, dense, frequent

Motifs in transcribed areas with multiplicities above what is expected by chance

Present in almost all known human genes

# Pattern discovery

TEIRESIAS pattern discovery tool

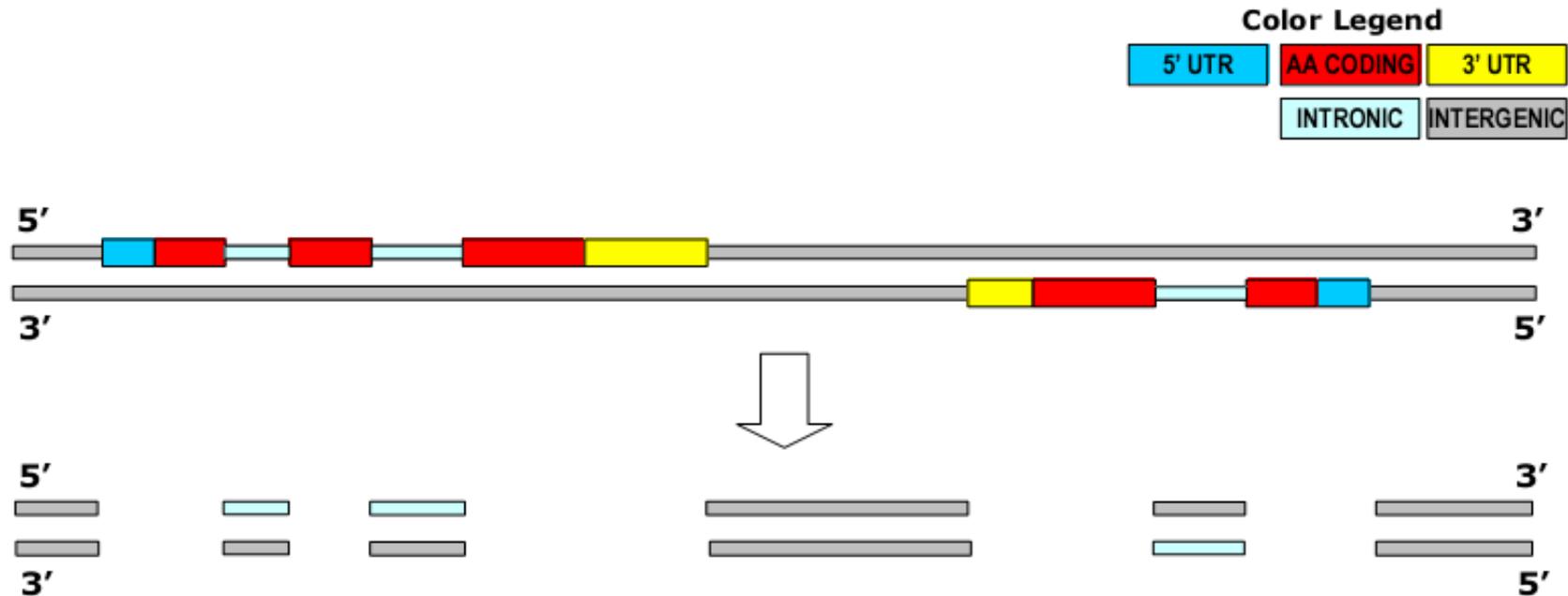
Search in intergenic and intronic DNA and reverse complement

$L = 16+ \text{ nt}$

min 40 times in genome

$P_{\text{init}} = 66 \text{ million patterns}$

# Nongenic regions



**Fig. 3. Preprocessing of the intergenic/intronic sequences prior to carrying out pattern discovery. All sequence segments that correspond to the untranslated and coding regions of known genes or that are the reverse complement of the untranslated and coding regions of known genes are removed. Top: genomic input before the preprocessing. Bottom: input on which pattern discovery will be run.**

# Patterns in UTR and CR

Generating and filtering pyknon collection:

Occurs at least once in genic regions

No collisions/overlaps with other patterns

Longer and more frequent patterns preferred

$$P_{\text{CR}} = 12267$$

$$P_{\text{3'UTR}} = 54396$$

$$P_{\text{5'UTR}} = 67544$$

$$P_{\text{CR}} \cup P_{\text{3'UTR}} \cup P_{\text{5'UTR}} = 127998 \text{ pyknons}$$

# Properties

16000 transcripts contain at least 4 pyknons, 2200 contain 20 or more.

Combinatory arrangement, mosaic.

Preferred distances between adjacent pyknons 18, 22, 24, 26, 29, 30, 31 nt

Associated with specific biological processes (Gene Ontology db)

# Combinatory arrangement



Fig. 1. Pyknons in the 3' UTRs of the apoptosis inhibitor *birc4* (shown above the horizontal line) and nine other genes. The sequences below the line contain some of *birc4*'s pyknons, but in different arrangements; they also contain instances of other pyknons that are not present in *birc4*'s 3' UTR. The 10 3' UTRs are pyknon mosaics. The shown pyknons, whether highlighted or in dark gray, have 40 or more instances in the genome's intergenic/intronic regions and additional copies in the untranslated and coding regions of these and other genes. We highlight only those pyknons that appear two or more times in the shown 3' UTRs. The light gray string -(xx)- indicates that xx nucleotides separate the pyknons that surround it. To appreciate the importance of this picture, it suffices to track the number of copies and relative position of TGCCTCCAGCCTGGG, TAATCCAGCAGCTTTGGGA, GGCTGAGGCAGGAGAAT, and GAGGTTGCAGTGAGCC.

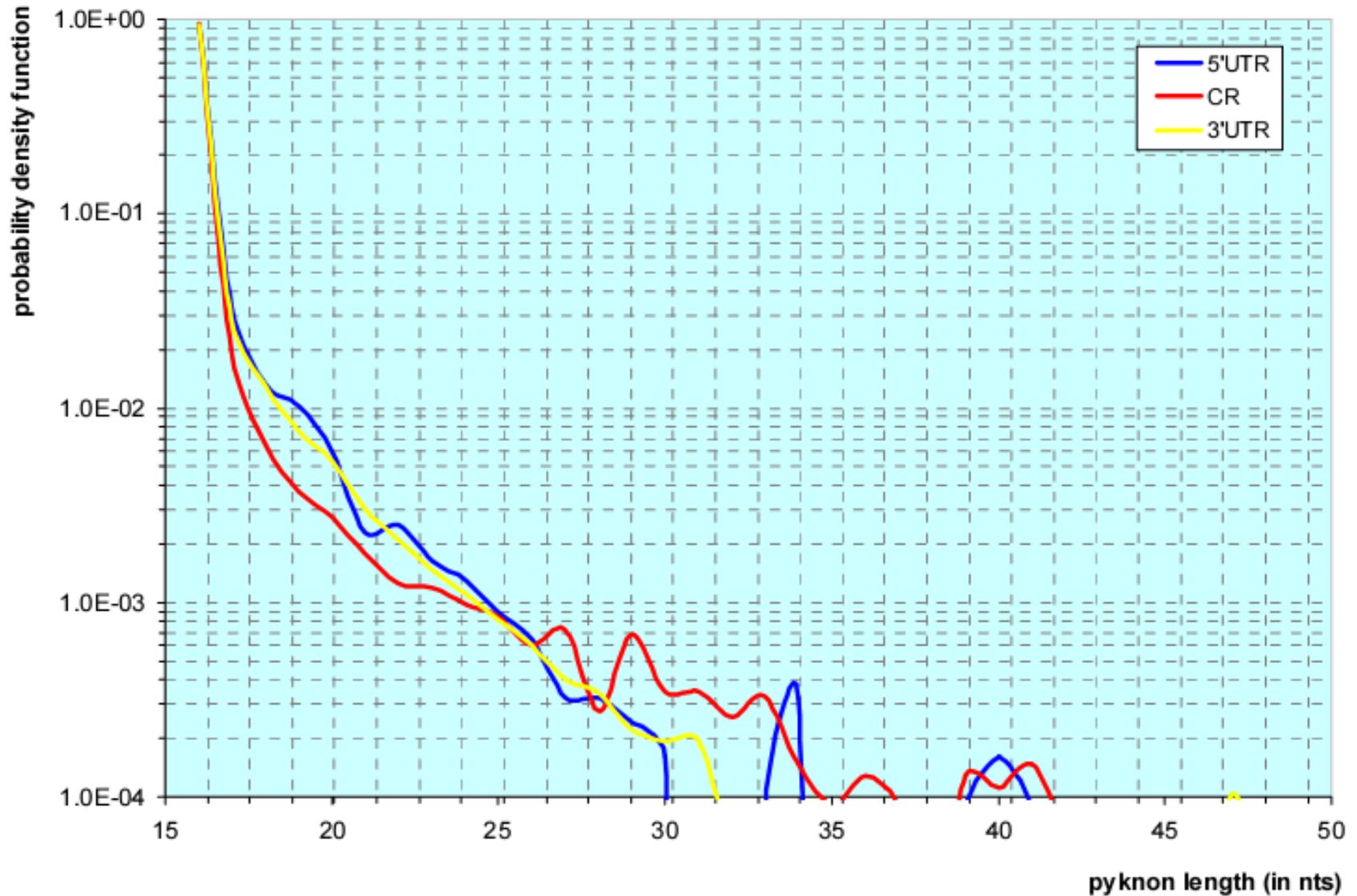


Fig. 9. The probability density function for the lengths of the pyknons—it is shown separately for each of the set  $P_{5'UTR}$  (blue),  $P_{CR}$  (red), and  $P_{3'UTR}$  (yellow).

**Fig. 2.** Probability density functions for the distance between the starting points of consecutive instances of pyknons, shown separately for 5' UTRs, CRs, and 3' UTRs. The distributions have long tails, and only a portion is shown. Note the peaks at  $x = 18, 22, 24, 26, 29, 30,$  and  $31$ .

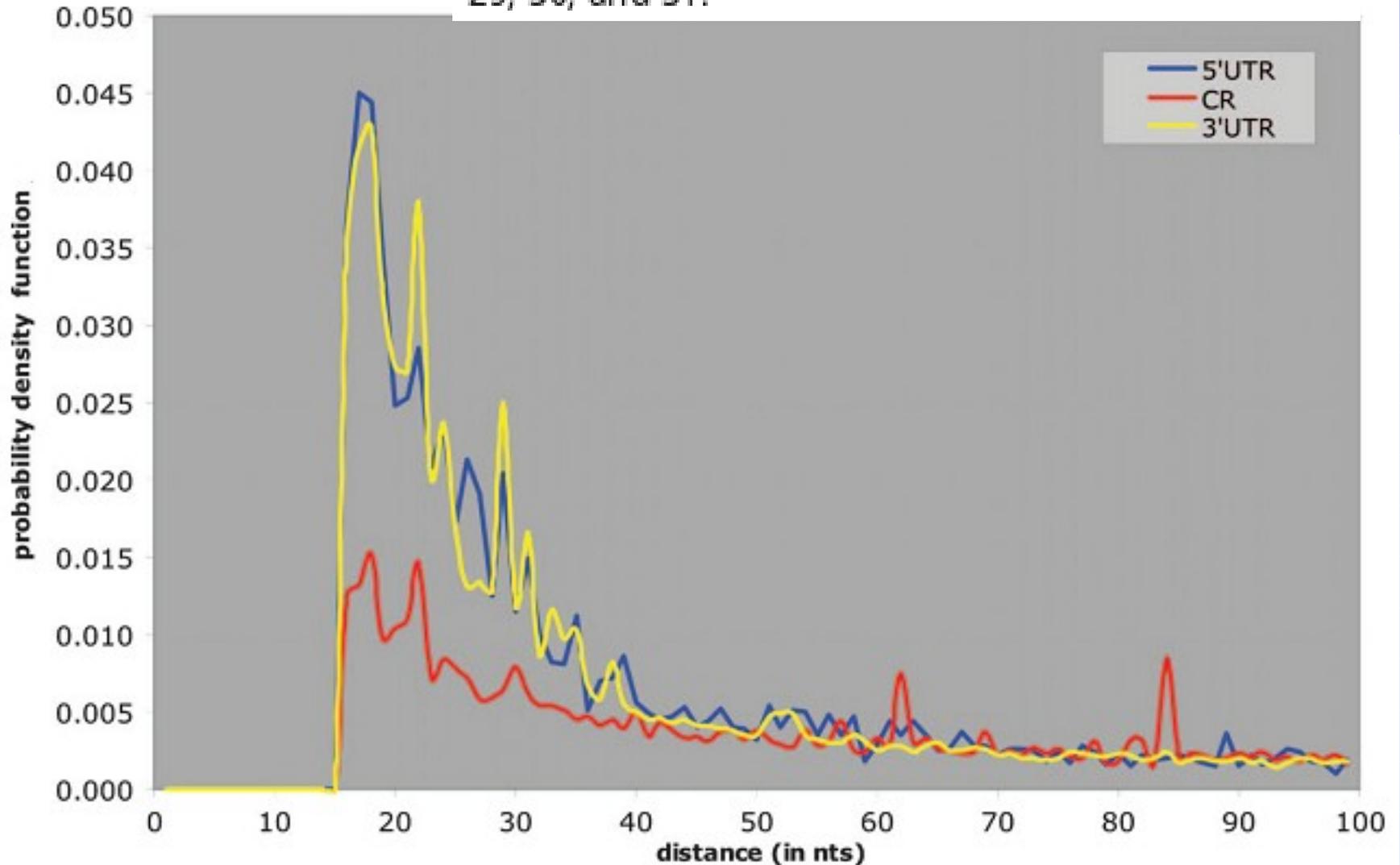


Table 1. Partial list of biological processes whose corresponding genes show significant enrichment (green cells) or depletion (red cells) in pyknon instances in their 5' UTR, CR, or 3' UTR

GO Term	5' UTR		Coding		3' UTR	
	ENRICHMENT DEPLETION	$ \log(P \text{ value}) $	ENRICHMENT DEPLETION	$ \log(P \text{ value}) $	ENRICHMENT DEPLETION	$ \log(P \text{ value}) $
DNA catabolism		19.88				23.05
muscle cell differentiation		27.11		28.08		
regulation of transcription, DNA-dependent		10.88		5.49		27.11
regulation of physiological process		8.66		5.14		22.11
nucleo-(base,side,tide)& nucleic acid metabolism		11.57		46.78		10.27
regulation of metabolism		9.58		8.95		27.33
DNA transposition				229.02		13.86
DNA metabolism		2.84		155.71		8.61
DNA replication				154.68		7.10
morphogenesis		3.27		4.43		9.93
organogenesis		6.58		2.34		9.26
defense response to bacteria		4.33		10.88		34.41
detection of external stimulus		5.18		13.97		15.72
organismal physiological process		3.06		15.84		11.86
response to external stimulus		7.91		34.01		23.25

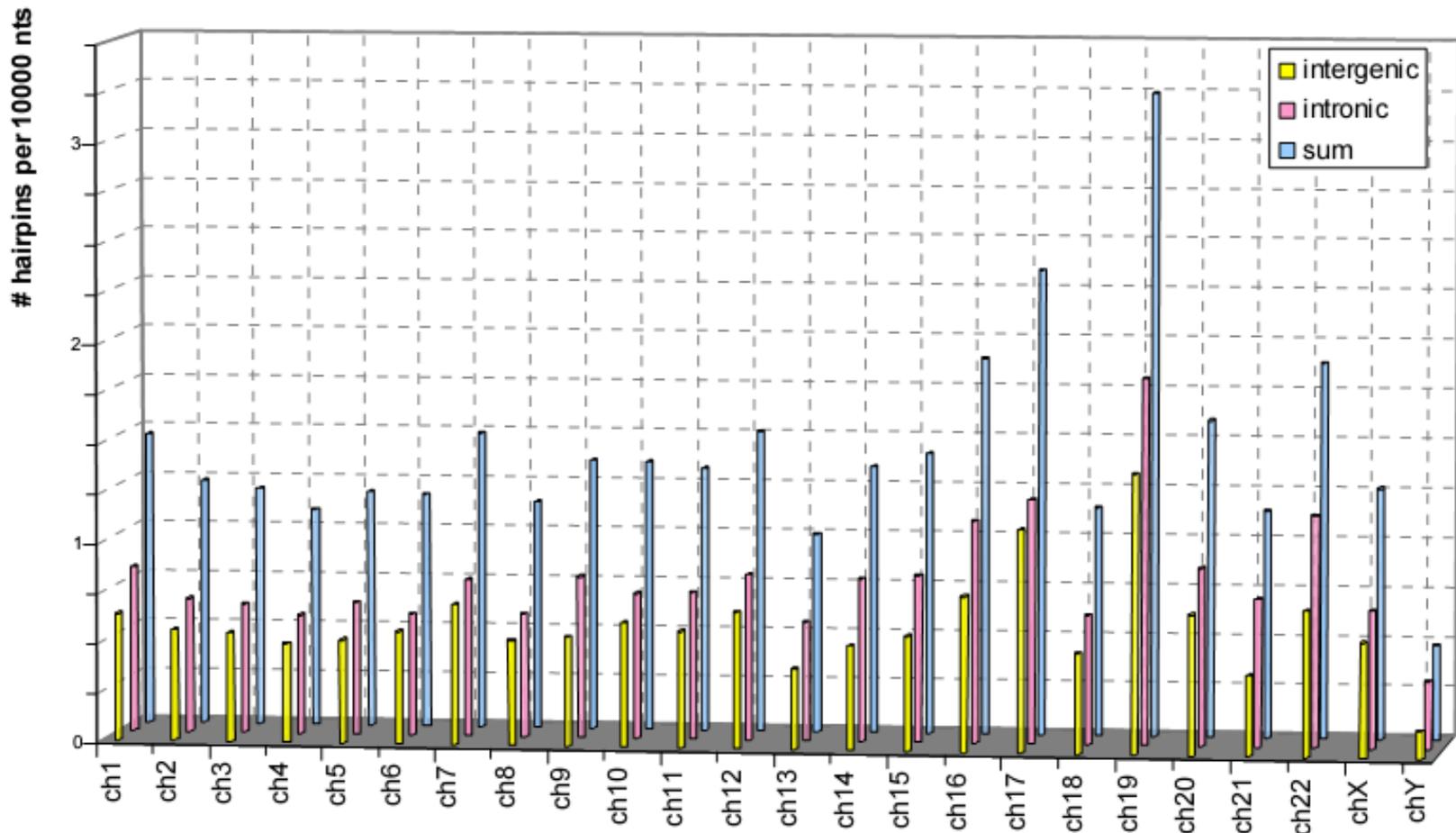
# Relation to microRNAs

Possibly linked to PTGS – post transcriptional gene silencing – 30% of pyknons may form hairpins in introns

120184 pyknons are at least 70% similar to known miRNAs

Related to 8-mer motifs in human, mouse, rat and dog 3'UTRs (overlap with 56 of 72 known motifs)

Distinct from ultraconserved elements



**Fig. 16.** Number of intergenic/intronic neighborhoods, generally 60 to 80 nucleotides in length, each of which contains the reverse-complement of a pyknon and is predicted to form double-stranded, energetically stable, hairpin-shaped RNA secondary structures—the number is given per 10,000 nts of sequence. Results are shown separately for each chromosome and for the intergenic (yellow) and intronic (pink) regions. The average number of hairpins per 10,000 nts independent of provenance is also shown in blue (“sum”).

# Cross-species conservation

**Table 2. Number of positions per 10,000 nucleotides that are covered by instances of the human pyknons**

Pattern set/ region	Positions covered in corresponding region of listed genome per 10,000 nucleotides						
	HSA	CFA	MMU	RNO	GGA	DME	CEL
$P_{5'UTR}/5'UTR$	382.2	43.0	20.7	14.2	9.7	22.4	5.8
$P_{CR}/CR$	304.1	88.3	57.4	61.1	28.4	25.2	14.7
$P_{3'UTR}/3'UTR$	733.3	152.4	82.6	64.9	42.1	65.9	57.1

For each of the three regions and each genome in turn we search region  $X$  of the genome with the patterns contained in the set  $P_X$ , where  $X = \{5' UTR, CR, 3' UTR\}$ . HSA, human; CFA, dog; MMU, mouse; RNO, rat; GGA, chicken; DME, fruit-fly; CEL, worm. See also text.

# Cross-species conservation

**Table 3. Number of human pyknons that are conserved in the human genome and the corresponding region of the  $j$ th genome for seven genomes and for each of 5' UTR, CR, and 3' UTR**

Genome	No. of human pyknons with instances in the corresponding region			Total size of intergenic/intronic region (both strands)	Intergenic/intronic positions (both strands) covered by	
	$P_{5'UTR}$	$P_{CR}$	$P_{3'UTR}$		all human pyknons	pyknons "in common"
HSA	12,267	54,396	67,544	6,093,304,675	692,393,548	692,393,548
MMU	400	8,767	6,160	5,216,777,897	89,568,584	45,996,326
RNO	170	3,424	1,644	5,409,179,291	82,635,080	25,134,158
CFA	234	6,170	1,351	4,826,002,769	87,572,989	7,912,193
GGA	51	1,786	718	1,855,717,211	9,262,198	577,232
DME	174	1,335	1,175	228,181,521	1,562,508	559,698
CEL	20	996	790	170,879,577	1,634,993	174,174

Shown is the number of intergenic/intronic positions in the  $j$ th genome that are covered by (i) all human pyknons, and (ii) only those human pyknons that are also present in the  $j$ th genome's 5' UTRs/CRs/3' UTRs. HSA, human; CFA, dog; MMU, mouse; RNO, rat; GGA, chicken; DME, fruit fly; CEL, worm. See also text.

# Conclusion

Links between coding and non-coding sequences in human genome

Pyknons enriched in UTR

Placement is not accidental and likely serves a specific purpose

Pyknon instances could be binding sites for small RNAs

Human pyknons in other organisms show enrichment in same types of genes yet are not the result of syntenic conservation

Dynamic view of genome, controlled response to environmental changes

# Repeats

1% of 127988 pyknons occurs exclusively inside repeat elements

79 pyknons never occur in repeat regions

98,9% occur both in repeat elements and in repeat-free regions

~40% have most instances in repeat-free regions

“Classically is expected that single instance of miRNA regulates a set of transcripts ... In contrast each pyknon is found more frequently in nongenic positions rather than inside transcripts”