European Journal of Human Genetics (2012) 20, 878–883
© 2012 Macmillan Publishers Limited All rights reserved 1018-4813/12
www.nature.com/ejhg

ARTICLE

Characterization of the intronic portion of cadherin superfamily members, common cancer orchestrators

Patrícia Oliveira[1,2], Remo Sanges[3], David Huntsman[4], Elia Stupka[5] and Carla Oliveira*[,1,6]

Cadherins are cell–cell adhesion proteins essential for the maintenance of tissue architecture and integrity, and their impairment is often associated with human cancer. Knowledge regarding regulatory mechanisms associated with cadherin misexpression in cancer is scarce. Specific features of the intronic-structure and intronic-based regulatory mechanisms in the cadherin superfamily are unidentified. This study aims at systematically characterizing the intronic portion of cadherin superfamily members and the identification of intronic regions constituting putative targets/triggers of regulation, using a bioinformatic approach and biological data mining. Our study demonstrates that the cadherin superfamily genes harbour specific characteristics in comparison to all *non-cadherin* genes, both from the genomic and transcriptional standpoints. Cadherin superfamily genes display higher average total intron number and significantly longer introns than other genes and across the entire vertebrate lineage. Moreover, in the human genome, we observed an uncommon high frequency of MIR (mammalian-wide interspersed repeats) and MaLR (mammalian-wide interspersed repeats, a subtype of LTR) regulatory-associated repetitive elements at 5′-located introns, concomitantly with increased *de novo* intronic transcription. Using this approach, we identified cadherin intronic-specific sites that may constitute novel targets/triggers of cadherin superfamily expression regulation. These findings pinpoint the need to identify mechanisms affecting particularly MIR and MaLR elements located in introns 2 and 3 of human cadherin genes, possibly important in the expression modulation of this superfamily in homeostasis and cancer.

European Journal of Human Genetics (2012) **20,** 878–883; doi:10.1038/ejhg.2012.11; published online 8 February 2012

Keywords: cadherin; cancer; intronic-based regulatory elements; MIR; MaLR; transcription

INTRODUCTION

Cadherins are transmembrane glycoproteins involved in biological functions from tissue morphogenesis to cancer.[1] All cadherin proteins share one or more copies of a 110-residue extracellular peptide (cadherin repeat-EC), responsible for mediating calcium-dependent homophilic/heterophilic cell-to-cell adhesion.[1] Many cadherin superfamily members have been proved or suggested to work as tumour suppressor genes and oncogenes in different cancer contexts. Changes affecting cadherin expression are of particular relevance in epithelial cancers, which constitute approximately 80–90% of all human cancers.[2] In this context, the non-homeostatic loss of cellular adhesion is frequently the master trigger for invasion and metastization.[3] Both genetic and epigenetic changes affecting cadherin genes can occur during carcinogenesis with the single purpose of disturbing cellular adhesion and allowing the escape of cancerous cells from primary tumours to more distant locations.[2,4]

E-cadherin (OMIM*192090), a classical tumour suppressor gene, possesses anti-invasive and anti-metastatic properties,[5] and the clinical turning point in carcinoma progression and metastasis is mediated by its disruption in 90% of all epithelial cancers.[5,6] Although classical gene inactivation (mutation, gene loss and promoter hypermethylation) and transcriptional and post-transcriptional mechanisms (transcription repressors, RNA and protein quality control) hamper normal E-cadherin expression and function,[5,7] these phenomena are insufficient to explain E-cadherin impairment both in the development and overall tumour progression. Regulation by non-coding RNAs (microRNAs), alternative transcripts (antisense transcripts) and alternative translated isoforms (antagonistic isoforms) have recently emerged as a new layer to explain gene and protein expression alterations,[8] nevertheless their impact on E-cadherin expression and function control in cancer is still poorly understood. A study by Stemmler *et al*[9] have shown that intron 2 of *CDH1* (ENSG00000039068), the gene encoding E-cadherin, entailed unknown regulatory sequences required to initiate transcriptional activation and to maintain its expression in mouse embryo differentiated epithelia, highlighting the importance of intronic elements in gene expression regulation.

Another interesting phenomena related with the members of the cadherin superfamily is the so-called cadherin switch, often observed in different types of cancer.[10,11] For example, E and/or P-cadherin expression is often replaced by N-cadherin in malignant breast and prostate cancers,[5] and desmocollin 2 expression is replaced by *de novo* expression of desmocollins 1 and 3 in colorectal cancer.[10] Regardless of the numerous observations in cadherin superfamily members' misexpression in cancer and other diseases, the mechanisms that control these effects are vastly unknown. Growing evidence attribute protein functional impairment to regulation by or at non-protein coding intronic and intergenic sequences.[12] Therefore, the aim of this study was to systematically characterize the intronic portion of cadherin

[1]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Rua Dr Roberto Frias, s/n, Porto, Portugal; [2]Instituto de Ciência Biomédicas Abel Salazar, Largo Professor Abel Salazar, 2, Porto, Portugal; [3]Stazione Zoologica Anton Dohrn, Villa Comunale, Napoli, Italy; [4]British Columbia Cancer Agency, Vancouver, British Columbia, Canada; [5]Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Via Olgettina 60, Milan, Italy; [6]Faculdade de Medicina da Universidade do Porto, Alameda Professor Hernâni Monteiro, Porto, Portugal
*Correspondence: Dr C Oliveira, Cancer Genetics, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Rua Dr Roberto Frias, s/n, Porto 4200-465, Portugal. Tel: +351 225 570 700; Fax: +351 225 570 799; E-mail: carlaol@ipatimup.pt
Received 25 August 2011; revised 20 December 2011; accepted 6 January 2012; published online 8 February 2012

superfamily members, in order to identify regions constituting putative targets of regulation. Using a bioinformatic approach, mining current gene annotation data,[13] as well as biological data from the ENCODE project,[14] we investigated genomic and transcriptional intronic-related features of this gene superfamily.

MATERIALS AND METHODS

Selection of cadherin and *non-cadherin* genes across species – the seven-domain approach

Using the Ensembl database (v52, v60)[15] and *InterPro* annotation of protein domains,[16] a keyword search was performed in order to obtain all domains that are present within cadherins from six species that have an extensive genomic infrastructure (such as expressed sequence tag (EST) and cDNA collections) and high coverage after genome sequencing/assembly: *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Danio rerio*, *Xenopus tropicalis* and *Ciona intestinalis*. These six species provide a general view on chordate evolution starting from *C. intestinalis* (commonly used to explore the evolutionary origins of the chordate lineage[17]) up to *H. sapiens*. The following keywords were used based on cadherin superfamily literature information: *cadherin*, *protocadherin*, *desmosome*, *desmoglein*. Seven domains were thus obtained: IPR000233; IPR002126; IPR006644; IPR009122; IPR013164; IPR014868; and IPR015919. Using PERL scripts and Ensembl API, we collected all non-repeated protein-coding genes, which contained at least one of the seven domains selected as well as both an ATG and STOP codons (data set A). The canonical transcript for each gene was selected based on three criteria in the following order: coding sequence length; number of exons; and transcript length (Supplementary Table 1). For each species, all protein-coding genes that were not selected as cadherins (absent from data set A, given the total absence of the seven *InterPro* domains selected) were classified as *non-cadherin* genes, which were used as reference for comparisons with the cadherin superfamily. The same selection criteria were used to select a canonical transcript for each *non-cadherin* gene. A new cadherin data set (data set B) of genes corresponding to proteins studied by van Roy and group[18] was built by manual identification (NCBI database) and using PERL scripts (to inquire Ensembl database).

Length class and longest *vs* non-longest division of introns

To calculate intron density as a function of intron length, all cadherin and *non-cadherin* introns extracted from Ensembl were partitioned into three length classes: introns were ranked according to their size and subsequently clustered, such that each length class possessed the same total nucleotide number (bp), which corresponds to one-third of the sum of the length of all introns extracted.[19] This ranking and clustering was done for three species given the extensive cDNA evidence (*H. sapiens*, *M. musculus*, *D. rerio*), and distinct length classes were obtained, given interspecies intronic variation (Supplementary Table 3). After ranking and clustering all introns from each species, each intron was sorted depending on being annotated to a cadherin or *non-cadherin* gene. All introns were also sorted within each gene depending on being the longest of all introns or one of the *non-longest* introns. Only genes with at least two introns were used in this part of the sorting analysis. Four groups were thus obtained: cadherin longest introns; cadherin *non-longest* introns; *non-cadherin* longest introns; and *non-cadherin non-longest* introns.

Non-coding sequence conservation analysis

To determine intronic sequence conservation within cadherin and *non-cadherin* introns, all available data on mammalian conservation from the Ensembl database[15] was used. The frequency of genomic evolutionary rate profiling (GERP) constrained elements across 12 mammalian species was analysed to inquire intronic sequence conservation (obtained using Ensembl's *Enredo-Pecan-Ortheus* pipeline on 12 eutherian mammals[20]). Cadherin and *non-cadherin* introns from each of the three length classes described were analysed separately. Moreover, both all cadherin and *non-cadherin* introns and only the longest introns were analysed. A data set was also produced by randomly collecting 1000 sets of 1000 *non-cadherin* introns, which were also analysed in terms of sequence conservation across mammalians (random data set). For each length class, an average frequency of GERP constrained elements was computed and plotted. For the random data set, the average frequency of GERP constrained elements added or subtracted by 3 SDs was calculated and plotted, thus allowing graphical assessment of significant differences in mammalian conservation.

Total intron number analysis

With Ensembl API and PERL scripts, the total number of introns in each selected canonical transcript was extracted from the six species selected previously (cadherins and *non-cadherins*). The resulting distributions were plotted and compared using the Wilcoxon rank-sum test[21] and the *P*-values corrected using the Bonferroni correction. Three adhesion-related families were selected following the same strategy as for cadherin data set A: ADAM[22] (IPR006586-44 genes); integrins[23] (INTs, IPR000413-19 genes); and tyrosine kinases[24] (TKs, IPR008266-99 genes). Three families unrelated with cellular adhesion were also selected: ARF[25] (IPR006688-106 genes); MHC[26] (IPR011162-90 genes); and POU[27] (IPR013847-102 genes). From within all adhesion-related genes (n=162 genes, derived from ADAM, INT and TK families), adhesion-unrelated genes (n=298 genes, derived from ARF, MHC and POU families), 1000 randomized sets, of 104 genes each, were selected (because data set A included 104 genes). The total intron number of each gene from each family/data set was extracted using the Ensembl[15] database. For the randomized data sets, we then calculated the overall average intron number, the minimum, maximum and average inter-quartile range (IQR).[28]

Intron length analysis across cadherin and non-cadherin introns

With Ensembl API and PERL scripts we extracted the length of all introns present in each selected canonical transcript from six selected species. Cadherin (data set A) and *non-cadherin* introns were separated according to the three length classes described and length status (longest and *non-longest*). The number of introns in each group (length class and length status) was computed. The number of introns in each group was compared (cadherins longest introns *vs non-cadherins* longest introns; cadherins *non-longest* introns *vs non-cadherins non-longest* introns). The Test of Equal or Given Proportions[21] was performed and the *P*-value corrected using the Bonferroni correction. The length of the longest intron and the average length of all *non-longest* introns were calculated for each cadherin (data set A) and *non-cadherin* gene. The distributions obtained for the lengths of longest introns from cadherin and *non-cadherin* genes were compared. The same was done for the average lengths of *non-longest* introns. The comparisons were done using the Wilcoxon rank-sum test,[21] and were performed for all cadherin and *non-cadherin* introns from the six species queried. The average ratio between longest introns from cadherins and *non-cadherins* was calculated: (1) all average lengths for longest introns obtained previously were summed and divided by the total number of cadherin longest introns annotated; (2) the same was done for *non-cadherin* introns; (3) these two values were divided thus obtaining the overall ratio between cadherin and *non-cadherin* longest introns. The same was done for *non-longest* introns.

Analysis of the position of the longest intron

Using PERL scripts and the Ensembl database (v52, v60), the position of the longest intron was assessed in both cadherin (data set A) and *non-cadherin* genes: ie, position 1 corresponds to the first intron in the gene (most 5′) immediately after exon 1. The Wilcoxon rank-sum test[21] was selected to compare the obtained intron position distributions. The same was done only for length class 3 longest introns.

Analysis of the frequency of regulatory elements and repetitive elements

Several regulatory and repetitive elements were analysed in terms of frequency. Data for both these types of elements were obtained using the Ensembl database (v52, elements inquired in Supplementary Tables 8 and 9). The analysis was done by comparing the frequency of base pairs overlapping each feature studied in human cadherin and *non-cadherin* introns as well as with the random data set of introns described previously. Introns were studied separately according to the three length classes described earlier. Moreover, introns were studied as a whole as well as separated into longest and *non-longest*.

Intron length, repetitive elements and CAGE data analysis

Introns from each of assessed position were analysed in terms of length and number of *Alu*, *MIR* and *MaLR* elements normalized to intron length of origin. Introns were also assessed in terms of intronic transcription, using the available data from the CAGE experiments within the *ENCODE* project[14,29,30] (files analysed are described in Supplementary Table 10). We focused on CAGE data collected using the 'normal' lymphoblastoid cell line GM12878: long poly-A-negative RNA from both genomic strands and from the nucleic and cytosolic fractions. For each intron position, cadherin and *non-cadherin* CAGE tag distributions were computed, all of which normalized to intron length of origin, thus avoiding intron length–related bias. The obtained distributions were compared using the Wilcoxon rank-sum test[21] and *P*-values corrected using the Bonferroni correction.

RESULTS

Selection of cadherin superfamily members

To characterize the genetic architecture of as many members as possible belonging to the cadherin superfamily, we generated criteria that could successfully gather all genes within this superfamily, using the Ensembl database.[13,15] We combined literature data and protein domain databases (*InterPro*[16]) to collect genes that coded for protein sequences currently classified as cadherins from six distinct genomes, resulting in seven distinct cadherin-related *InterPro* domains. Selected genes were further curated to assure that only protein-coding genes were selected and included in data set A (Supplementary Table 1). All protein-coding genes in each genome that were not selected for the cadherins' data set A were grouped in a control data set named *non-cadherin*s. To validate data set A, we compared it with a previously published data set (data set B) based on cadherin superfamily protein-related information.[18] For the *H. sapiens* genome, 90% of cadherins in data set B matched those of data set A. The remaining 10%, corresponded to (1) four protein sequences whose genomic locus we were not able to identify and; (2) four protein sequences that either our approach did not consider to be a cadherin (ie, it did not possess any of the seven *InterPro* domains selected) or were excluded from data set A given the absence of an annotated ATG/STOP codon. Five of the six genomes analysed, with the exception of *X. tropicalis*, produced a high overlap between data sets A and B (Supplementary Table 2).

Cadherin introns do not exhibit significant intronic sequence conservation

We investigated whether the DNA sequence of cadherin introns was being conserved across vertebrate species. We compared the intronic sequence conservation (*Genomic Evolutionary Rate Profiling* constrained elements) across 12 mammalian species[15,31] by studying 3 separated groups of human introns: (1) cadherin introns (data set A); (2) *non-cadherin* introns; and (3) 1000 data sets of 1000 randomly chosen *non-cadherin* introns (random data set). Sequence conservation was analysed by separating all cadherin and *non-cadherin* introns into three length classes, each encompassing the same number of nucleotides.[19] Length classes were named 1–3, with the latter encompassing the longer introns in each genome (Supplementary Table 3). By analysing separately each intron length class, we aimed at preventing any length-related bias. In addition, given that most genes in several genomes present their longest intron at the most 5′-end of the gene and that 5′-introns frequently encompass important non-coding regulatory elements,[32] we also analysed the sequence conservation of the longest intron from each cadherin and *non-cadherin* genes. We observed that cadherin introns' frequency of constrained elements was never significantly distinct from that of *non-cadherin* introns. This was valid for all length classes and for longest introns. In addition, no significant differences were observed in terms of intronic sequence conservation when comparing cadherin introns with the collected random data sets (Supplementary Figure 1).

Cadherins show a higher average intron number than non-cadherins as well as a particular pattern of intron distribution

Using the total intron number as a measure of gene architecture, we have observed that overall, the cadherin superfamily of genes presented a higher average intron number than *non-cadherin* genes in all species analysed (>11 introns for cadherins and <9 for *non-cadherins*, *P*-value ranging from 2.76E−08 to 1.15E−02, Figure 1, Supplementary Figure 2 and Supplementary Table 4). Moreover, the distribution of total intron number was wider in comparison to *non-cadherin* genes and particularly prominent in human and mouse genomes (Figure 1 and Supplementary Figure 2). The analysis of the

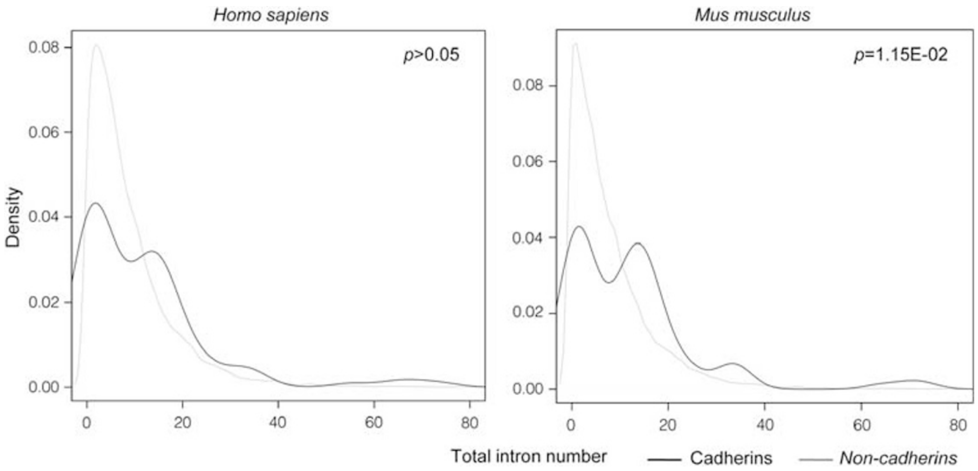


Figure 1 Total intron number distribution in human and mouse *non-cadherin* genes and cadherin superfamily genes. The grey full line represents *non-cadherin* genes and the black full line represents cadherin genes (data set A). Values in the top right corner correspond to *P*-values obtained when comparing data set A against *non-cadherins*.

distribution of total intron number revealed that while *non-cadherins* displayed a single and highly enriched peak corresponding to genes with less than 10 introns, the intron number distribution of *cadherin* genes resulted in four distinct peaks (human and mouse genomes): peak 1 corresponded to <10 introns; peak 2 corresponded to 12–20 introns; peak 3 corresponded to 30–40 introns; and peak 4 corresponded to 50–80 introns (Figure 1). Concerning *cadherin* genes found in non-mammalian species (except *C. intestinalis*), the peaks observed were less distinct than in mammals, yet still present and the total intron number in cadherins was also significantly distinct from *non-cadherin* genes (*P*-value ranging from 1.46E−03 to 2.76E−08, Supplementary Figure 1 and Supplementary Table 4).

To understand whether the wider range of total intron number resulted mainly from the comparison of a single family of genes against all genes in the genome, we compared the total intron number in cadherins to that of other human gene families/sets: (1) adhesion-related ADAM,[22] INT[23,24] and TKs;[24] (2) adhesion-unrelated ARF,[25] MHC[26] and POU[27]); (3) 1000 randomized sets of 104 human genes selected from within all adhesion-related genes and from within all adhesion-unrelated genes. By comparing intron number distribution with the IQRs (which measures data dispersion[28]) from *cadherin* genes with those of the remaining families of genes, we observed that while cadherins displayed an IQR=15.3, adhesion-related families displayed an IQR$<$8.5, adhesion-unrelated families of genes displayed an IQR$\leq$3 and random data sets exhibited average IQRs not higher than 9.9 (Supplementary Table 5). These results showed that *cadherin* genes displayed a wider total intron number distribution in comparison with all other families/data sets, supporting our previous observation for comparison with *non-cadherin* genes, and proved that cadherins are in fact a particular family of genes in terms of total intron number distribution.

Cadherin genes have significantly longer introns than other genes in the vertebrate lineage

The assessment of cadherin genetic structure revealed that cadherin family members such as *CDH1*, *CDH2* and *CDH3* presented unusually large introns.[9,33] Given the growing evidence of functional regulatory features located within introns,[34,35] we analysed whether the cadherin superfamily of genes displayed significantly longer introns than other genes in the genomes of *H. sapiens*, *M. musculus* and *D. rerio*, using the three length classes previously described (Supplementary Table 3). We further analysed separately the longest intron from each gene and all other introns, classified as *non-longest*. This comparison showed a significant enrichment of all cadherin introns (both longest and *non-longest*) in length class 3 for all three genomes (*P* ranging from 1.32E−15 to 6.57E−07, Figure 2, Supplementary Figure 3 and Supplementary Table 6). All *non-cadherin* introns analysed (except *D. rerio non-cadherin* longest introns) in turn were significantly more present in length class 1 (*P* ranging from 1.32E−15 to 2.08E−06, Supplementary Table 6). These data indicate that cadherin introns are significantly longer than those of *non-cadherin* genes (regardless of the longest/*non-longest* status, data not shown) and that this feature is conserved, being therefore potentially relevant to the entire vertebrate lineage. We further calculated the average length of longest introns from cadherin and *non-cadherin* genes and computed a ratio (average length of longest cadherin introns divided by average length of longest *non-cadherin* introns). The same was done for *non-longest* introns. The distributions obtained were compared, and it revealed that five of six species displayed significantly higher ratios either for cadherin longest and *non-longest* introns (*P* ranging from 1.96E−11 to 4.92E−02, Supplementary Table 7). An average length ratio per species was also computed and plotted to ease visualization (Supplementary Figure 4). The higher length ratios for cadherin introns were observed across all vertebrate species analysed, but not in *C. intestinalis*, an invertebrate species, reinforcing that cadherin's intron length may be a pervasive vertebrate feature.

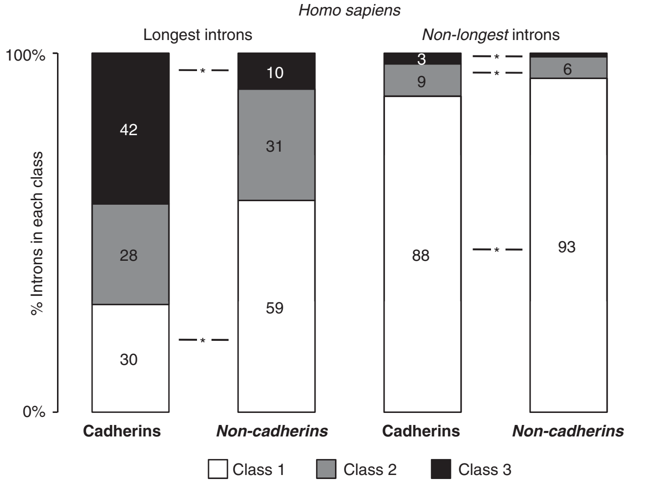


Figure 2 Distribution across length classes of human cadherin and *non-cadherin* longest and *non-longest* introns. White columns correspond to length class 1 introns; grey columns correspond to length class 2 introns; and black columns correspond to length class 3 introns. Data labels refer to the percentage of introns in each length class, type of gene and length status. Asterisks stand for significantly distinct comparisons ($P<0.05$).

Cadherin longest introns are positioned preferably at the 5′ start of the gene and are MIR and MaLR rich and Alu poor

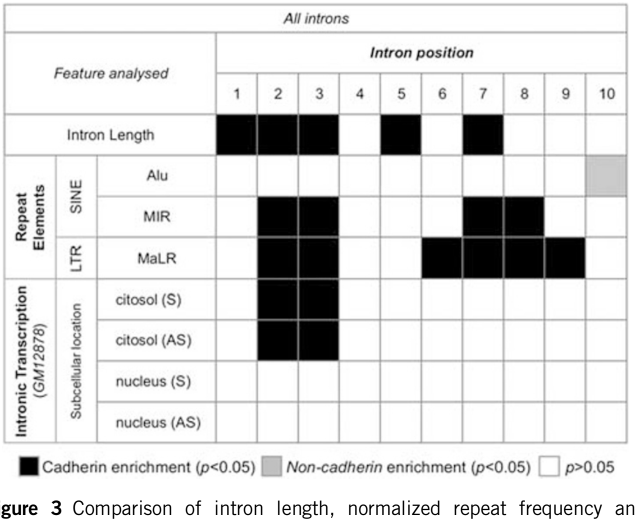
Prior reported studies indicated that in eukaryotes the first introns (positioned closer to the 5′-end of genes) tend to be longer,[32] and often harbour relevant regulatory elements essential for gene expression control.[19,32,34] Therefore, we next investigated the position of the longest intron of cadherins in comparison to the annotated transcription start site (*TSS*). Our data indicated that the majority of cadherins' longest introns (regardless of class) were positioned in the closest 5′-position in relation to the *TSS*, as were non-cadherin longest introns in all species analysed ($P>0.05$, Supplementary Figure 5 and data not shown). These observations pointed out that the position of the longest introns in cadherins followed the known trend in several vertebrate genomes inquired.

We next tested whether 5′-located long introns would accommodate important regulatory-associated annotated features, and determined the frequency of such features within cadherin introns. We analysed the presence of (1) DNAse1 hypersensitive sites, which mark for accessible chromatin;[36] (2) several histone methylation and acetylation marks, which commonly underlie promoter elements;[37] (3) ESTs, which mark for transcription;[38] and (4) DNA repetitive sequences, which have been shown to impact gene regulation.[39,40] We observed that, for all above-mentioned regulatory features, human cadherin introns displayed similar frequencies in comparison to all *non-cadherin* introns and to the random data set ($P>0.05$, data not shown), with the exception of DNA repetitive sequences. Human cadherin and *non-cadherin* introns revealed significant differences in terms of frequency of two specific families of repeats: short interspersed nuclear elements (SINE) and long terminal repeats (LTRs). SINE elements, in particular Alu and MIR (mammalian-wide interspersed repeats), were found to be present in a significantly distinct manner in cadherin introns: (1) Alu elements were significantly less frequent in cadherin introns than in *non-cadherin* introns (*P*=3.94E−14 and 1.14E−13 for length classes 1 and 2, respectively, Supplementary Figure 6A); (2) MIR elements were significantly more frequent in cadherin introns (*P*=1.38E−04 for length class 1, Supplementary Figure 6B) in comparison with *non-cadherin* introns. For LTR elements, a significantly higher frequency of MaLR (mammalian-wide interspersed repeats, a subtype of LTR) was observed in both cadherins' smaller and longer introns in comparison with *non-cadherin* introns (*P*=1.05E−03 and 3.57E−02 for length classes 1 and 3, respectively, Supplementary Figure 6C).

In order to integrate intron length and repeat frequency, we reanalysed all these features in terms of intron positioning. We observed that in terms of intron length, cadherin longest introns were significantly longer than *non-cadherin* longest introns found at positions 1, 2, 3, 5 and 7 (*P* ranging from 9.93E−09 to 3.92E−02, Figure 3). Next, we observed that both MIR and MaLR elements' frequency (normalized to intron length) was found to be significantly enriched in introns positioned closer to the 5′ start of cadherins in comparison with *non-cadherin* genes. In particular, cadherin introns in positions 2, 3, 7 and 8 were significantly enriched for MIR elements (*P* ranging from 2.30E−06 to 4.21E−02, Figure 3) and cadherin introns in positions 2, 3, 6, 7, 8 and 9 were significantly enriched in MaLR elements (*P* ranging from 1.00E−09 to 7.58E−03, Figure 3). Cadherin introns were significantly impoverished in Alu elements in position 10 in comparison with *non-cadherin* introns (*P*=3.99E−03, Figure 3).

Human cadherin introns exhibit significantly increased MIR and MaLR frequency with simultaneous increased intronic transcription

MIR and MaLR repeat elements are involved in genome novelty by alternative regulation phenomena as well as by promoting exonization.[39–41] Therefore, we next analysed whether the unusual frequency of Alu, MIR and MaLR elements within cadherin introns was correlated with differential intronic transcription arising from cadherin introns.



All introns												
Feature analysed			Intron position									
			1	2	3	4	5	6	7	8	9	10
Intron Length			■	■	■		■		■			
Repeat Elements	SINE	Alu										▨
		MIR		■	■				■	■		
	LTR	MaLR		■	■			■	■	■	■	
Intronic Transcription (*GM12878*)	Subcellular location	citosol (S)		■	■							
		citosol (AS)		■	■							
		nucleus (S)										
		nucleus (AS)										

■ Cadherin enrichment (p<0.05) ▨ *Non-cadherin* enrichment (p<0.05) □ p>0.05

Figure 3 Comparison of intron length, normalized repeat frequency and normalized intronic transcription between human cadherin and *non-cadherin* introns. Human cadherin and non-cadherin introns were separated according to their position in each corresponding gene (position 1 corresponds to the most 5′ intron). Black squares correspond to a significantly enriched feature for cadherin introns (*P*<0.05); grey squares correspond to a significantly enriched feature for non-cadherin introns (*P*<0.05); and white squares correspond a non-significantly distinct feature (*P*>0.05).

To assess cadherin intronic transcription, we mined the data from the pilot stage of the ENCODE project,[14] in particular data obtained by 5′ cap analysis gene expression (CAGE) performed by Carninci and group[29] at the RIKEN Institute. This technique allowed for the detection of new *TSS*, and we focused on the data obtained using RNA extracted from cytosolic and nucleic fractions of the lymphoblastic human normal cell line *GM12878* to assess transcription.[14] We verified that cadherin introns at positions 2 and 3 displayed a significant *TSS* enrichment in the cytosolic fraction of the cell line *GM12878*, arising from both RNA strands (*P* ranging from 1.51E−04 to 4.01E−02, Figure 3). Taken together, a correlation was observed for cadherin introns found at positions 2 and 3, for which cadherin introns were significantly longer, carried increased frequency of MIR and MaLR elements and increased levels of novel transcription initiation than *non-cadherin* introns (Figure 3).

DISCUSSION

The herein presented systematic characterization of the cadherin superfamily of genes encompassed the analysis of the intron sequence conservation; total intron number as a measure of gene structure; intron length and positioning; annotated regulatory and putative regulatory elements; and *de novo* intronic transcription initiation. We observed that this superfamily displays a higher average intron number than *non-cadherins*, a particular pattern of intron distribution, and significantly longer introns than the rest of the genome, throughout the vertebrate lineage, emphasizing an overall gene structure conservation in cadherins genes without intronic sequence conservation. This type of analysis has not been reported for other gene families and therefore it is difficult to extrapolate its implications. Nevertheless, the maintenance of many and large fragments of intronic DNA across evolution may indicate an unanticipated importance of these specific structural features. In fact, lack of sequence conservation does not imply lack of functionality of underlying elements and may rather derive from a rapid sequence evolution crucial for species adaptation.[12] We also observed that the unusually long introns were preferably located in the 5′ start of cadherin genes in line with what is currently described for overall genomes.[32] In fact, first introns are thought to encompass relevant (if not fundamental) regulatory elements.[19,32,34] This could therefore support the fact that cadherin introns' length and total intron number have been conserved during evolution to maintain essential underlying regulatory features.

The search for regulatory elements was only performed on the human genome due to the larger bioinformatic data availability (derived from projects such as ENCODE[14]) and revealed that neither the cadherins' intron sequence was being significantly conserved nor typical insulator/enhancer underlying elements, histone methylation marks or even accessible chromatin areas. Nevertheless, cadherin introns carried significantly more MIR and MaLR repetitive sequences and less Alu elements than all other genes in the human genome. All these repetitive elements are known to be involved in novel regulatory mechanisms and in exonization.[39,40] Given that the occurrence of exonization leads to the creation of new transcriptionally active regions, we analysed cadherin introns in terms of intron position and observed introns in positions 2 and 3 were not only longer but harboured a concomitant increased frequency of MIR, MaLR and CAGE tags in comparison with the rest of the human genome. This clearly suggests that the long cadherin introns at these two positions may in fact encode novel transcribed regulatory elements associated with MIR and MaLR repetitive elements. Supporting this observation, it has been shown that *CDH1* intron 2 entailed unknown regulatory sequences required to initiate transcriptional activation and to maintain its expression in mouse embryo differentiated epithelia.[9] In fact, several other studies have also revealed the relevance of the presence of introns (and consequently of putative intron-based elements) for basic biological processes such as mRNA stability and gene expression. In *Arabidopsis thaliana*, the removal of introns from the *ERECTA* gene leads to a dramatic decrease in its mRNA production, which is in addition much less stable and prone to degradation.[42] In the green alga *Ostreococcus lucimarinus*, a positive correlation between intron presence and increased gene expression has been observed.[43] Moreover, following the same idea, the few annotated intron-containing genes in yeast are responsible for almost one-third of all mRNA transcription,[44] and both human and plant intron-containing genes encode for more stable mRNA transcripts than genes without introns.[45]

In conclusion, our study demonstrates that the cadherin superfamily of genes harbours highly specific characteristics from the genomic and transcriptional standpoints, namely high frequency of specific repetitive elements within cadherin 5′-located long introns combined with an unusual frequency of novel transcription initiation. These findings lay the ground for discovering novel areas important in fine-tuning the expression of this gene family as well as intronic-based regulatory mechanisms, particularly in introns 2 and 3, important for expression of the cadherin superfamily of genes in biological events such as cadherin switching or cadherin gene loss/functional impairment in homeostasis and disease.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge the Portuguese Foundation for Science and Technology (FCT) for: project funding (PTDC/SAU-GMG/72168/2006, PTDC/SAU-GMG/110785/2009), PhD grant to PO (SFRH/BD/32984/2006) and salary support to Carla Oliveira from POPH – QREN/Type 4.2, European Social Fund and Portuguese Ministry of Science and Technology (MCTES). PO is a PhD student of the Graduate Program in Areas of Basic and Applied Biology (GABBA, 10th edition).

1 Halbleib JM, Nelson WJ: Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes Dev* 2006; **20**: 3199–3214.
2 Christofori G, Semb H: The role of the cell-adhesion molecule E-cadherin as a tumour suppressor gene. *Trends Biochem Sci* 1999; **24**: 73–76.
3 Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. *Cell* 2011; **144**: 646–674.
4 Nieto MA: Epithelial-mesenchymal transitions in development and disease: old views and new perspectives. *Int J Dev Biol* 2009; **53**: 1541–1547.
5 Berx G, van Roy F: Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harb Perspect Biol* 2009; **1**: a003129.
6 Vleminckx K, Vakaet Jr L, Mareel M, Fiers W, van Roy F: Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell* 1991; **66**: 107–119.
7 Cano A, Perez-Moreno MA, Rodrigo I *et al*: The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol* 2000; **2**: 76–83.
8 Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS: Non-coding RNAs: regulators of disease. *J Pathol* 2010; **220**: 126–139.
9 Stemmler MP, Hecht A, Kemler R: E-cadherin intron 2 contains cis-regulatory elements essential for gene expression. *Development* 2005; **132**: 965–976.
10 Khan K, Hardy R, Haq A, Ogunbiyi O, Morton D, Chidgey M: Desmocollin switching in colorectal cancer. *Br J Cancer* 2006; **95**: 1367–1370.
11 Tomita K, van Bokhoven A, van Leenders GJ *et al*: Cadherin switching in human prostate cancer progression. *Cancer Res* 2000; **60**: 3650–3654.
12 Pang KC, Frith MC, Mattick JS: Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* 2006; **22**: 1–5.
13 Flicek P, Aken BL, Ballester B *et al*: Ensembl's 10th year. *Nucleic Acids Res* 2010; **38**: D557–D562.
14 Consortium EP: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799–816.
15 Flicek P, Aken BL, Beal K *et al*: Ensembl 2008. *Nucleic Acids Res* 2008; **36**: D707–D714.
16 Hunter S, Apweiler R, Attwood TK *et al*: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009; **37**: D211–D215.
17 Dehal P, Satou Y, Campbell RK *et al*: The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 2002; **298**: 2157–2167.
18 Hulpiau P, van Roy F: Molecular evolution of the cadherin superfamily. *Int J Biochem Cell Biol* 2009; **41**: 349–369.
19 Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N, Pozzoli U: Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum Mol Genet* 2005; **14**: 2533–2546.
20 Hubbard TJ, Aken BL, Ayling S *et al*: Ensembl 2009. *Nucleic Acids Res* 2009; **37**: D690–D697.
21 R_Development_Core_Team: R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* 2008.
22 Edwards DR, Handsley MM, Pennington CJ: The ADAM metalloproteinases. *Mol Aspects Med* 2008; **29**: 258–289.
23 Zaidel-Bar R, Geiger B: The switchable integrin adhesome. *J Cell Sci* 2010; **123**: 1385–1388.
24 Zhao J, Guan JL: Signal transduction by focal adhesion kinase in cancer. *Cancer Metastasis Rev* 2009; **28**: 35–49.
25 Boman AL, Kahn RA: Arf proteins: the membrane traffic police? *Trends Biochem Sci* 1995; **20**: 147–150.
26 Clark DA, Chaouat G, Wong K, Gorczynski RM, Kinsky R: Tolerance mechanisms in pregnancy: a reappraisal of the role of class I paternal MHC antigens. *Am J Reprod Immunol* 2010; **63**: 93–103.
27 Andersen B, Rosenfeld MG: POU domain factors in the neuroendocrine system: lessons from developmental biology provide insights into human disease. *Endocr Rev* 2001; **22**: 2–35.
28 Upton G, Cook I: *Understanding Statistics*. Great Britain: Oxford University Press, 1996.
29 Kodzius R, Kojima M, Nishiyori H *et al*: CAGE: cap analysis of gene expression. *Nat Methods* 2006; **3**: 211–222.
30 Rhead B, Karolchik D, Kuhn RM *et al*: The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 2010; **38**: D613–D619.
31 Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005; **15**: 901–913.
32 Bradnam KR, Korf I: Longer first introns are a general property of eukaryotic gene structure. *PLoS One* 2008; **3**: e3093.
33 van Roy F, Berx G: The cell-cell adhesion molecule E-cadherin. *Cell Mol Life Sci* 2008; **65**: 3756–3788.
34 Mattick JS: The genetic signatures of noncoding RNAs. *PLoS Genet* 2009; **5**: e1000459.
35 Nagano T, Mitchell JA, Sanz LA *et al*: The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 2008; **322**: 1717–1720.
36 Sabo PJ, Hawrylycz M, Wallace JC *et al*: Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci USA* 2004; **101**: 16837–16842.
37 Reik W: Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 2007; **447**: 425–432.
38 Adams MD, Kelley JM, Gocayne JD *et al*: Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991; **252**: 1651–1656.
39 Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J: Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res* 2007; **17**: 1139–1145.
40 Piriyapongsa J, Polavarapu N, Borodovsky M, McDonald J: Exonization of the LTR transposable elements in human genome. *BMC Genomics* 2007; **8**: 291.
41 Prudhomme S, Oriol G, Mallet F: A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. *J Virol* 2004; **78**: 12157–12168.
42 Karve R, Liu W, Willet SG, Torii KU, Shpak ED: The presence of multiple introns is essential for ERECTA expression in Arabidopsis. *RNA* 2011; **17**: 1907–1921.
43 Lanier W, Moustafa A, Bhattacharya D, Comeron JM: EST analysis of *Ostreococcus lucimarinus*, the most compact eukaryotic genome, shows an excess of introns in highly expressed genes. *PLoS One* 2008; **3**: e2171.
44 Ares Jr M, Grate L, Pauling MH: A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* 1999; **5**: 1138–1139.
45 Wang HF, Feng L, Niu DK: Relationship between mRNA stability and intron presence. *Biochem Biophys Res Commun* 2007; **354**: 203–208.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)