Hum Genet (2006) 120: 1–21
DOI 10.1007/s00439-006-0180-7

REVIEW

Jian-Min Chen · Claude Férec · David N. Cooper

A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes I: general principles and overview

Received: 9 February 2006 / Accepted: 26 March 2006 / Published online: 28 April 2006

Abstract The 3′ regulatory regions (3′ RRs) of human genes play an important role in regulating mRNA 3′ end formation, stability/degradation, nuclear export, subcellular localization and translation and are consequently rich in regulatory elements. Although 3′ RRs contain only ~0.2% of known disease-associated mutations, this is likely to represent a rather conservative estimate of their actual prevalence. In an attempt to catalogue 3′ RR-mediated disease and also to gain a greater understanding of the functional role of regulatory elements within 3′ RRs, we have performed a systematic analysis of disease-associated 3′ RR variants; 121 3′ RR variants in 94 human genes were collated. These included 17 mutations in the upstream core polyadenylation signal sequence (UCPAS), 81 in the upstream sequence (USS) between the translational termination codon and the UCPAS, 6 in the left arm of the 'spacer' sequence (LAS) between the UCPAS and the pre-mRNA cleavage site (CS), 3 in the right arm of the 'spacer' sequence (RAS) or downstream core polyadenylation signal sequence (DCPAS) and 7 in the downstream sequence (DSS) of the 3′-flanking region, with 7 further mutations being treated as isolated examples. All the UCPAS mutations and the rather unusual cases of the *DMPK*, *SCA8*, *FCMD* and *GLA* mutations exert a significant effect on the mRNA phenotype and are usually associated with monogenic disease. By contrast, most of the remaining variants are polymorphisms that exert a comparatively minor influence on mRNA expression, but which may nevertheless predispose to or otherwise modify complex clinical phenotypes. Considerable efforts have been made to validate/elucidate the mechanisms through which the 3′ untranslated region (3' UTR) variants affect gene expression. It is hoped that the integrative approach employed here in the study of naturally occurring variants of actual or potential pathological significance will serve to complement ongoing efforts to identify all functional regulatory elements in the human genome.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-006-0180-7 and is accessible for authorized users.

J.-M. Chen · C. Férec
INSERM, U613, 29220 Brest, France

J.-M. Chen (✉) · C. Férec
Etablissement Français du Sang—Bretagne,
46 rue Félix Le Dantec, 29220 Brest, France
E-mail: Jian-Min.Chen@univ-brest.fr
Tel.: +33-2-98445064
Fax: +33-2-98430555

J.-M. Chen · C. Férec
Faculté de Médecine de Brest et des Sciences de la Santé,
Université de Bretagne Occidentale, 29238 Brest, France

D. N. Cooper
Institute of Medical Genetics, Cardiff University, Heath Park,
Cardiff CF14 4XN, UK

C. Férec
Laboratoire de Génétique Moléculaire et d'Histocompatibilité,
CHRU Brest, Hôpital Morvan, 29220 Brest, France

Introduction

Regulatory variants: definition and distribution

Regulatory variants may be defined as DNA sequence changes that have occurred within a gene's regulatory elements and which serve to alter either the expression of an mRNA transcript or one of its encoded isoforms (Pastinen and Hudson 2004). Although most known regulatory variants are found within promoters and 5′ untranslated regions (5′ UTRs) and may thus exert their effects by interfering with either transcription or translation (Pickering and Willis 2005), variants within the coding sequences and intronic regions are also known to affect allelic expression [e.g. coding sequence variants that affect splicing (Cartegni et al. 2002) or which lead to nonsense-mediated mRNA decay (Mendell et al. 2004), and intronic variants that modulate either splicing

(Baralle and Baralle 2005) or transcription factor binding (Liao et al. 2004)]. By contrast, the 3′ UTR has been a relatively neglected source of regulatory variants. This is perhaps surprising since not only is the average length of 3′ UTRs of protein-coding genes some 3–5 times longer than that of their 5′ counterparts (Mignone et al. 2002), but there is also increasing evidence to show that 3′ UTRs are particularly rich in regulatory elements. It therefore follows that 3′ UTRs are likely to be a fertile hunting ground for regulatory variants.

3′ UTRs: diverse roles in the regulation of gene expression

3′ UTRs are involved in regulating gene expression at multiple levels: at the pre-mRNA level, 3′ UTRs are involved in mRNA 3′ end formation and polyadenylation (Zhao et al. 1999) whereas at the mature mRNA level, 3′ UTRs determine such properties as mRNA stability/degradation, nuclear export, subcellular localization and translation efficiency (Conne et al. 2000; Mignone et al. 2002; Chabanon et al. 2004). These diverse regulatory roles are executed via *cis*-acting elements in the 3′ UTRs that interact with a multitude of *trans*-acting factors in a given cellular environment. Whereas the function of a DNA regulatory motif is essentially conferred by its primary structure, that of an RNA regulatory motif often relies on a combination of primary and secondary structure (Mignone et al. 2002).

A recent comparative analysis of the 3′ UTRs of some 17,700 human genes with their counterparts in the mouse, rat and dog genomes identified 106 highly conserved motifs in human 3′ UTRs, many of which may turn out to be microRNA (miRNA) targets (Xie et al. 2005). A subsequent study of the ‘transcriptional landscape’ of the mammalian genome has revealed that many non-protein-coding RNAs initiate from sites in the 3′ UTRs of protein-coding genes (Carninci et al. 2005). Further, 3′ UTRs have also been found to constitute preferred targets of *cis*-encoded natural antisense transcripts (Sun et al. 2005). These findings suggest that 3′ UTRs may play a rather more important role in regulating gene expression than perhaps has hitherto been appreciated.

How do the 3′ UTRs of protein-coding genes perform such diverse biological functions?

3′ UTRs are not under the same rigid structural constraints as promoter or coding sequences which are obliged to accommodate transcriptional or translational machinery (Conne et al. 2000). In other words, 3′ UTRs are subject to weaker purifying selection than coding and promoter sequences and have consequently been found to be less highly conserved in comparisons of orthologous gene regions (Makalowski et al. 1996). 3′ UTR sequences may thus have been freed up to evolve more rapidly in response to the need to specify the fates of different mRNA species in different cell types at different developmental stages. This postulate appears to be supported by at least two lines of evidence. On the one hand, not only do 3′ UTRs manifest a higher density of both short insertion/deletion length variants and single nucleotide polymorphisms (SNPs; Imanishi et al. 2004) but they also appear to have a greater tolerance of repeat expansions (Missirlis et al. 2005) as compared with 5′ UTRs and coding sequences. Furthermore, whilst the average length of 5′ UTRs appears to be roughly constant over diverse taxonomic classes (ranging between 100 and 200 nucleotides), that of 3′ UTRs is much more variable, ranging from ~200 nucleotides in plants and fungi to ~1,000 nucleotides in humans (Mignone et al. 2002).

Aims of this meta-analytical study

Given the importance of 3′ UTRs in regulating gene expression, Conne et al. (2000) suggested that the 3′ UTR might serve as a “molecular ‘hotspot’ for pathology”. Indeed, a number of examples of human diseases arising from anomalies within 3′ UTRs were already known at this time and Conne et al. (2000) suggested that a more systematic study would be likely to reveal the existence of other ‘such 3′ UTR-mediated diseases’. Noting that most of the 106 highly conserved motifs found in human 3′ UTRs did not match known regulatory motifs, Xie et al. (2005) opined that “the next challenge will be to develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion”. While such an approach is clearly warranted, we nevertheless feel that other complementary strategies should also be adopted. In this regard, we have performed a systematic analysis of the considerable number of disease-associated sequence variants that have now been identified in human gene 3′ regulatory regions (3′ RRs; see Fig. 1 for definition), with a view not only to cataloguing 3′ RR-mediated disease but also to gaining a greater understanding of the functional role of both the known and the putative regulatory elements that have been disrupted.

Data source and analysis

Data collection

Original publications reporting all the disease-associated 3′ RR variants collated in the Human Gene Mutation Database (HGMD; http://www.hgmd.org; Stenson et al. 2003) as of Jan. 31, 2006 were manually screened. Since the HGMD is a non-redundant mutation database, considerable efforts were made not only to check the original publications that first reported each variant, but also to follow up subsequent reports that provided either (1) functional data or (2) additional information relating

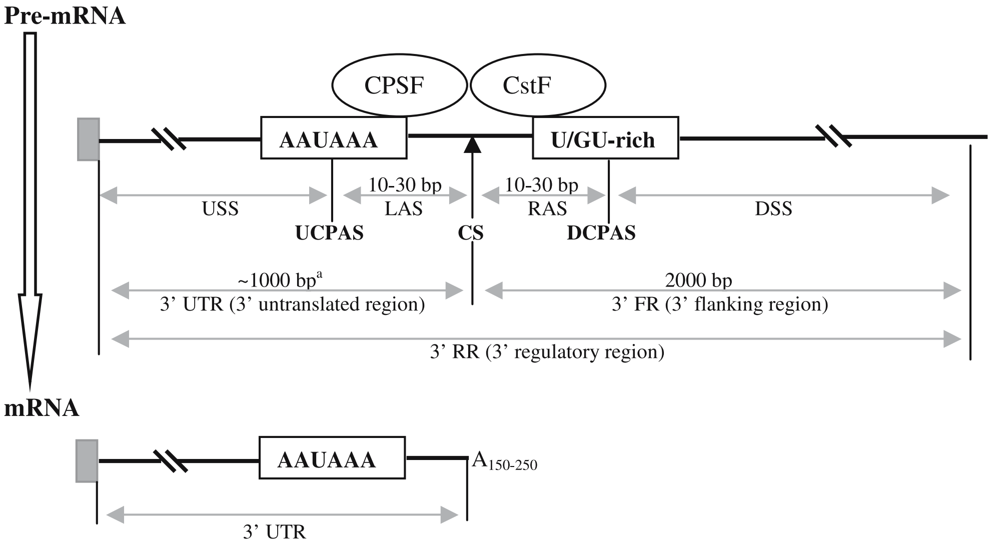

Fig. 1 Schematic representation of the primary and processed transcripts of an archetypal human protein-coding gene depicting the 3′ regulatory domains and motifs relevant to this study. *Shaded boxes*, coding region of the last exon [note that in some genes (e.g. *IL12*B and *PTGDS*), the last exon corresponds precisely to the entire 3′ untranslated region (3′ UTR)]; *USS*, upstream sequence between the translational termination codon and the upstream core polyadenylation signal (UCPAS) exemplified by the most frequently occurring AAUAAA hexamer; *LAS and RAS*, left and right arms of the 'spacer' sequence between the UCPAS and the downstream core polyadenylation signal (DCPAS) exemplified by the best described U/GU-rich element; *CS*, cleavage site or poly(A) addition site (indicated by an *upward pointing arrow*). The 3′ flanking region (3′ FR) was arbitrarily designated as the 2 kb region 3′ to the pre-mRNA CS. *DSS*, downstream sequence of the 3′ FR. The 3′ regulatory region (3′ RR) and 3′ UTR are also indicated. The *trans*-acting factors, cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulating factor (CstF) that specifically bind to UCPAS and DCPAS, respectively, are also illustrated. [a]Average length of complete human 3′ UTRs in accordance with Mignone et al. (2002)

to a putative disease association. This was achieved by a combination of a keyword search in PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = PubMed&itool = toolbar) and the perusal of articles citing a given publication (reference citations for some papers are available in the linked-ISI Web of Science and other citation tracking systems but may also be traced through Google™ Advanced Scholar Search (http://scholar.google.com/advanced_scholar_search)). As a result of this process, some additional 3′ RR variants that had not originally been logged in HGMD were identified; these were also included for the purposes of analysis. It should also be noted that the term 'disease-associated' as used in this study is quite loose; some of the collated variants may be disease-causing (as in the case of monogenic disorders), disease-predisposing (as in the case of polygenic or complex disorders), disease-modifying (i.e. the variant itself does not cause or predispose to disease but rather affects the severity, age of onset or clinical outcome of the disease), or even 'disease-protective', whereas other variants may only serve as disease markers. Together, these efforts yielded a total of 121 variants in the 3′ RRs of 94 human protein-coding genes (Table 1). Although this dataset should not be regarded as an exhaustive list of reported 3′ RR variants, it is nevertheless likely to be the most comprehensive available. Based upon the data currently logged in HGMD, we estimate, rather conservatively, that at least 0.2% of disease-associated mutations reside within 3′ RRs.

Classification of 3′ RR variants with respect to their locations

The collated variants were classified into distinct categories depending upon their respective locations within the 3′ RRs (Fig. 1). This classificatory scheme is discussed below.

Function-based domain and motif definition of the 3′ RR

The 3′ ends of all fully processed eukaryotic mRNAs [with the exception of most histone mRNAs] have a poly(A) tail (Gilmartin 2005). The poly(A) tail is synthesized by poly(A) polymerase, has a length ranging in mammalian mRNAs from ~150 to 250 nucleotides [See Supplementary Online Text for mRNAs with < 20-nt poly(A) tails] and is critical for the regulation of mRNA stability and translation, probably by providing a binding site for poly(A)-binding proteins (Coller and Parker 2004; Kuhn and Wahle 2004).

The cellular process that adds a poly(A) tail to the maturing mRNAs, known as polyadenylation, proceeds in two steps: the endonucleolytic cleavage of pre-mRNA and the subsequent addition of poly(A) tail to the newly formed 3′ end (Zhao et al. 1999). In mammalian cells, three elements define the core polyadenylation signal: (1) the highly conserved AAUAAA hexamer or a close variant thereof (termed the upstream core

Table 1 Catalogue of 3′ regulatory region (3′ RR) variants analysed in this study

Disease(s)/trait(s) under investigation[a]	Gene	Variant[b]	Main reference(s)
UCPAS[c] *(n = 17)*			
Arylsulphatase A pseudodeficiency	*ARSA*	AATAAC > AGTAAC	Gieselmann et al. (1989)
Immune dysfunction, polyendocrinopathy, enteropathy, X-linked	*FOXP3*	AATAAA > AATGAA	Bennett et al. (2001)
α^+ thalassaemia	*HBA2*	AATAAA > AATAAG	Higgs et al. (1983)
α^+ thalassaemia	*HBA2*	AATAAA > AATGAA	Yuregir et al. (1992)
α^+ thalassaemia	*HBA2*	AATAAA > AATA	Harteveld et al. (1994)
α^+ thalassaemia	*HBA2*	16 bp deletion including the 1st base of UCPAS	Tamary et al. (1997)
β-thalassaemia	*HBB*	AATAAA > AACAAA	Orkin et al. (1985)
β-thalassaemia	*HBB*	AATAAA > AATGAA	Jankovic et al. (1990)
β-thalassaemia	*HBB*	AATAAA > AATAGA	Jankovic et al. (1990)
β-thalassaemia	*HBB*	AATAAA > AATAAG	Rund et al. (1992)
β-thalassaemia	*HBB*	AATAAA > A	Rund et al. (1992)
β-thalassaemia	*HBB*	AATAAA > AAAA	Kimberland et al. (1995)
β-thalassaemia	*HBB*	AATAAA > GATAAA	Waye et al. (2001)
β-thalassaemia	*HBB*	AATAAA > AAAAAA	Jacquette et al. (2004)
β-thalassaemia	*HBB*	AATAAA > AATATA	Giordano et al. (2005)
Insulin-like growth factor 1 deficiency	*IGF1*	AATATA > AAAATA	Bonapace et al. (2003)
X-linked severe combined immunodeficiency	*IL2RG*	AATAAA > AATAAG	Hsu et al. (2000)
LAS (n = 6)			
Gene expression	*ADIPOR1*	C > G polymorphism	Wang et al. (2004)
Venous thrombosis	*F2*	G20210A	Poort et al. (1996)
Venous thrombosis; stroke; pregnancy complications	*F2*	C20209T	Warshawsky et al. (2002), Schrijver et al. (2003), Itakura et al. (2005), Soo et al. (2005), Wylenzek et al. (2005)
Unknown	*F2*	A20207C	Meadows et al. (2002), Ceelie et al. (2005)
β-thalassemia intermedia	*HBB*	TTGCA > CTGAA	Heath et al. (2001)
Insulin resistance	*RETN*	ATG repeat polymorphism	Pizzuti et al. (2002)
USS (n = 81; J.M. Chen, C. Férec, D.N. Cooper, submitted)			
RAS or DCPAS (n = 3)			
Intrarenal segmental arterial thrombosis; Budd-Chiari syndrome; pregnancy complications	*F2*	C20221T	Wylenzek et al. (2001), Balim et al. (2003), Schrijver et al. (2003)
Unknown	*F2*	A20218G	Meadows et al. (2002), Ceelie et al. (2005)
Venous thrombosis	*FGG*	10034C > T polymorphism	Uitte de Willige et al. (2005)
DSS (n = 7)			
Graves' disease; type-1 diabetes; autoimmune hypothyroidism	*CTLA4*	G > A polymorphism	Ueda et al. (2003), Anjos et al. (2004), Furugaki et al. (2004), Zhernakova et al. (2005)
Lung and oral cancers	*CYP1A1*	C > T polymorphism	e.g. Kawajiri et al. (1990), Hayashi et al. (1991), Sato et al. (1999), Tanimoto et al. (1999)
Type 2 diabetes	*FABP3*	Single T deletion polymorphism	Shin et al. (2003)
δ-thalassaemia	*HBD*	G > A mutation	Moi et al. (1992)
Airway hyperresponsiveness	*KCNS3*	C > T polymorphism	Hao et al. (2005)
Chronic respiratory disease	*SERPINA1*	G > A polymorphism	Morgan et al. (1993)
Psoriasis?	*SLC9A3R1*	SNP9	See text
Isolated examples (n = 7)			
Myotonic dystrophy, type 1	*DMPK*	CTG repeat expansion	Brook et al. (1992)
Fukuyama-type congenital muscular dystrophy	*FCMD*	SVA insertion	Kobayashi et al. (1998)
Fabry disease	*GLA*	1277delAA	Yasuda et al. (2003)
Fabry disease	*GLA*	1284delACTT	Yasuda et al. (2003)
Recurrent spontaneous abortion; pre-eclampsia; and outcome of in vitro fertilization	*HLA-G*	14 bp insertion polymorphism	See text
Spinocerebellar ataxia type 8	*SCA8*	CTG repeat expansion	Koob et al. (1999)
Attention-deficit hyperactivity disorder	*SLC6A3*	40-bp variable number of tandem repeat polymorphism	See text

[a] Some of the disease- or trait-association data are controversial or should be regarded merely as putative until independent confirmation

[b] Nomenclature in accordance with the original publication

[c] See Fig. 1 for term definition

polyadenylation signal, UCPAS) located 10–30 nucleotides upstream of the polyadenylation addition site or cleavage site (CS), (2) a less conserved U/GU-rich element (termed the downstream core polyadenylation signal, DCPAS) located 10–30 nucleotides downstream of the CS, and (3) the CS itself (Fig. 1; Zhao et al. 1999).

Except for the preferential use of a CA dinucleotide immediately 5′ to the CS (Sheets et al. 1990), sequence between the UCPAS and the DCPAS is poorly conserved and has not yet been shown to contain additional experimentally confirmed regulatory motifs (Chen et al. 1995; Zarudnaya et al. 2003; Tian et al. 2005). It would nevertheless appear that the distance between the UCPAS and the DCPAS is critical for correct cleavage and polyadenylation to occur (Chen et al. 1995). We have therefore termed this region the 'spacer' and it may be further divided into two portions relative to the CS, viz. the left arm of the spacer (LAS) and the right arm of the spacer (RAS) (Fig. 1).

The 3′ RR comprises the 3′ UTR and 3′ FR (Fig. 1). Whilst the 3′ UTR of a gene can be unequivocally defined as the region stretching from the first nucleotide 3′ to the translational termination codon to the CS, there is no consensus definition of the 3′ FR. Here, we arbitrarily define the 3′ FR as a 2 kb sequence tract downstream of the CS; this allows the inclusion not only of the region known to contain downstream core and auxiliary *cis*-elements which regulate mRNA 3′ end formation, but also those variants that could disrupt more distant transcription enhancer or repressor sites.

Practical considerations for assigning 3′ RR variants to specific domains or motifs

Alternative polyadenylation has been observed in many genes and this phenomenon is generally considered to be a further means of gene regulation (Edwalds-Gilbert et al. 1997; Tian et al. 2005; Yan and Marr 2005). However, to date, no consensus as to how to document the complex patterns of alternative polyadenylation has been proposed. Here we propose a new classificatory system depending upon whether or not alternative splicing is involved.

First, alternatively spliced transcripts that end in different 3′ terminal exons invariably promote alternative polyadenylation and we have termed these, type I *alternative splicing-coupled alternative polyadenylation* (ASCAP) isoforms (Fig. 2a). Second, alternatively spliced transcripts that end in a common 3′ terminal exon but which use different sets of UCPASs and CSs for 3′ end formation have been termed *type II ASCAP* isoforms (Fig. 2a). By contrast, alternatively spliced transcripts that not only end in a common 3′ terminal exon but also use the same set of UCPAS and CS for 3′ end formation were termed *alternative splicing-coupled identical polyadenylation* (ASCIP) isoforms (Fig. 2a). Third, the length of a 3′ terminal exon in a given protein-coding transcriptional framework may vary due to the alternative use of different sets of tandemly arrayed UCPASs and CSs in the downstream sequence; these alternatively polyadenylated isoforms may simply be termed *non-ASCAP* (Fig. 2b). Finally, it is important to make a clear distinction between the alternative polyadenylation-associated use of alternative CSs and the heterogeneity of CSs. By definition, alternative polyadenylation, irrespective of whether ASCAP or

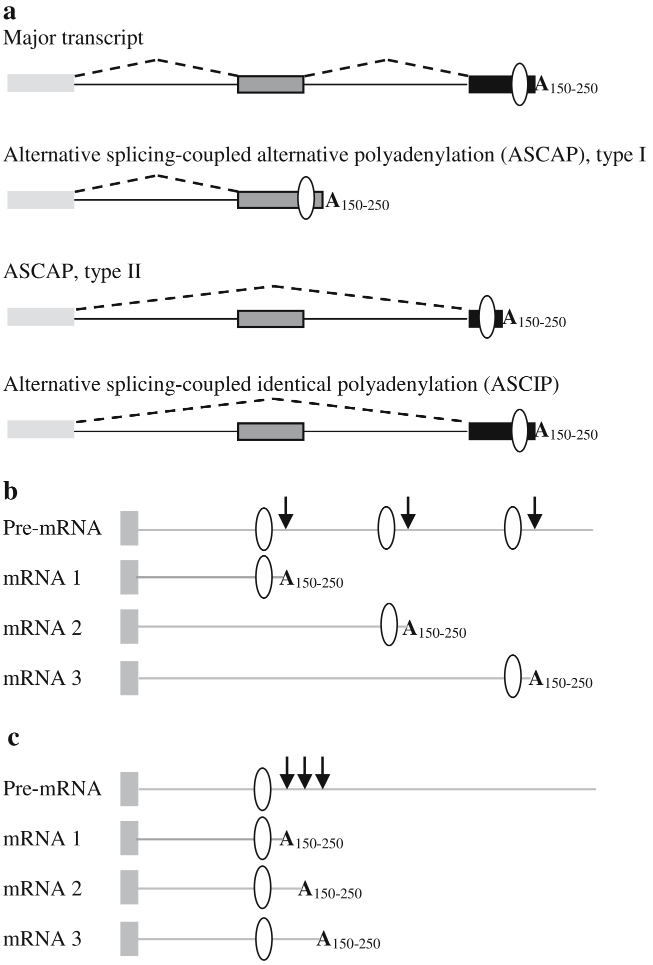

Fig. 2 **a** Schematic representation of the different types of mRNA polyadenylation when alternative splicing is involved. Type I *alternative splicing-coupled alternative polyadenylation* (ASCAP) isoforms end in different 3′ terminal exons; type II ASCAP isoforms end in a common 3′ terminal exon but use different sets of UCPASs and CSs for 3′ end formation; *alternative splicing-coupled identical polyadenylation* (ASCIP) refers to the use of a same set of UCPAS and CS for 3′ end formation within a common 3′ terminal exon. Exons are indicated by *boxes* whereas splicing is denoted by *dotted lines*. *Oval* indicates UCPAS. Note that multiple patterns of alternative splicing may exist for each of the three illustrated types. **b** Schematic representation of *non-alternative splicing-coupled alternative polyadenylation* (non-ASCAP), exemplified by three mRNA species that were generated from the differential use of three tandemly arrayed UCPASs and CSs in the context of a given transcriptional unit. **c** Schematic representation of the heterogeneity of CS in the context of a common UCPAS. Symbols used in **b** and **c**: *shaded box*, coding region of the last exon; *oval*, UCPAS; *arrow*, CS

non-ASCAP, involves the use of different sets of UCPASs and CSs. In addition, some 50% of human mRNAs have more than one CS (Beaudoing and Gautheret 2001; Pauws et al. 2001; Tian et al. 2005) in the context of a common UCPAS motif (Fig. 2c).

Based upon the above classificatory system, all the collated 3′ RR variants were precisely localised within GenBank-available genomic DNA sequences and/or mRNA sequences in the following way: first, when a variant occurred in genes with alternatively spliced isoforms, it was usually assigned to the main transcript (i.e. the most extensively studied transcript with respect to physiological expression and function); second, when a variant occurred in genes with non-ASCAP isoforms, it was assigned to the shortest 3′ UTR whilst its possible relevance to the context of longer UTRs is specified only when deemed to be of potential importance; and finally, the CS of a given mRNA with heterogeneous CSs was assigned 3′ to a CA dinucleotide (in accordance with Sheets et al. 1990) wherever possible; otherwise the location of the 5′-most CS was used to represent the location of the CS (in accordance with Tian et al. 2005).

In addition, for the UCPAS, LAS and 3′ FR variants, mRNAs were systematically aligned with both ESTs and genomic sequences available in GenBank in order to locate the UCPAS hexamers and CSs accurately. This was not however performed for some of the USS variants located at least 100 bp 5′ to the end of the GenBank-annotated reference mRNAs.

Distribution of reported 3′ RR variants in the different domains or motifs

The distribution of reported 3′ RR variants in the different domains or motifs is not proportional to the lengths of the respective domains or motifs (Table 1). This may be largely accounted for by the bias inherent in mutation screening but may also reflect the differential functional importance of the different domains. For example, only ten variants have been found in the 2,000 bp 3′ FR (including LAS, DCPAS and DSS). Apart from being screened much less often, this region clearly has a much more limited regulatory role than the 3′ UTR since it is not included in the mature mRNA. Finally, it should be noted that seven variants (two of which occurred at the 3′ end of the unique 3′ UTR-lacking *GLA* mRNA and five others within the USSs of their respective genes) have been treated as isolated examples owing to their unique features (Table 1).

Systematic evaluation of 3′ RR variants with respect to their potential functional consequences

All the collated variants were evaluated individually in the context of the initial disease association reports, replication studies, the physiological role of the gene involved, the expression level of mRNA/protein in vivo, and in vitro functional analysis data whenever available. In particular, considerable efforts were made to validate/elucidate the mechanisms underlying the bona fide or putative functional variants wherever appropriate. Here it should be emphasized that most of the functional analyses were performed in transient transfection assays, in which a fragment of the studied 3′ UTR was inserted downstream of a reporter gene. There are always concerns as to whether in vitro results can be extrapolated to the in vivo situation (Pastinen and Hudson 2004). Nevertheless, every care has been taken to avoid potentially misleading interpretations by considering the abovementioned factors. It is also important to emphasize that linkage disequilibrium (LD) is not only often seen with common variants but is also frequently encountered in the context of rare mutations. However, in the absence of appropriate functional analytical data, all 3′ RR variants in LD with other variants were regarded as being of potential functional importance and were therefore analysed in the same way as those that had not been shown to be in LD with other variants.

Disease-associated variants within the UCPAS

The UCPAS hexamer serves as the binding site for CPSF (Fig. 1) and constitutes the most frequently encountered motif in the 3′ UTRs of protein-coding genes (Graber et al. 1999; Zhao et al. 1999; Beaudoing et al. 2000; Tian et al. 2005; Xie et al. 2005). It is therefore not surprising that a considerable number (17) of functional variants within the UCPAS have been reported (Table 1). It should however be pointed out that whereas some 92% of human protein-coding genes contain the consensus UCPAS hexamer (AATAAA) motif or a close variant thereof (Tian et al. 2005; see also Table 2 for the 12 most frequently encountered UCPAS hexamer variants), the remaining 8% of genes may employ an alternative tripartite mechanism for poly(A) site recognition (Venkataraman et al. 2005). In this study, however, no tripartite-like structures were identified in any of the genes in which the UCPAS hexamers and cleavage sites (CSs) were accurately located.

HBA2: the gene in which the first disease-causing UCPAS mutation was characterized

Two functional human α-globin genes, α2 (*HBA2*; MIM# 141850) and α1 (*HBA1*; MIM# 141800) are tandemly repeated on chromosome 16pter-p13.3. Mutations in the duplicated *HBA2* and *HBA1* genes can result in diminished (α^+ thalassaemia) or absent (α^0 thalassaemia) α-chain production.

A homozygous A > G mutation in a patient with α^+ thalassaemia that altered the UCPAS of the *HBA2* gene from AATAAA to AATAAG (Fig. 3), was the first UCPAS mutation reported to cause a human inherited

Table 2 Most frequently encountered UCPAS hexamers in human protein-coding genes and disease-associated UCPAS mutations

Genome wide-analysis[a]		This study	
UCPAS	Frequency (%)	Disease-associated mutations[b]	Frequency (times)[c]
AATAAA	53.18		
ATTAAA	16.78		
TATAAA	4.37		
AGTAAA	3.72		
AAGAAA	2.99		
AATATA	2.13	AATATA	1
AATACA	2.03		
CATAAA	1.92		
GATAAA	1.75	GATAAA	1
AATGAA	1.56	AATGAA	3
TTTAAA	1.20		
ACTAAA	0.93		
AATAGA	0.60	AATAGA	1
		AATAAG	3
		AACAAA	1
		AAAAAA	1

[a] Data from Tian et al. (2005)

[b] Only single base-pair substitutions that have occurred within the canonical AATAAA motif have been included (see Fig. 3)

[c] A specific mutation, irrespective of how many times it has been reported in the same gene, was counted only once

disease (Higgs et al. 1983). Since the *HBA1* gene in the same patient was inactivated by a homozygous frameshift mutation, the observed α^+ phenotype must have been due solely to the mutant *HBA2* gene. Indeed, analysis of cytoplasmic mRNA obtained from peripheral blood reticulocytes of the patient revealed that the *HBA2* gene was still expressed, albeit at a significantly reduced level (some 5% of normal; Higgs et al. 1983). In a HeLa cell-based transient expression system, the 3′ end of wild-type *HBA2* mRNA was shown to correspond to the usual poly(A) addition site, whereas the majority of the mutant mRNA extended beyond this point (Higgs et al. 1983; Whitelaw and Proudfoot 1986). However, mutant *HBA2* mRNA transcripts, obtained from the peripheral blood reticulocytes of the patient, were identical in size to those from normal reticulocytes. They would thus appear to represent those transcripts that have not only been cleaved correctly at the usual polyadenylation site but which have also been normally polyadenylated (Higgs et al. 1983). By contrast, the abnormal (i.e. extended) *HBA2* transcripts, observed in transfected HeLa cells but not in the patient's peripheral blood reticulocytes, were relatively unstable as compared with the normally formed transcripts (Higgs et al. 1983). This kind of aberrant mRNA with an extended 3′ UTR is probably degraded by the same cellular machinery that is responsible for nonsense-mediated mRNA decay (Muhlrad and Parker 1999; Amrani et al. 2006).

Three other *HBA2* UCPAS mutations have also been reported (Fig. 3). Although these mutations were not functionally characterized, they may cause disease through a similar mechanism to that discussed above.

HBB: the gene with the most reported UCPAS mutations

β-thalassaemia is an autosomal recessive disease characterized by defective synthesis of the β-chains (encoded by the *HBB* gene; MIM# 141900) of the haemoglobin tetramer. As in the case of α-thalassaemia, β-thalassaemia can also be divided into two types: β^+, in which some chains are present, and β^0, in which β chain synthesis is absent.

Orkin et al. (1985) reported an AATAAA → AACAAA mutation in the UCPAS of the *HBB* gene in a patient with β^+-thalassaemia (Fig. 3). This mutation was shown to result in an extended transcript (~1,500 bp, cf. the normal transcript of ~600 bp) not only in vitro but also in vivo. Two additional β^+-thalassaemia-causing UCPAS mutations in the *HBB* gene, AATAAA → AATGAA and AATAAA → AATAGA (Jankovic et al. 1990; Fig. 3), were also found to be associated with extended unstable ~1,500 bp transcripts in vivo.

Rund et al. (1992) reported a further two *HBB* UCPAS mutations, AATAAA → AATAAG and AATAAA → A in patients with β^+-thalassaemia (Fig. 3). A detailed analysis of in vivo-expressed mRNA in these patients enabled Rund et al. (1992) to draw several important conclusions:

- A significant amount of correct cleavage and polyadenylation occurred in spite of the AATAAA → AATAAG mutation. This indicates that the highly conserved AATAAA hexamer may not be an absolute requirement for correct mRNA 3′ end formation.
- Only four extended *HBB* transcripts with lengths of ~1,500, ~1,650, ~2,450 and ~2,900 bp, respectively, were present in RNA derived from patients carrying the AATAAA → AATAAG mutation, although up to 14 cryptic AATAAA sequences are present distal to the normal UCPAS within the range of the extended transcripts. This serves to demonstrate that not all UCPAS hexamers within 3′ UTRs are functional and that additional contextual signals may be required for correct mRNA 3′ end formation to occur.
- The human *HBB* gene primary transcript can extend to > 5 kb beyond the CS. This illustrates the point that transcriptional termination is not a spatially precise event but rather can extend over hundreds or even thousands of nucleotides (Birnstiel et al. 1985; Kim and Martinson 2003).

Although four additional *HBB* UCPAS mutations have been reported in β^+-thalassaemia patients (Fig. 3), no associated information about in vivo mRNA expression and in vitro characterization has been made available.

The *HBB* gene itself harbours a total of nine UCPAS variants whereas the other five genes listed exhibit only eight between them (Fig. 3). Of the many factors that could have contributed to the preponderance of disease-causing UCPAS mutations in the *HBB* gene, perhaps the most important is likely to be that the *HBB* gene coding region is fairly short and uncomplex, thereby increasing the relative mutational target size proffered by the UCPAS. However, in comparison with the *HBA2* gene, which is similar in size and structure to *HBB* but harbours only four UCPAS mutations, it may be significant that the *HBB* gene does not have a functionally equivalent paralogue.

IL2RG: difficulties in assessing the significance of UCPAS mutation-associated transcripts of unusual length

An A > G mutation that changed the UCPAS of the *IL2RG* gene from AATAAA → AATAAG was identified in a male patient with sporadic X-linked severe combined immunodeficiency (Hsu et al. 2000; Tsai et al. 2002; Fig. 3). Northern blotting revealed a low amount of a 4 kb mRNA species (cf. the normal 1.8 kb *IL2RG* mRNA) in a B-cell line derived from the patient and this abnormally long mRNA was accounted for by invoking extended transcription (Hsu et al. 2000). However, in the process of cloning the human *IL2RG* gene, Takeshita et al. (1992) had already demonstrated that in addition to the main ~1.8 kb transcript, a second ~3.6 kb mRNA species was also detectable in normal human peripheral blood leukocytes; whereas the shorter species corresponds to the cloned *IL2RG* gene (GenBank accession number D11086.1), the origin of the longer one remains unknown.

Given that the lengths of the two longer transcripts represent only rough estimates from Northern blotting, we consider that it is likely that the two longer transcripts described by Takeshita et al. (1992) and Hsu et al. (2000) represent one and the same transcript. Since the human *IL2RG* gene spans about 4 kb on the X chromosome (GenBank accession number NC_000023.8), we surmise that the longer mRNA species may represent a non-processed nuclear transcript of the *IL2RG* gene. To provide some support for this postulate, we searched the EST-human database in GenBank (as of July 8, 2005) using BLASTN against ±20 bp flanking the cleavage site of the human *IL2RG* gene and indeed found a clone (GenBank accession number CT002657.1) that is characteristic of a non-processed nuclear transcript viz. it comprises the 3′ end sequence of the last intron, the last exon (coding region plus 3′ UTR), and sequence 3′ to the cleavage site of the *IL2RG* gene.

The AATAAA → AATAAG mutation could indeed have resulted in the production of extended transcripts. Such transcripts were however probably unstable and therefore unlikely to be detectable in vivo. Finally, this mutation certainly did not render the gene completely

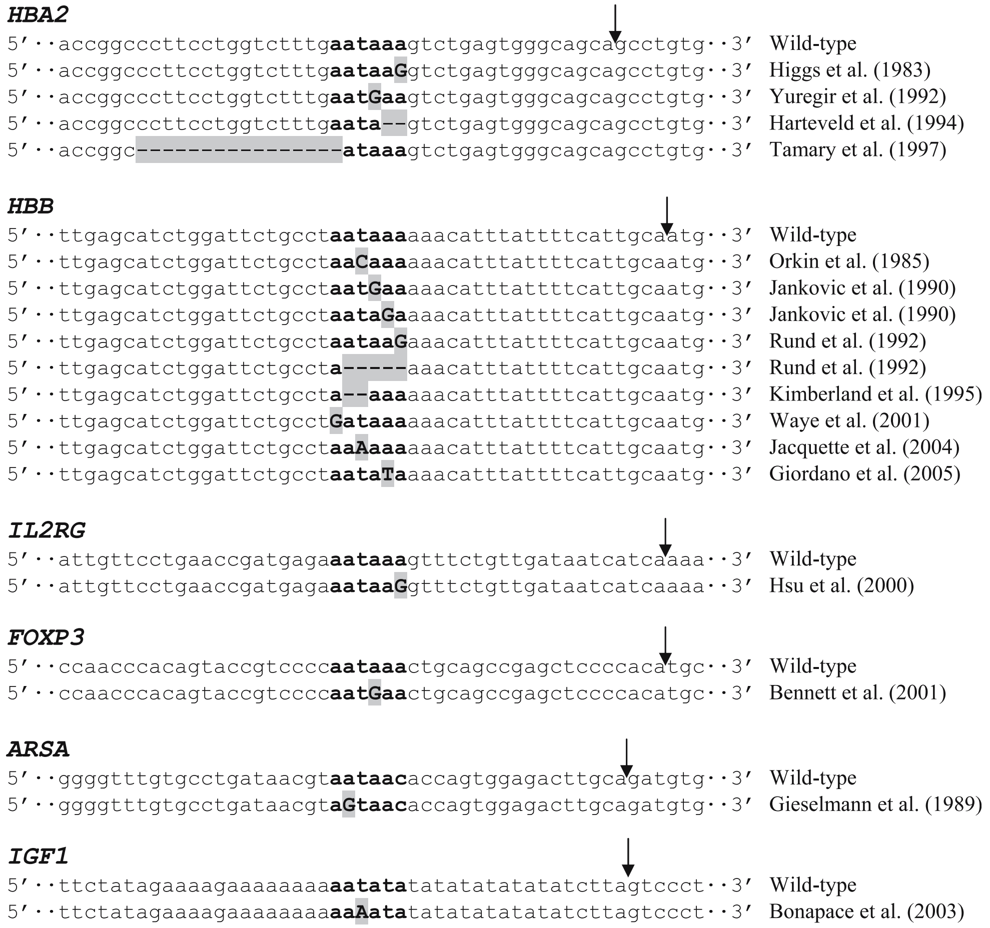

Fig. 3 Naturally occurring human UCPAS variants. The UCPAS motif is highlighted in *bold*. Point mutations are denoted by *upper case letters* and *shaded*. Deleted nucleotides are *barred* and *shaded*. The CSs are indicated by *downward pointing arrows*. Reference is made to the publications that first reported these mutations. Note also that both the *ARSA* and *IGF1* genes utilised non-canonical UCPAS hexamers and have alternatively polyadenylated transcripts

```
HBA2
5′··accggcccttcctggtctttgaataaagtctgagtgggcagcagcctgtg··3′  Wild-type
5′··accggcccttcctggtctttgaataaGgtctgagtgggcagcagcctgtg··3′  Higgs et al. (1983)
5′··accggcccttcctggtctttgaatGaagtctgagtgggcagcagcctgtg··3′  Yuregir et al. (1992)
5′··accggcccttcctggtctttgaata--gtctgagtgggcagcagcctgtg··3′  Harteveld et al. (1994)
5′··accggc----------------ataaagtctgagtgggcagcagcctgtg··3′  Tamary et al. (1997)

HBB
5′··ttgagcatctggattctgcctaataaaaaacatttattttcattgcaatg··3′  Wild-type
5′··ttgagcatctggattctgcctaaCaaaaaacatttattttcattgcaatg··3′  Orkin et al. (1985)
5′··ttgagcatctggattctgcctaatGaaaaacatttattttcattgcaatg··3′  Jankovic et al. (1990)
5′··ttgagcatctggattctgcctaataGaaaacatttattttcattgcaatg··3′  Jankovic et al. (1990)
5′··ttgagcatctggattctgcctaataaGaaacatttattttcattgcaatg··3′  Rund et al. (1992)
5′··ttgagcatctggattctgccta-----aaacatttattttcattgcaatg··3′  Rund et al. (1992)
5′··ttgagcatctggattctgccta--aaaaaacatttattttcattgcaatg··3′  Kimberland et al. (1995)
5′··ttgagcatctggattctgcctGataaaaaacatttattttcattgcaatg··3′  Waye et al. (2001)
5′··ttgagcatctggattctgcctaaAaaaaaacatttattttcattgcaatg··3′  Jacquette et al. (2004)
5′··ttgagcatctggattctgcctaataTaaaacatttattttcattgcaatg··3′  Giordano et al. (2005)

IL2RG
5′··attgttcctgaaccgatgagaaataaagtttctgttgataatcatcaaaa··3′  Wild-type
5′··attgttcctgaaccgatgagaaataaGgtttctgttgataatcatcaaaa··3′  Hsu et al. (2000)

FOXP3
5′··ccaacccacagtaccgtccccaataaactgcagccgagctccccacatgc··3′  Wild-type
5′··ccaacccacagtaccgtccccaatGaactgcagccgagctccccacatgc··3′  Bennett et al. (2001)

ARSA
5′··ggggtttgtgcctgataacgtaataacaccagtggagacttgcagatgtg··3′  Wild-type
5′··ggggtttgtgcctgataacgtaGtaacaccagtggagacttgcagatgtg··3′  Gieselmann et al. (1989)

IGF1
5′··ttctatagaaaagaaaaaaaaaatatatatatatatatatcttagtccct··3′  Wild-type
5′··ttctatagaaaagaaaaaaaaaAatatatatatatatatatcttagtccct··3′  Bonapace et al. (2003)
```

non-functional because the patient expressed trace amounts of IL2RG (Tsai et al. 2002).

FOXP3: an illustration of the necessity to analyse entire 3′ UTRs

IPEX (Immune dysfunction, polyendocrinopathy, enteropathy, X-linked) is caused by mutations in the *FOXP3* gene (MIM# 300292). In a five-generation family with IPEX, no mutations were found in the *FOXP3* gene after sequencing all 11 exons including at least 50 bp flanking each intron/exon boundary, 531 bp promoter sequence and 487 bp 3′ UTR. Since the 3′ UTR sequence originally screened did not contain a canonical UCPAS, Bennett et al. (2001) extended their analysis to further downstream 3′ UTR sequence and consequently found an A > G mutation within the first UCPAS encountered, some 972 bp 3′ to the translational termination codon (AATAAA > AATGAA; Fig. 3). That this substitution was disease-causing was supported by the following observations: (1) it segregated with the disease; (2) it was absent in 318 control chromosomes; (3) the next potential UCPAS motif was not encountered for a further 5.1 kb and (4) *FOXP3* mRNA expression was not only clearly reduced in the primary peripheral blood mononuclear cells from the heterozygous mother but also undetectable in those from the affected son (Bennett et al. 2001).

The mutant UCPAS sequence observed in this case, AATGAA, is used as a normal polyadenylation signal in 1.6% of human genes (Table 2). This lends further support to the notion that UCPAS function must be context-dependent.

ARSA: the mutational loss of one UCPAS did not lead to the increased use of alternative UCPASs

Metachromatic leukodystrophy (MLD) is an autosomal recessive inborn error of metabolism resulting from the deficiency of the lysosomal enzyme, arylsulphatase A (ARSA). ARSA deficiency is consequent to mutations in the *ARSA* gene (MIM# 607574) but not all mutations resulting in ARSA deficiency cause MLD. In this regard, ~10% of healthy individuals exhibit ARSA activity levels comparable to those seen in MLD patients; this non-pathogenic reduction in ARSA activity is caused by homozygosity for the so-called ARSA pseudodeficiency allele (ARSA-PD; e.g. Gieselmann et al. 1989; Harvey et al. 1998).

Northern blotting using *ARSA* cDNA probes has detected three different RNA species of 2.1, 3.7 and 4.8 kb in poly(A)$^+$ RNA from normal human fibroblasts (Gieselmann et al. 1989; Kreysing et al. 1990). These probably represent either type II ASCAP isoforms (refer to Fig. 2a) or non-ASCAP mRNA isoforms (refer to Fig. 2b), both because a probe derived from DNA sequence 300 nucleotides downstream of the first UCPAS (i.e. that used for the 2.1 kb species) hybridises only to the two larger transcripts and because an intron-specific probe does not hybridise to any of the three transcripts (Kreysing et al. 1990). The 2.1 kb mRNA species, whose sequence is known (Kreysing et al. 1990), accounts for 71–90% of the total *ARSA* mRNA (Gieselmann et al. 1989; Harvey et al. 1998). The full-length sequences of the two larger transcripts have still not been determined.

An A > G mutation in an *ARSA*-PD allele that changed the UCPAS of the major mRNA species from AATAAC to AGTAAC has been described (Gieselmann et al. 1989; Fig. 3). Although the expression of the 2.1 kb mRNA species was seriously reduced in *ARSA*-PD homozygous individuals, that of the other two longer transcripts was unaltered (Gieselmann et al. 1989; Harvey et al. 1998). This example therefore indicates that the loss of one UCPAS may not necessarily lead to a compensatory increase in the use of other alternative UCPASs. In addition, it is worth noting that AATAAC is not among the most frequently encountered UCPAS hexamers in human genes (Table 2).

IGF1: an UCPAS mutation with probable dual effect

The insulin-like growth factor 1 gene (*IGF1*; MIM# 147440) encodes three major mRNA species of 1.1, 1.3 and 7.6 kb, respectively. Since the 1.1 and 7.6 kb transcripts share the same coding sequence (exons 1, 3, 4 and 6; gene structure in accordance with Sussenbach et al. 1992) but differ in terms of their 3′ UTRs, they therefore represent non-ASCAP isoforms (refer to Fig. 2b). The 1.3 kb species is a type I ASCAP isoform (refer to Fig. 2a) which comprises exons 1, 3, 4 and 5, and will not be considered further.

In a patient with short stature, sensorineural deafness and delayed psychomotor development, Bonapace et al. (2003) identified a homozygous AATATA → AAAATA UCPAS mutation in the *IGF1* gene with respect to the 1.1 kb transcript (Fig. 3). This transversion is probably disease-causing because (1) it was the only mutation found within the entire coding region, promoter and 3′ UTR of the *IGF1* gene, (2) it was not present in 100 unrelated healthy controls and (3) the patient manifested a significant deficiency of serum IGF1. The original authors attempted to assess the mutation's effect on mRNA transcription and maturation by RT-PCR, using total RNA prepared from leukocytes from the homozygous patient, heterozygous mother and father and a normal control. The normal control displayed a single band of 450 bp, the patient a single band of 340 bp, and the mother and father both bands (see Fig. 3 in Bonapace et al. 2003). Although the sequences for the two RT-PCR primers were not provided in the original publication, it may be inferred that the two primers must have targeted the 5′ end of exon 6 and sequences downstream of the pre-mRNA CS of the *IGF1* gene with respect to the 1.1 kb mRNA [this means that the normally expressed 1.1 kb mRNA could not have been amplified by this pair of primers]. Although the 450 bp band corresponds to a cDNA sequence amplified from the normally expressed 7.6 kb mRNA species, our evaluation of the sequence of the 340 bp band (see Fig. 4 in Bonapace et al. 2003) revealed, however, that it was probably derived from non-specific PCR amplification. It would thus appear that no specific amplification was obtained from the homozygous patient. This, when combined with the significant IGF1 deficiency in the patient, strongly suggests that the expression of both the 1.1 and 7.6 kb mRNA species were adversely affected. In other words, the observed point mutation could have abrogated not only 3′ end formation and stability of the shorter mRNA species (as a UCPAS mutation), but also that of the longer transcript (as a USS mutation).

Finally, it is important to note that, in the context of the 1.1 kb mRNA species, the AATATA > AAAATA mutation yields an alternative AATATA motif 2 bp downstream from the physiological UCPAS hexamer (Fig. 3). Clearly, this newly created hexamer motif is unable to compensate for the loss of the physiological one, most probably due to a sub-optimal distance between the new UCPAS and the normal CS (Chen et al. 1995).

*NAT1*10* is a common USS polymorphism rather than an UCPAS mutation as originally claimed

The gene encoding arylamine *N*–acetyl transferase 1, known as *NAT1* (MIM# 108345), is highly polymorphic. Approximately 30% of Europeans carry a *NAT1*10* allele in which the canonical UCPAS AATAAA has become altered to AAAAAA. A number of studies have found associations of this allele with colorectal and bladder cancer, but results have often been contradictory (e.g. Gu et al. 2005). Irrespective of the status of these disease associations, the T > A transversion associated with the *NAT1*10* allele would be expected to affect 3′ end formation of its associated mRNA. However, as noted by de Leon et al. (2000), an aat triplet lies immediately 5′ to the original UCPAS (AATAAA) and, together with the first three A nucleotides of the mutant sequence, would be expected to generate a new canonical UCPAS (aatAAA) only 3 nucleotides 5′ to the original UCPAS (i.e. aatAATAAA → aatAAAAAA). The in vivo efficacy of the newly generated UCPAS was thought to be unclear because the 5′ shift of three nucleotides may give rise to a sub-optimal distance between the UCPAS and the normal CS (de Leon et al. 2000).

de Leon et al. (2000) addressed this issue by means of two complementary approaches. First, they demonstrated that the recruitment of the immediate upstream aat to the formation of a new AATAAA polyadenylation signal site does not disrupt the predicted secondary structure of the *NAT1*10* pre-mRNA which is indistinguishable from that of the wild-type sequence. Then they performed transfection studies in COS-1 cells: the amount of NAT1*10 protein in COS-1 cell cytosol was identical to that of the wild-type protein. The authors therefore concluded that the aatAATAAA → aatAAAAAA change in *NAT1*10* had been fully compensated for by the newly generated polyadenylation signal site.

We were intrigued by the extremely mild effect on gene expression conferred by this mutation that occurred within a canonical AATAAA element. We therefore traced the original report that cloned the *NAT1* gene and found that the gene was sequenced at the genomic rather than the cDNA level (Blum et al. 1990). Thus, the 3′ UTR of the gene was formerly unknown and the ‘canonical UCPAS AATAAA’, referred to subsequent work, was regarded as only ‘putative’ in the original study (Blum et al. 1990). Evaluation of the mRNA sequences and ESTs available in UniGene Hs. 155956 (for *Homo sapiens NAT1*; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = unigene&cmd = search&term = hs.155956; as of Aug. 18, 2005) allowed us to conclude that, on the basis of all informative sequences with an unambiguous poly(A) tail, the physiological UCPAS of the *NAT1* gene should be assigned as the ATTAAA motif located 154 bp 3′ to the so-called ‘canonical AAT AAA motif’ (see Supplementary Table S1). Consequently, *NAT1*10* should be re-defined as an USS variant, the above functional interpretation of de Leon et al. (2000) being simply spurious.

Correlation of genome-wide and biochemical analysis with naturally occurring mutation data

As indicated in Table 2, the canonical AATAAA sequence accounts for ~53% of the presumably functional UCPAS hexamers detected in the 3′ UTRs of human protein-coding genes. The most frequent UCPAS variant, ATTAAA, has a frequency of ~16% whereas the 11 next most frequently encountered UCPAS hexamers exhibit frequencies of 0.6–4.4%. Early in vitro studies conclusively demonstrated that, with the exception of ATTAAA, all variants of the canonical AATAAA sequence strongly inhibit pre-mRNA processing (Wickens and Stephenson 1984; Wilusz et al. 1989; Sheets et al. 1990). This is also the case for the functionally analysed pathogenic UCPAS mutations. It therefore follows that the less frequently used non-canonical UCPAS hexamers may well have become evolutionarily adapted (and hence perhaps even optimised) to the expression of their respective genes. Alternatively, it may be that the use of an apparently sub-optimal UCPAS site represents a gene-specific strategy for modulating the level of gene expression. Nevertheless, it seems very likely that the less frequently used non-canonical UCPAS hexamers serve as comparatively poor binding sites for CPSF; thus, CPSF would be more likely, in those genes using non-canonical hexamers as physiological UCPASs, to be able to bind to other potential UCPAS elements. It is precisely this type of gene that is often associated with alternative polyadenylation (e.g. Xiong et al. 1991; Zhao et al. 1999; Hall-Pogar et al. 2005). Interestingly, of the six genes listed in Fig. 3, the first four use AATAAA as an

UCPAS and have no alternatively polyadenylated mRNA species, whereas the remaining two genes utilize non-canonical UCPAS variant sites and give rise to at least two alternatively polyadenylated mRNA isoforms.

From Table 2, we can see that the pathogenic point mutations that have been reported within canonical AATAAA UCPAS hexamers tend to generate motifs that correspond to the less frequently used non-canonical UCPAS hexamers. Further, it has been clearly shown that, in a given sequence context, the effect of single nucleotide substitution within the UCPAS AATAAA hexamer on cleavage and polyadenylation depends critically on its position within the hexamer; and the in vitro phenotypic severity follows the order of positions 3, 4, 6 > 5 > 1 > 2 (Sheets et al. 1990). Intriguingly, of the 12 non-canonical UCPAS hexamers found in human genes, the 2nd position in the motif is the most frequently affected whereas the 6th and 4th positions are the least frequently affected (Table 2). By contrast, of the pathogenic canonical (AATAAA) hexamer point mutations, the 2nd position is least frequently affected whereas the 6th and 4th positions are the most frequently affected (Table 2). Finally, biochemical analysis has revealed that a substituting G in any given position is often associated with a severe phenotype by comparison with other base substitutions (Sheets et al. 1990). Not surprisingly, therefore, G is the most frequently encountered substituting base among the pathogenic point mutations known to have occurred within canonical AATAAA UCPAS hexamers (Table 2).

Variants that have occurred within the LAS

Currently, the only known role of the short LAS is involvement in mRNA 3′ end formation at the pre-mRNA level (discussed below). A recent bioinformatics analysis has identified several well conserved motifs in the region -40/−1 (the first nucleotide 5′ to the CS as –1; nomenclature used consistently in this section) of human genes (Hu et al. 2005). However, since (1) the analysed region is 10–30 bp longer than the LAS as defined here and (2) the relative position of each identified motif within the analysed region is unknown, the utility of this information is seriously limited. Nevertheless, none of the six known LAS variants were located in these motifs and their possible functional consequences will therefore only be discussed in the context of our current understanding of the role of the LAS in mRNA 3′ end formation.

Three point mutations involving the dinucleotide immediately upstream of the CS

Preferential use of a CA dinucleotide immediately upstream of the CS

To assess the evolutionary conservation of the nucleotides to which the poly(A) is added, Sheets et al. (1990) aligned 63 vertebrate cDNA sequences with their corresponding genomic sequences. Whenever the start of the poly(A) tail coincided with an A in the genomic sequence, this A was assigned as the nucleotide to which the poly(A) stretch was added post-transcriptionally [in accordance with the observation that the first nucleotide of the poly(A) tail in most adenovirus L3 and SV40 late mRNAs is likely to be template-encoded (Moore et al. 1986; Sheets et al. 1987)]. Having found a C residue at position –2 in 59% of the genes examined and an A at position –1 in 71%, Sheets et al. (1990) concluded that CA is the most common dinucleotide to which poly(A) is added. Using a large set of human genes (NB. the exact number of genes used for the relevant analysis cannot be determined from the original publication), Zhang (1998) found that a C residue at position –2 and an A residue at position –1 accounted for 46 and 64% of nucleotides at these sites, respectively. Since the C at position –2 and the A at position –1 were counted separately in both the above cited studies, we have no idea as to the proportion of human genes that use a CA dinucleotide immediately before the CS. Using the same principle employed by Sheets et al. (1990), we systematically annotated the CSs of all the genes known to contain UCPAS, LAS and 3′ FR variants and of some of the genes containing USS variants; a CA dinucleotide was found to be used in 25 (37%) of the 67 genes analysed.

Since a significant fraction of CSs are not preceded by a CA dinucleotide and since >50% of human protein-coding genes harbour multiple mRNA CSs (Tian et al. 2005), it would appear that a CA dinucleotide cannot be an absolute requirement for correct cleavage, a situation analogous to the use of both the canonical AATAAA sequence and its variants by human genes. In this regard, a detailed biochemical analysis of a series of constructs, in which the wild-type CA dinucleotide at positions –2 and –1 of the SV40 late polyadenylation signal was changed to all 16 possible combinations, clearly demonstrated that (1) mutations at position –2 or/and –1 had only minor effects on the overall efficiency of cleavage; (2) whilst cleavage occurred predominantly at certain sites, additional minor sites of cleavage over a ~6 base-pair interval were also observed and (3) CS usage at position –1 was found to be in the order of preference A > U > C > > G (Chen et al. 1995).

F2: two variants that probably increase the efficiency of CS recognition

G20210A, occurring immediately 5′ to the CS of the prothrombin gene (*F2*; MIM# 176930), displays an allele frequency of 1–4% in the Caucasian population and represents a moderate risk factor for venous thrombosis (Poort et al. 1996). Consistent with the observations that (1) the G20210A polymorphism changes the CG dinucleotide to a preferentially used CA dinucleotide (Fig. 4), (2) it is associated with a gain-of-function, and

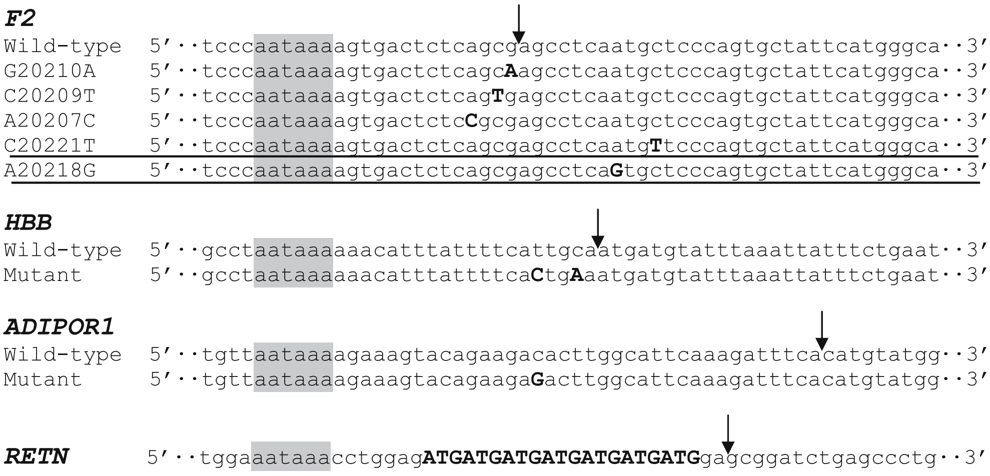
F2
Wild-type 5′··tcccaataaaagtgactctcagcgagcctcaatgctcccagtgctattcatgggca··3′
G20210A 5′··tcccaataaaagtgactctcagcAagcctcaatgctcccagtgctattcatgggca··3′
C20209T 5′··tcccaataaaagtgactctcagTgagcctcaatgctcccagtgctattcatgggca··3′
A20207C 5′··tcccaataaaagtgactctcCgcgagcctcaatgctcccagtgctattcatgggca··3′
C20221T 5′··tcccaataaaagtgactctcagcgagcctcaatgTtcccagtgctattcatgggca··3′
A20218G 5′··tcccaataaaagtgactctcagcgagcctcaGtgctcccagtgctattcatgggca··3′
HBB
Wild-type 5′··gcctaataaaaaacatttattttcattgcaatgatgtatttaaattatttctgaat··3′
Mutant 5′··gcctaataaaaaacatttattttcaCtgAaatgatgtatttaaattatttctgaat··3′
ADIPOR1
Wild-type 5′··tgttaataaaagaaagtacagaagacacttggcattcaaagatttcacatgtatgg··3′
Mutant 5′··tgttaataaaagaaagtacagaagaGacttggcattcaaagatttcacatgtatgg··3′
RETN 5′··tggaaataaacctggagATGATGATGATGATGATGATGgagcggatctgagccctg··3′

Fig. 4 Naturally occurring LAS variants. The UCPAS motif is *shaded*. The cleavage sites are indicated by *downward pointing arrows*. Point mutations in *F2, HBB*, and *ADIPOR1* and the ATG repeat polymorphism exemplified by the 7-repeat allele in *RETN* are highlighted in *bold upper case letters*. The two 3′ flanking region variants in the *F2* gene are also included (*underlined*)

(3) heterozygous carriers of the A allele exhibit on average a 25% higher plasma prothrombin level as compared to wild-type individuals, this polymorphism might reasonably be expected to give rise to improved cleavage site by enhancing CS recognition thereby leading to increased mRNA accumulation and protein synthesis. This indeed turned out to be the case, as has been demonstrated by several in vitro studies (Gehring et al. 2001; Pollak et al. 2002; Ceelie et al. 2004; Danckwardt et al. 2004; Sachchithananthan et al. 2005). Thus, in the words of Danckwardt et al. (2004), "the physiological G at the cleavage site at position 20210 is the functionally least efficient nucleotide to support 3′ end processing but has evolved to be physiologically optimal".

C20209T, which affects the –2 position relative to the CS of *F2* (Fig. 4) and which has been noted at low frequency in the African-American population (~0.4%), probably acts as a modifier of thrombotic risk (Warshawsky et al. 2002; Schrijver et al. 2003; Itakura et al. 2005; Soo et al. 2005; Wylenzek et al. 2005). This variant, if it is indeed of functional significance, should also increase the efficiency of pre-mRNA cleavage and polyadenylation.

HBB: a mutation that may decrease the efficiency of CS recognition

Two heterozygous single nucleotide substitutions in *cis* that altered the end of the *HBB* 3′ UTR from ttgCA to ctgAA (Fig. 4) [combined with a null mutation on the other allele] were identified in a patient with *β*-thalassemia intermedia (Heath et al. 2001). Note that this mutation changed the preferred CA dinucleotide to a sub-optimal AA (Sheets et al. 1990). Given both the loss-of-function nature of known *HBB* mutations and the patient's mild phenotype (a moderate microcytic, hypochromic anaemia with normal iron and elevated HbF; Heath et al. 2001), this change could lead to a moderate decrease in the efficiency of pre-mRNA CS recognition.

Two point mutations occurring in residues upstream of the –2/–1 dinucleotide

To date, we are not aware of any reports of the functional characterization of artificially introduced single nucleotide substitutions with respect to the remaining positions in the LAS (i.e. from +1 downstream of the UCPAS hexamer to position –3 relative to CS). However, substitution of a short sequence tract ranging from 3 to 5 bp in the middle of the LAS of the human complement C2 (*C2*) gene has been found to affect adversely cleavage and polyadenylation efficiency. Moreover, this effect was proportional to the number of bases substituted (Moreira et al. 1995). Based upon the finding that the shortest 3-base substitution resulted in only a two-fold reduction in efficiency as compared with the wild-type sequence (Moreira et al. 1995), and the results of Chen et al. (1995), it is not unreasonable to conclude that nucleotide substitutions in the LAS, from +1 downstream of the UCPAS hexamer to position –3 relative to the CS, are unlikely to have a major effect on mRNA cleavage and polyadenylation. Indeed, the A20207C variant, located at position –4 of *F2* (Fig. 4) and so far only reported in abstract form (Meadows et al. 2002), does not appear to result in major changes to the position of the poly(A) attachment site, the efficiency of polyadenylation, or protein synthesis (Ceelie et al. 2005). The functional consequences of a common C > G SNP, located deep within the LAS of the *APIPOR1* gene (MIM# 605441; Fig. 4) and reported to be associated with reduced gene expression in cell lines from diabetic African-Americans as compared with control cell lines (Wang et al. 2004), appear to be more elusive.

RETN: an ATG repeat variation polymorphism that may exert its effects by altering the optimal distance between the UCPAS and the CS

A specific spacing between the UCPAS and the CS has been found to be essential for efficient cleavage and polyadenylation (Chen et al. 1995). However, until now, no sequence variants have been reported that exert their effects by altering the spacing between the UCPAS and the CS. Our systematic analysis may have identified just such an example.

Pizzuti et al. (2002) identified three alleles involving an ATG triple repeat in the 3′ UTR of the resistin gene (*RETN*; MIM# 605565) in a control population: 8, 7 and 6 repeats with allele frequencies of 0.3, 94.5 and 5.2%, respectively. The 6-repeat allele was found to be associated with a decreased risk of insulin resistance by comparison with the 7-repeat allele. Since the ATG repeat maps to the LAS of the *RETN* gene (Fig. 4), the shortest allele could well lead to decreased *RETN* expression by reducing the optimal distance between the UCPAS and the CS.

Variants located within the USS

There is no doubt that the USS, which accounts for some 97% of the average length of 3′ UTRs in human protein-coding genes (Fig. 1), is heavily involved in gene regulation at the post-transcriptional level. To better understand the relatively large number (81) of USS variants (most of which are polymorphisms), we first performed a comprehensive survey of well-defined *cis*-regulatory elements within the 3′ UTRs of protein-coding genes; this has served to strengthen the notion that RNA regulatory elements function in the context of a specific secondary structure. We then attempted to validate/decipher the potential functional consequences of the collated USS variants by systematically evaluating the primary sequences (against the well-defined *cis*-regulatory motifs) within the context of predicted RNA secondary structures; this enabled us to identify (1) consistent patterns of secondary structural change that may allow the discrimination of non-functional USS variants from their functional counterparts and (2) potential novel regulatory motifs within the 3′ UTRs. The details of this part will be published elsewhere (J.M. Chen, C. Férec, D.N. Cooper, submitted).

3′ FR variants

The 3′ flanking sequence, arbitrarily defined here as the region containing up to 2,000 nucleotides 3′ to the CS, was further divided into three sub-domains or motifs viz. RAS, DCPAS and DSS (Fig. 1).

Unlike the UCPAS, the DCPAS is poorly conserved evolutionarily. This notwithstanding, two main DCPAS types, a U-rich element and a GU-rich element, have been described (Zhao et al. 1999; Zarudnaya et al. 2003; Hu et al. 2005). In addition, some putative *cis*-acting elements in the region from +1 to +100 (the first nucleotide 3′ to CS is numbered +1; nomenclature used consistently in this section) of human genes have also been identified by comparative sequence analysis (Hu et al. 2005). Further, some secondary and higher-order structures formed by sequences mapping within the region +1/+100 have been reported to be involved in mRNA 3′ end formation (Zarudnaya et al. 2003).

It is highly likely that elements which regulate mRNA 3′ end formation are also going to occur in the region downstream of +100 [NB. the human *HBB* gene primary transcript can extend to >5 kb beyond the CS (Rund et al. 1992)] but such elements have been reported only rarely (Dye and Proudfoot 2001; Plant et al. 2005). Rather, most studies have focused upon putative regulatory elements that function as transcriptional enhancers or repressors (see below).

Given the poorly defined nature of the DCPAS and hence the RAS, three of the 10 3′ FR variants (all occurring ≤ 12 bp 3′ to the CS) were assigned as RAS or DCPAS mutations whereas the remaining seven (all occurring ≥74 bp 3′ to the CS) were assigned as DSS mutations.

RAS or DCPAS mutations that could affect mRNA 3′ end formation

F2: C20221T is probably a pathological mutation in a CstF binding site but the functional consequences of A20218G are less clear

A C20221T mutation in the *F2* gene was identified in a 9-year-old child with acute vascular rejection and intrarenal segmental arterial thrombosis of an allogeneic kidney transplant (Wylenzek et al. 2001), a 28-year-old man with Budd-Chiari syndrome (Balim et al. 2003), and a 40-year-old women with pregnancy complications (Schrijver et al. 2003), respectively. The clinical phenotypes of these patients are consistent with a gain-of-function mutation. Interestingly, the C20221T mutation, which occurred at position 11 3′ to the CS of *F2* (Fig. 4), generates a four nucleotide sequence tract TGTT that could potentially enhance CstF binding. Indeed, when tested in vitro, the 20221T allele increases the mRNA expression level by 2.6-fold as compared with the wild-type allele (Danckwardt et al. 2004).

The *F2* A20218G variant (Fig. 4) was cited by Ceelie et al. (2005) as having been first reported in abstract form (Meadows et al. 2002). When tested by in vitro functional analysis, this variant did not lead to major changes in the position of the poly(A) attachment site, in the effectiveness of polyadenylation or in protein expression (Ceelie et al. 2005).

FGG: reduced expression of the γB isoform may be due to increased expression of the type I ASCAP γA isoform carrying a putative gain-of-function mutation in the CstF binding site

The fibrinogen γ gene (*FGG*; MIM# 134850) gives rise to two type I ASACP isoforms, γA and γB (Fornace et al. 1984). [γB corresponds to γ′ in Wolfenstein-Todel and

Mosesson (1981) and Uitte de Willige et al. (2005)]. γA, the major isoform (comprises some 90% of the fibrinogen γ chain in plasma; Wolfenstein-Todel and Mosesson 1981), consists of exons 1–10. By contrast, γB retains intron 9 and polyadenylation occurs in a site within this intron (see Fig. 1 in Uitte de Willige et al. 2005).

A *FGG* haplotype, *FGG*-H2, has recently been reported to be associated with reduced γB levels and elevated total fibrinogen levels that are in turn associated with an increased risk of venous thrombosis (Uitte de Willige et al. 2005). Uitte de Willige et al. (2005) have proposed that the reduced expression of FGG γB might be attributable to the increased expression of γA: the *FGG*-H2 haplotype contains a 10034C > $\underline{\text{T}}$ polymorphism (located 12 bp downstream of the CS in the context of the γA isoform; Supplementary Table S2) which results in the gain of a CstF consensus 2a sequence viz. YGTG$\underline{\text{T}}$YTTYAYTGNNYGT (Beyer et al. 1997). In other words, the T allele may increase the efficiency of 3′ end formation of the γA mRNA isoform, which in turn leads to a competitive decrease in the formation of the γB mRNA isoform.

DSS variants that may affect binding sites for transcriptional enhancers or repressors

It is now apparent that *cis*-acting transcriptional enhancers or repressors can modulate gene expression over very long distances e.g. tens of kilobases or even up to a megabase away from their target genes (reviewed by West and Fraser 2005). Here, we focus exclusively upon the proximal 3′ flanking regions. As we shall see below, three of the seven disease-associated DSS variants have been shown to affect transcription factor-binding sites. The remaining four variants—a G > A SNP (236 bp downstream of the CS) in *CTLA4* (MIM# 123890), a C > T polymorphism (+ 242 bp) in *CYP1A1* (MIM# 108330), a single T deletion (+ 485 bp) in *FABP3* (MIM# 134651), and a C > T polymorphism (+ 707 bp) in *KCNS3* (MIM# 603888) (Table 1; see also Supplementary Table S3)—will not be addressed further owing to the lack of supporting functional analysis.

SERPINA1: an intriguing illustration of the complexity of gene regulation

α1-antitrypsin is an acute-phase reactant. Its concentration in the plasma can increase three to fourfold during inflammation; this effect is mediated primarily by the cytokine, interleukin-6, that interacts with the tissue-specific transcription factor NF-IL6 (Morgan et al. 1997). A G > A substitution at position 1255 in the 3′ flanking region of the α1-antitrypsin gene (*SERPINA1*; MIM# 107400; Supplementary Table S4) occurs in $\sim$5% of apparently healthy individuals. However, it is present in $\sim$17% of patients with chronic respiratory disease (Morgan et al. 1993). Since the G > A transition does not affect the basal expression of the protein (the plasma concentration of α1-antitrypsin is normal in individuals carrying the polymorphism), this variant probably predisposes individuals to chronic respiratory disease by modulating the increase in plasma α1-antitrypsin concentration during inflammation (Morgan et al. 1993). Indeed, this polymorphic variant, which occurs within a DNA binding site for the ubiquitous transcription factor Oct-1 (i.e. ATTTCGA > ATTTCAA) was found to disrupt the functional cooperativity between Oct-1 and NF-IL6, resulting in IL6-deficient acute-phase response (Morgan et al. 1997).

HBD: a variant that affects a putative context-specific transcriptional repressor binding site

A G > A substitution 64 bp downstream of the CS of the δ-globin gene (*HBD*; MIM# 142000; Supplementary Table S5) was identified in subjects with δ-thalassaemia (Moi et al. 1992). The wild-type $\underline{\text{G}}$ contributes to the sequence TACAGATA$\underline{\text{G}}$G which is very similar to the GATA box core sequence (A/T)GATA(A/G), a binding site for GATA-1. That the GATA-1 protein binds more tightly to the *HBD* mutant A allele than to its wild-type counterpart (Moi et al. 1992) suggests that GATA-1 may act as a transcriptional repressor in this specific context (Cao and Moi 2002).

SLC9A3R1: a SNP that disrupts a putative RUNX1-binding site

Psoriasis is a common, immunologically-mediated, hyperproliferative skin disorder that is inherited as a multifactorial trait. Recently, several non-coding SNPs within the second susceptibility locus, *PSORS2* (MIM# 602723), have been reported to be associated with the disease (Helms et al. 2003). The A allele of SNP9, which lies 237 bp downstream of the CS of the *SLC9A3R1* gene (MIM# 604990) and serves to eliminate a putative binding site for the Runt-related transcription factor, RUNX1 (see Supplementary Table S6), has received special attention for two reasons. First, *SLC9A3R1* encodes a PDZ domain-containing phosphoprotein that is implicated in diverse aspects of epithelial membrane biology and immune synapse formation in T cells (Itoh et al. 2002). Second, DNA variants in RUNX1 [a transcription factor that is essential for haematopoietic and endothelial cell development; Lacaud et al. 2002)] binding sites have been associated with systematic lupus erythematosus (Prokunina et al. 2002) and rheumatoid arthritis (Tokuhiro et al. 2003). Interestingly, electrophoretic mobility shift assays indicated specific binding of nuclear extracts from a human T cell line (Jurkat) to the wild-type RUNX1-binding site (G allele), but not to the psoriasis-associated A allele at any concentration of nuclear extract. Moreover, although no significant differences in reporter gene activity were observed in transient transfection assays in the Jurkat T cell line with constructs, respectively, expressing the A or G alleles, the wild-type G allele showed the most prominent increase in reporter gene activity when RUNX1 and its coactivator CBFβ were co-transfected (Helms et al. 2003).

The above observations have strongly suggested that the SNP9-associated RUNX1-binding site has a functional role. However, this disease association has not been confirmed by subsequent studies (Capon et al. 2004; Butt et al. 2005; Hosomi et al. 2005; Huffmeier et al. 2005; Hwu et al. 2005; Morar et al. 2006; Stuart et al. 2006).

Isolated examples

DMPK and *SCA8*: trinucleotide CTG repeat expansions within the 3′ UTR

CTG repeat expansion within the 3′-UTR of the dystrophia myotonica protein kinase gene (*DMPK*; MIM# 605377) causes autosomal dominant myotonic dystrophy, type 1 (Brook et al. 1992). This constitutes a novel pathological mechanism resulting from a triplet repeat expansion that acts at the RNA level: CUG expansion sequesters essential cellular RNA-binding proteins resulting in abnormal splicing of multiple transcripts involved in the disease (Day and Ranum 2005). The CTG expansion within the 3′-UTR of the *SCA8* gene (MIM# 603680) causing spinocerebellar ataxia type 8 (Koob et al. 1999) may involve a similar pathogenic RNA mechanism (Mutsuddi et al. 2004).

FCMD: a large LINE-1 *trans*-driven SVA insertion in the 3′-UTR

Disease-causing mutations resulting from long interspersed element-1 (LINE-1 or L1)-dependent retrotransposition constitute ~0.1% of the lesions listed in HGMD (Stenson et al. 2003; Chen et al. 2005a). Of the ~50 such events known (Chen et al. 2005b), one involves the integration of a 3,062 bp SVA element into the 3′-UTR of the *FCMD* gene (MIM# 607440) causing autosomal recessive Fukuyama-type congenital muscular dystrophy (Kobayashi et al. 1998). It is thought that this large SVA insertion may either inhibit transcriptional elongation or cause abnormal polyadenylation, resulting in the complete loss of gene expression (reviewed in Chen et al. 2006).

GLA: a human gene lacking a 3′-UTR

The human *GLA* gene (MIM# 301500) is thought to be unique on account of the complete absence of a 3′ UTR: the polyadenylation signal ATTAAA actually lies within the coding sequence whilst the translational termination codon TAA corresponds to the end of the mature mRNA sequence. Two micro-deletions at the 3′ end of the *GLA* gene—a dinucleotide (AA) deletion occurred within the UCPAS ATTAAA and a tetranucleotide (ACTT) deletion located immediately upstream of the translational termination codon TAA—were identified separately in two unrelated men with classical Fabry disease. Both mutations were found to affect adversely 3′ end formation of the *GLA* mRNA (Yasuda et al. 2003).

HLA-G: a 14 bp deletion/insertion length polymorphism within the 3′ UTR regulates alternative splicing and mRNA stability

Human leukocyte antigen-G (HLA-G; MIM# 142871), a non-classical major histocompatibility complex (MHC) class I molecule, is characterised by reduced polymorphism (15 alleles) and only seven different protein isoforms (LeMaoult et al. 2003). The presence of a 14 bp insertion polymorphism in the 3′ UTR of the *HLA-G* gene (Harrison et al. 1993) in certain alleles promotes a further splicing event viz. a total of 92 bases flanking the insertion site was spliced out due to the activation of a downstream cryptic acceptor splice site (Hiby et al. 1999; see also Fig. 1 in Rousseau et al. 2003); this also affects mRNA stability (Hviid et al. 2003; Rousseau et al. 2003). The 14 bp insertion/deletion length polymorphism may play a role in recurrent spontaneous abortion, pre-eclampsia, and outcome of in vitro fertilization (Hviid 2004; Hviid et al. 2004; Tripathi et al. 2004; Hviid and Christiansen 2005).

SLC6A3: controversial association of a 3′ UTR VNTR polymorphism with ADHD

Given its important role in regulating dopamine neurotransmission, the dopamine transporter gene (*SLC6A3*; MIM# 126455) has been extensively studied as one of the candidate genes that could predispose to attention-deficit/hyperactivity disorder (ADHD; Madras et al. 2005). Not only is the association of the 10-repeat allele of a 40-bp variable number of tandem repeat (VNTR) polymorphism in the *SLC6A3* 3′ UTR with ADHD controversial (e.g. Bellgrove et al. 2005; Feng et al. 2005; Langley et al. 2005; Purper-Ouakil et al. 2005; Kim et al. 2006), but inconsistent results have also been generated in in vitro transient expression analyses of the 10-repeat allele (e.g. Fuke et al. 2001; Mill et al. 2005; VanNess et al. 2005).

Concluding remarks and perspectives

Despite their evident complexity both in terms of biology and pathology, we have performed the first systematic analysis of the known (and highly diverse) naturally-occurring 3′ RR variants. This has generated many useful insights, some of which will be briefly discussed below.

3′ RR variants do not usually lead to the complete loss of gene expression

With the exception of the rather unusual cases of *DMPK*, *SCA8* and *FCMD* mutations, none of the known disease-associated 3′ RR variants/mutations result in the complete functional loss of the genes involved. This is perhaps not altogether surprising since even the highly conserved AATAAA hexamer appears not to be an absolute requirement for correct mRNA 3′ end formation, let alone the other functional elements.

Functionality versus causality

Most of the USS variants studied might reasonably have been expected to have had some functional consequences for gene expression, a conclusion based upon the variants' high probability of changing either the primary sequences of a variety of regulatory motifs present within this region or/and the associated mRNA secondary structures (J.M. Chen, C. Férec, D.N. Cooper, submitted). However, most of these variants as well as those in the LAS and the 3′ FR have usually been found to exert a comparatively minor influence on mRNA expression. This could perhaps provide a partial explanation for why such a high proportion of initially reported disease associations involving 3′ RR variants end up not being independently confirmed; these variants may predispose to, 'protect from', or modify disease susceptibility but only in combination with other genetic and environmental factors, which may vary quite dramatically between different populations.

Our current understanding of the functionality of 3′ RR variants is far from complete

Our current understanding of the functional roles of the diverse 3′ RR variants (and accordingly the biology of the 3′ RR) is far from complete. For example, none of the in vitro characterised functional USS variants have yet been found to affect mRNA cellular localisation. It may be that this is due to the still rather limited number of such variants which are known. However, any such effects might well have been overlooked in studies that have focussed almost exclusively on the analysis of mRNA stability and/or the level of the synthesized protein. Different types of enhancers, repressors and silencers should also exist within the 3′ UTRs of human genes. However, owing to the presence of well characterized functional elements that operate post-transcriptionally, the identification of such bona fide regulatory elements is likely to be rather challenging. The 3′ UTR is known to represent a preferred target for regulation by *cis*-encoded natural antisense transcripts (Sun et al. 2005), a phenomenon that has so far scarcely been investigated and which could yet prove to be quite widespread. It may be that a given 3′ UTR motif will turn out to have multiple roles, depending upon the cell type, subcellular location, developmental stage, and the concomitant availability or otherwise of *trans*-acting protein or RNA factors. We would also like to emphasize that whilst miRNA-binding sites are likely to exist in the 3′ UTRs of many human genes (Lim et al. 2005), much of the new lexicon of miRNA-binding sites still remains to be defined. Many of these sites could harbour mutations such as that found to cause Tourette syndrome (Abelson et al. 2005) and it may be that this category of lesion will turn out to be especially important in certain disease states.

Toward the identification of all functional regulatory elements in the human genome

The study of naturally occurring disease-associated mutations in the 3′ RRs of human genes should facilitate the identification of novel regulatory sequence elements within these regions. Conversely, our improving knowledge of these long neglected regulatory regions should help to guide mutation screening programs such that an increasing number of mutations within 3′ RRs are likely to come to clinical attention in the coming years. This parallelism should ensure not only that the mechanisms through which newly detected 3′ RR mutations exert their pathological effects will become steadily clearer, but also that we shall in time arrive at a better understanding of how both the primary and secondary structure of these regions influences their normal function.

Of the different approaches to studying *cis*-acting regulatory elements, each has its own strengths and weaknesses (Pastinen and Hudson 2004). The use of all these approaches in concert promises to provide us with new insights into the emerging vocabulary of 3′ RR elements. It is hoped that the integrative approach employed here in the study of naturally occurring variants of actual or potential pathological significance will serve to complement ongoing efforts to identify all functional regulatory elements in the human genome (e.g. ENCODE Project Consortium 2004).

Acknowledgements JMC is a visiting Professor of Genetics supported by the Ministère de la Jeunesse, de l'Éducation Nationale et de la Recherche, France. This work was supported by the INSERM (Institut National de la Santé et de la Recherche Médicale), France. The authors are most grateful to Peter Stenson for provision of HGMD data.

References

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS 4th, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW (2005) Sequence variants in *SLITRK1* are associated with Tourette's syndrome. Science 310:317–320

Amrani N, Dong S, He F, Ganesan R, Ghosh S, Kervestin S, Li C, Mangus DA, Spatrick P, Jacobson A (2006) Aberrant termination triggers nonsense-mediated mRNA decay. Biochem Soc Trans 34:39–42

Anjos SM, Tessier MC, Polychronakos C (2004) Association of the cytotoxic T lymphocyte-associated antigen 4 gene with type 1 diabetes: evidence for independent effects of two polymorphisms on the same haplotype block. J Clin Endocrinol Metab 89:6257–6265

Balim Z, Kosova B, Falzon K, Bezzina Wettinger S, Colak Y (2003) Budd-Chiari syndrome in a patient heterozygous for the point mutation C20221T of the prothrombin gene. J Thromb Haemost 1:852–853

Baralle D, Baralle M (2005) Splicing in action: assessing disease causing sequence changes. J Med Genet 42:737–748

Beaudoing E, Gautheret D (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res 11:1520–1526

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res 10:1001–1010

Bellgrove MA, Hawi Z, Kirley A, Fitzgerald M, Gill M, Robertson IH (2005) Association between dopamine transporter (*DAT1*) genotype, left-sided inattention, and an enhanced response to methylphenidate in attention-deficit hyperactivity disorder. Neuropsychopharmacology 30:2290–2297

Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, Shigeoka AO, Ochs HD, Chance PF (2001) A rare polyadenylation signal mutation of the *FOXP3* gene (AAUAAA → AAUGAA) leads to the IPEX syndrome. Immunogenetics 53:435–439

Beyer K, Dandekar T, Keller W (1997) RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3′-end processing of pre-mRNA. J Biol Chem 272:26769–26779

Birnstiel ML, Busslinger M, Strub K (1985) Transcription termination and 3′ processing: the end is in site! Cell 41:349–359

Blum M, Grant DM, McBride W, Heim M, Meyer UA (1990) Human arylamine *N*-acetyltransferase genes: isolation, chromosomal localization, and functional expression. DNA Cell Biol 9:193–203

Bonapace G, Concolino D, Formicola S, Strisciuglio P (2003) A novel mutation in a patient with insulin-like growth factor 1 (IGF1) deficiency. J Med Genet 40:913–917

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, Sohn R, Zemelman B, Snell RG, Rundle SA, Crow S, Davies J, Shelbourne P, Buxton J, Jones C, Juvonen V, Johnson K, Harper PS, Shaw DJ, Housman DE (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell 68:799–808

Butt C, Sun S, Greenwood C, Gladman D, Rahman P (2005) Lack of association of *SLC22A4*, *SLC22A5*, *SLC9A3R1* and RUNX1 variants in psoriatic arthritis. Rheumatology (Oxford) 44:820–821

Cao A, Moi P (2002) Regulation of the globin genes. Pediatr Res 51:415–421

Capon F, Helms C, Veal CD, Tillman D, Burden AD, Barker JN, Bowcock AM, Trembath RC (2004) Genetic analysis of *PSORS2* markers in a UK dataset supports the association between *RAPTOR* SNPs and familial psoriasis. J Med Genet 41:459–460

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y FANTOM Consortium; RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group) (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Ceelie H, Spaargaren-van Riel CC, Bertina RM, Vos HL (2004) G20210A is a functional mutation in the prothrombin gene; effect on protein levels and 3′-end formation. J Thromb Haemost 2:119–127

Ceelie H, Spaargaren-Van Riel CC, Lyon E, Bertina RM, Vos HL (2005) Functional analysis of two polymorphisms in the 3′-UTR of the human prothrombin gene. J Thromb Haemost 3:806–808

Chabanon H, Mickleburgh I, Hesketh J (2004) Zipcodes and postage stamps: mRNA localisation signals and their *trans*-acting binding proteins. Brief Funct Genomic Proteomic 3:240–256

Chen F, MacDonald CC, Wilusz J (1995) Cleavage site determinants in the mammalian polyadenylation signal. Nucleic Acids Res 23:2614–2620

Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN (2005a) Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat 25:207–221

Chen JM, Stenson PD, Cooper DN, Ferec C (2005b) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet 117:411–427

Chen JM, Ferec C, Cooper DN (2006) LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. J Biomed Biotechnol (http://www.hindawi.com/GetSpecialIssueArticles.aspx?journal = JBB&volume = 2006&si = 1)

Coller J, Parker R (2004) Eukaryotic mRNA decapping. Annu Rev Biochem 73:861–890

Conne B, Stutz A, Vassalli JD (2000) The 3′ untranslated region of messenger RNA: a molecular 'hotspot' for pathology? Nat Med 6:637–641

Danckwardt S, Gehring NH, Neu-Yilik G, Hundsdoerfer P, Pforsich M, Frede U, Hentze MW, Kulozik AE (2004) The prothrombin 3′ end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. Blood 104:428–435

Day JW, Ranum LP (2005) RNA pathogenesis of the myotonic dystrophies. Neuromusc Disord 15:5–16

Dye MJ, Proudfoot NJ (2001) Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. Cell 105:669–681

Edwalds-Gilbert G, Veraldi KL, Milcarek C (1997) Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res 25:2547–2561

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640

Feng Y, Wigg KG, Makkar R, Ickowicz A, Pathare T, Tannock R, Roberts W, Malone M, Kennedy JL, Schachar R, Barr CL (2005) Sequence variation in the 3′-untranslated region of the dopamine transporter gene and attention-deficit hyperactivity disorder (ADHD). Am J Med Genet B Neuropsychiatr Genet 139:1–6

Fornace AJ Jr, Cummings DE, Comeau CM, Kant JA, Crabtree GR (1984) Structure of the human gamma-fibrinogen gene. Alternate mRNA splicing near the 3′ end of the gene produces gamma A and gamma B forms of gamma-fibrinogen. J Biol Chem 259:12826–12830

Fuke S, Suo S, Takahashi N, Koike H, Sasagawa N, Ishiura S (2001) The VNTR polymorphism of the human dopamine transporter (*DAT1*) gene affects gene expression. Pharmacogenomics J 1:152–156

Furugaki K, Shirasawa S, Ishikawa N, Ito K, Ito K, Kubota S, Kuma K, Tamai H, Akamizu T, Hiratani H, Tanaka M, Sasazuki T (2004) Association of the T-cell regulatory gene *CTLA4* with Graves' disease and autoimmune thyroid disease in the Japanese. J Hum Genet 49:166–168

Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, Kulozik AE (2001) Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat Genet 28:389–392

Gieselmann V, Polten A, Kreysing J, von Figura K (1989) Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and *N*-glycosylation site. Proc Natl Acad Sci USA 86:9436–9440

Gilmartin GM (2005) Eukaryotic mRNA 3′ processing: a common means to different ends. Genes Dev 19:2517–2521

Giordano PC, Bouva MJ, Van Delft P, Akkerman N, Kappers-Klunne MC, Harteveld CL (2005) A new polyadenylation site mutation associated with a mild beta-thalassemia phenotype. Haematologica 90:551–552

Graber JH, Cantor CR, Mohr SC, Smith TF (1999) *In silico* detection of control signals: mRNA 3′-end-processing sequences in diverse species. Proc Natl Acad Sci USA 96:14055–14060

Gu J, Liang D, Wang Y, Lu C, Wu X (2005) Effects of *N*-acetyl transferase 1 and 2 polymorphisms on bladder cancer risk in Caucasians. Mutat Res 581:97–104

Hall-Pogar T, Zhang H, Tian B, Lutz CS (2005) Alternative polyadenylation of cyclooxygenase-2. Nucleic Acids Res 33:2565–2579

Hao K, Niu T, Xu X, Fang Z, Xu X (2005) Single-nucleotide polymorphisms of the *KCNS3* gene are significantly associated with airway hyperresponsiveness. Hum Genet 116:378–383

Harrison GA, Humphrey KE, Jakobsen IB, Cooper DW (1993) A 14 bp deletion polymorphism in the *HLA-G* gene. Hum Mol Genet 2:2200

Harteveld CL, Losekoot M, Haak H, Heister GA, Giordano PC, Bernini LF (1994) A novel polyadenylation signal mutation in the alpha 2-globin gene causing alpha thalassaemia. Br J Haematol 87:139–143

Harvey JS, Carey WF, Morris CP (1998) Importance of the glycosylation and polyadenylation variants in metachromatic leukodystrophy pseudodeficiency phenotype. Hum Mol Genet 7:1215–1219

Hayashi S, Watanabe J, Nakachi K, Kawajiri K (1991) Genetic linkage of lung cancer-associated *Msp*I polymorphisms with amino acid replacement in the heme binding region of the human cytochrome P450IA1 gene. J Biochem (Tokyo) 110:407–411

Heath JA, Beaverson K, Giardina P, Boehm C, Cutting G (2001) A novel beta-thalassemia intermedia phenotype containing Nt494 + 129T → C and NT494 + 132C → A mutations in *cis* and a Nt168C → T (β^o 39 point) mutation in *trans*. Am J Hematol 67:57–58

Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JA, Robarge J, Ott J, Kwok PY, Menter A, Bowcock AM (2003) A putative RUNX1 binding site variant between *SLC9A3R1* and *NAT9* is associated with susceptibility to psoriasis. Nat Genet 35:349–356

Hiby SE, King A, Sharkey A, Loke YW (1999) Molecular studies of trophoblast *HLA-G*: polymorphism, isoforms, imprinting and expression in preimplantation embryo. Tissue Antigens 53:1–13

Higgs DR, Goodbourn SE, Lamb J, Clegg JB, Weatherall DJ, Proudfoot NJ (1983) Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature 306:398–400

Hosomi N, Fukai K, Oiso N, Kato A, Fukui M, Ishii M (2005) No association between atopic dermatitis and the *SLC9A3R1-NAT9* RUNX1 binding site polymorphism in Japanese patients. Clin Exp Dermatol 30:192–193

Hsu AP, Tsai EJ, Anderson SM, Fischer RE, Malech H, Buckley RH, Puck JM (2000) Unusual X-linked SCID phenotype due to mutation of the poly-A addition signal of *IL2RG* (abstract 206). Am J Hum Genet 67(Suppl 2):50

Hu J, Lutz CS, Wilusz J, Tian B (2005) Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. RNA 11:1485–1493

Huffmeier U, Traupe H, Burkhardt H, Schurmeier-Horst F, Lascorz J, Bohm B, Lohmann J, Stander M, Wendler J, Kelsch R, Baumann C, Kuster W, Wienker TF, Reis A (2005) Lack of evidence for genetic association to RUNX1 binding site at *PSORS2* in different German psoriasis cohorts. J Invest Dermatol 124:107–110

Hviid TV (2004) *HLA-G* genotype is associated with fetoplacental growth. Hum Immunol 65:586–593

Hviid TV, Christiansen OB (2005) Linkage disequilibrium between human leukocyte antigen (HLA) class II and HLA-G–possible implications for human reproduction and autoimmune disease. Hum Immunol 66:688–699

Hviid TV, Hylenius S, Rorbye C, Nielsen LG (2003) *HLA-G* allelic variants are associated with differences in the *HLA-G* mRNA isoform profile and *HLA-G* mRNA levels. Immunogenetics 55:63–79

Hviid TV, Rizzo R, Christiansen OB, Melchiorri L, Lindhard A, Baricordi OR (2004) HLA-G and IL-10 in serum in relation to *HLA-G* genotype and polymorphisms. Immunogenetics 56:135–141

Hwu WL, Yang CF, Fann CS, Chen CL, Tsai TF, Chien YH, Chiang SC, Chen CH, Hung SI, Wu JY, Chen YT (2005) Mapping of psoriasis to 17q terminus. J Med Genet 42:152–158

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda J, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo Mde F, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, de Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, Gopinath GR, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2:e162

Itakura H, Telen MJ, Hoppe CC, White DA, Zehnder JL (2005) Characterization of a novel prothrombin variant, Prothrombin C20209T, as a modifier of thrombotic risk among African-Americans. J Thromb Haemost 3:2357–2359

Itoh K, Sakakibara M, Yamasaki S, Takeuchi A, Arase H, Miyazaki M, Nakajima N, Okada M, Saito T (2002) Cutting edge: negative regulation of immune synapse formation by anchoring lipid raft to cytoskeleton through Cbp-EBP50-ERM assembly. J Immunol 168:541–544

Jacquette A, Le Roux G, Lacombe C, Goossens M, Pissard S (2004) Compound heterozygosity for two new mutations in the beta-globin gene [codon 9 (+TA) and polyadenylation site (AATAAA → AAAAAA)] leads to thalassemia intermedia in a Tunisian patient. Hemoglobin 28:243–248

Jankovic L, Efremov GD, Petkov G, Kattamis C, George E, Yang KG, Stoming TA, Huisman TH (1990) Two novel polyadenylation mutations leading to beta(+)-thalassemia. Br J Haematol 75:122–126

Kawajiri K, Nakachi K, Imai K, Yoshii A, Shinoda N, Watanabe J (1990) Identification of genetically high risk individuals to lung cancer by DNA polymorphisms of the cytochrome P450IA1 gene. FEBS Lett 263:131–133

Kim SJ, Martinson HG (2003) Poly(A)-dependent transcription termination: continued communication of the poly(A) signal with the polymerase is required long after extrusion *in vivo*. J Biol Chem 278:41691–41701

Kim JW, Kim BN, Cho SC (2006) The dopamine transporter gene and the impulsivity phenotype in attention deficit hyperactivity disorder: a case-control association study in a Korean sample. J Psychiatr Res [Epub ahead of print]

Kimberland ML, Boehm CD, Kazazian HH Jr (1995) Two novel beta-thalassemia alleles: poly A signal (AATAAA → AAAA) and -92 C → T. Hum Mutat 5:275–276

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, Hamano K, Sakakihara Y, Nonaka I, Nakagome Y, Kanazawa I, Nakamura Y, Tokunaga K, Toda T (1998) An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature 394:388–392

Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, Day JW, Ranum LP (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat Genet 21:379–384

Kreysing J, von Figura K, Gieselmann V (1990) Structure of the arylsulfatase A gene. Eur J Biochem 191:627–631

Kuhn U, Wahle E (2004) Structure and function of poly(A) binding proteins. Biochim Biophys Acta 1678:67–84

Lacaud G, Gore L, Kennedy M, Kouskoff V, Kingsley P, Hogan C, Carlsson L, Speck N, Palis J, Keller G (2002) Runx1 is essential for hematopoietic commitment at the hemangioblast stage of development in vitro. Blood 100:458–466

Langley K, Turic D, Peirce TR, Mills S, Van Den Bree MB, Owen MJ, O'Donovan MC, Thapar A (2005) No support for association between the dopamine transporter (*DAT1*) gene and ADHD. Am J Med Genet B Neuropsychiatr Genet 139:7–10

LeMaoult J, Le Discorde M, Rouas-Freiss N, Moreau P, Menier C, McCluskey J, Carosella ED (2003) Biology and functions of human leukocyte antigen-G in health and sickness. Tissue Antigens 62:273–284

de Leon JH, Vatsis KP, Weber WW (2000) Characterization of naturally occurring and recombinant human *N*-acetyltransferase variants encoded by *NAT1*. Mol Pharmacol 58:288–299

Liao G, Wang J, Guo J, Allard J, Cheng J, Ng A, Shafer S, Puech A, McPherson JD, Foernzler D, Peltz G, Usuka J (2004) *In silico* genetics: identification of a functional element regulating *H2-Ealpha* gene expression. Science 306:690–695

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433:769–773

Madras BK, Miller GM, Fischman AJ (2005) The dopamine transporter and attention-deficit/hyperactivity disorder. Biol Psychiatry 57:1397–1409

Makalowski W, Zhang J, Boguski MS (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res 6:846–857

Meadows CA, Warner D, Page S, Lyon E (2002) Detection of novel mutation using fluorescent hybridization probes and melting temperature analysis (abstract). J Mol Diagn 3:195

Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet 36:1073–1078

Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3:REVIEWS0004

Mill J, Asherson P, Craig I, D'Souza UM (2005) Transient expression analysis of allelic variants of a VNTR in the dopamine transporter gene (*DAT1*). BMC Genet 6:3

Missirlis PI, Mead CL, Butland SL, Ouellette BF, Devon RS, Leavitt BR, Holt RA (2005) Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. BMC Bioinformatics 6:145

Moi P, Loudianos G, Lavinha J, Murru S, Cossu P, Casu R, Oggiano L, Longinotti M, Cao A, Pirastu M (1992) Delta-thalassemia due to a mutation in an erythroid-specific binding protein sequence 3′ to the delta-globin gene. Blood 79:512–516

Moore CL, Skolnik-David H, Sharp PA (1986) Analysis of RNA cleavage at the adenovirus-2 L3 polyadenylation site. EMBO J 5:1929–1938

Morar N, Bowcock AM, Harper JI, Cookson WO, Moffatt MF (2006) Investigation of the chromosome 17q25 *PSORS2* locus in atopic dermatitis. J Invest Dermatol 126:603–606

Moreira A, Wollerton M, Monks J, Proudfoot NJ (1995) Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. EMBO J 14:3809–3819

Morgan K, Scobie G, Kalsheker NA (1993) Point mutation in a 3′ flanking sequence of the alpha-1-antitrypsin gene associated with chronic respiratory disease occurs in a regulatory sequence. Hum Mol Genet 2:253–257

Morgan K, Scobie G, Marsters P, Kalsheker NA (1997) Mutation in an alpha1-antitrypsin enhancer results in an interleukin-6 deficient acute-phase response due to loss of cooperativity between transcription factors. Biochim Biophys Acta 1362:67–76

Muhlrad D, Parker R (1999) Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance. RNA 5:1299–1307

Mutsuddi M, Marshall CM, Benzow KA, Koob MD, Rebay I (2004) The spinocerebellar ataxia 8 noncoding RNA causes neurodegeneration and associates with staufen in *Drosophila*. Curr Biol 14:302–308

Orkin SH, Cheng TC, Antonarakis SE, Kazazian HH Jr (1985) Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. EMBO J 4:453–456

Pastinen T, Hudson TJ (2004) *Cis*-acting regulatory variation in the human genome. Science 306:647–650

Pauws E, van Kampen AH, van de Graaf SA, de Vijlder JJ, Ris-Stalpers C (2001) Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res 29:1690–1694

Pickering BM, Willis AE (2005) The implications of structured 5′ untranslated regions on translation and disease. Semin Cell Dev Biol 16:39–47

Pizzuti A, Argiolas A, Di Paola R, Baratta R, Rauseo A, Bozzali M, Vigneri R, Dallapiccola B, Trischitta V, Frittitta L (2002) An ATG repeat in the 3′-untranslated region of the human resistin gene is associated with a decreased risk of insulin resistance. J Clin Endocrinol Metab 87:4403–4406

Plant KE, Dye MJ, Lafaille C, Proudfoot NJ (2005) Strong polyadenylation and weak pausing combine to cause efficient termination of transcription in the human Ggamma-globin gene. Mol Cell Biol 25:3276–3285

Pollak ES, Lam HS, Russell JE (2002) The G20210A mutation does not affect the stability of prothrombin mRNA *in vivo*. Blood 100:359–362

Poort SR, Rosendaal FR, Reitsma PH, Bertina RM (1996) A common genetic variation in the 3′-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. Blood 88:3698–3703

Prokunina L, Castillejo-Lopez C, Oberg F, Gunnarsson I, Berg L, Magnusson V, Brookes AJ, Tentler D, Kristjansdottir H, Grondal G, Bolstad AI, Svenungsson E, Lundberg I, Sturfelt G, Jonssen A, Truedsson L, Lima G, Alcocer-Varela J, Jonsson R, Gyllensten UB, Harley JB, Alarcon-Segovia D, Steinsson K, Alarcon-Riquelme ME (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. Nat Genet 32:666–669

Purper-Ouakil D, Wohl M, Mouren MC, Verpillat P, Ades J, Gorwood P (2005) Meta-analysis of family-based association studies between the dopamine transporter gene and attention deficit hyperactivity disorder. Psychiatr Genet 15:53–59

Rousseau P, Le Discorde M, Mouillot G, Marcou C, Carosella ED, Moreau P (2003) The 14 bp deletion-insertion polymorphism in the 3′ UT region of the *HLA-G* gene influences *HLA-G* mRNA stability. Hum Immunol 64:1005–1010

Rund D, Dowling C, Najjar K, Rachmilewitz EA, Kazazian HH Jr, Oppenheim A (1992) Two mutations in the beta-globin polyadenylylation signal reveal extended transcripts and new RNA polyadenylylation sites. Proc Natl Acad Sci USA 89:4324–4328

Sachchithananthan M, Stasinopoulos SJ, Wilusz J, Medcalf RL (2005) The relationship between the prothrombin upstream sequence element and the G20210A polymorphism: the influence of a competitive environment for mRNA 3′-end formation. Nucleic Acids Res 33:1010–1020

Sato M, Sato T, Izumo T, Amagasa T (1999) Genetic polymorphism of drug-metabolizing enzymes and susceptibility to oral cancer. Carcinogenesis 20:1927–1931

Schrijver I, Lenzi TJ, Jones CD, Lay MJ, Druzin ML, Zehnder JL (2003) Prothrombin gene variants in non-Caucasians with fetal loss and intrauterine growth retardation. J Mol Diagn 5:250–253

Sheets MD, Stephenson P, Wickens MP (1987) Products of *in vitro* cleavage and polyadenylation of simian virus 40 late pre-mRNAs. Mol Cell Biol 7:1518–1529

Sheets MD, Ogg SC, Wickens MP (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. Nucleic Acids Res 18:5799–5805

Shin HD, Kim LH, Park BL, Jung HS, Cho YM, Moon MK, Lee HK, Park KS (2003) Polymorphisms in fatty acid-binding protein-3 (*FABP3*) - putative association with type 2 diabetes mellitus. Hum Mutat 22:180

Soo PY, Patel RK, Best S, Arya R, Thein SL (2005) Detection of prothrombin gene polymorphism at position 20209 (PT20209C/T): pilot study in a black population in the United Kingdom. Thromb Haemost 93:179–180

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577–581

Stuart P, Nair RP, Abecasis GR, Nistor I, Hiremagalore R, Chia NV, Qin ZS, Thompson RA, Jenisch S, Weichenthal M, Janiga J, Lim HW, Christophers E, Voorhees JJ, Elder JT (2006) Analysis of RUNX1 binding site and *RAPTOR* polymorphisms in psoriasis: no evidence for association despite adequate power and evidence for linkage. J Med Genet 43:12–17

Sun M, Hurst LD, Carmichael GG, Chen J (2005) Evidence for a preferential targeting of 3′-UTRs by *cis*-encoded natural antisense transcripts. Nucleic Acids Res 33:5533–5543

Sussenbach JS, Steenbergh PH, Holthuizen P (1992) Structure and expression of the human insulin-like growth factor genes. Growth Regul 2:1–9

Takeshita T, Asao H, Ohtani K, Ishii N, Kumaki S, Tanaka N, Munakata H, Nakamura M, Sugamura K (1992) Cloning of the gamma chain of the human IL-2 receptor. Science 257:379–382

Tamary H, Klinger G, Shalmon L, Attias D, Fortina P, Kobayashi M, Surrey S, Zaizov R (1997) Alpha-thalassemia caused by a 16 bp deletion in the 3′ untranslated region of the alpha 2-globin gene including the first nucleotide of the poly A signal sequence. Hemoglobin 21:121–130

Tanimoto K, Hayashi S, Yoshiga K, Ichikawa T (1999) Polymorphisms of the *CYP1A1* and *GSTM1* gene involved in oral squamous cell carcinoma in association with a cigarette dose. Oral Oncol 35:191–196

Tian B, Hu J, Zhang H, Lutz CS (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res 33:201–212

Tokuhiro S, Yamada R, Chang X, Suzuki A, Kochi Y, Sawada T, Suzuki M, Nagasaki M, Ohtsuki M, Ono M, Furukawa H, Nagashima M, Yoshino S, Mabuchi A, Sekine A, Saito S, Takahashi A, Tsunoda T, Nakamura Y, Yamamoto K (2003) An intronic SNP in a RUNX1 binding site of *SLC22A4*, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nat Genet 35:341–348

Tripathi P, Abbas A, Naik S, Agrawal S (2004) Role of 14-bp deletion in the *HLA-G* gene in the maintenance of pregnancy. Tissue Antigens 64:706–710

Tsai EJ, Malech HL, Kirby MR, Hsu AP, Seidel NE, Porada CD, Zanjani ED, Bodine DM, Puck JM (2002) Retroviral transduction of *IL2RG* into CD34(+) cells from X-linked severe combined immunodeficiency patients permits human T- and B-cell development in sheep chimeras. Blood 100:72–79

Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KM, Smith AN, Di Genova G, Herr MH, Dahlman I, Payne F, Smyth D, Lowe C, Twells RC, Howlett S, Healy B, Nutland S, Rance HE, Everett V, Smink LJ, Lam AC, Cordell HJ, Walker NM, Bordin C, Hulme J, Motzo C, Cucca F, Hess JF, Metzker ML, Rogers J, Gregory S, Allahabadia A, Nithiyananthan R, Tuomilehto-Wolf E, Tuomilehto J, Bingley P, Gillespie KM, Undlien DE, Ronningen KS, Guja C, Ionescu-Tirgoviste C, Savage DA, Maxwell AP, Carson DJ, Patterson CC, Franklyn JA, Clayton DG, Peterson LB, Wicker LS, Todd JA, Gough SC (2003) Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. Nature 423:506–511

Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, Bertina RM (2005) Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. Blood 106:4176–4183

VanNess SH, Owens MJ, Kilts CD (2005) The variable number of tandem repeats element in *DAT1* regulates *in vitro* dopamine transporter density. BMC Genet 6:55

Venkataraman K, Brown KM, Gilmartin GM (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev 19:1315–1327

Wang H, Zhang H, Jia Y, Zhang Z, Craig R, Wang X, Elbein SC (2004) Adiponectin receptor 1 gene (*ADIPOR1*) as a candidate for type 2 diabetes and insulin resistance. Diabetes 53:2132–2136

Warshawsky I, Hren C, Sercia L, Shadrach B, Deitcher SR, Newton E, Kottke-Marchant K (2002) Detection of a novel point mutation of the prothrombin gene at position 20209. Diagn Mol Pathol 11:152–156

Waye JS, Eng B, Patterson M, Reis MD, Macdonald D, Chui DH (2001) Novel beta-thalassemia mutation in a beta-thalassemia intermedia patient. Hemoglobin 25:103–105

West AG, Fraser P (2005) Remote control of gene transcription. Hum Mol Genet 14(Spec No 1):R101–R111

Whitelaw E, Proudfoot N (1986) Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3′ end processing in the human alpha 2 globin gene. EMBO J 5:2915–2922

Wickens M, Stephenson P (1984) Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3′ end formation. Science 226:1045–1051

Wilusz J, Pettine SM, Shenk T (1989) Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. Nucleic Acids Res 17:3899–3908

Wolfenstein-Todel C, Mosesson MW (1981) Carboxy-terminal amino acid sequence of a human fibrinogen gamma-chain variant (gamma'). Biochemistry 20:6146–6149

Wylenzek M, Geisen C, Stapenhorst L, Wielckens K, Klingler KR (2001) A novel point mutation in the 3′ region of the prothrombin gene at position 20221 in a Lebanese/Syrian family. Thromb Haemost 85:943–944

Wylenzek C, Trubenbach J, Gohl P, Wildhardt G, Alkins S, Fausett MB, Decker J, Steinberger D (2005) Mutation screening for the prothrombin variant G20210A by melting point analysis with the Light Cycler system: atypical results, detection of the variant C20209T and possible clinical implications. Clin Lab Haematol 27:343–346

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434:338–345

Xiong Y, Connolly T, Futcher B, Beach D (1991) Human D-type cyclin. Cell 65:691–699

Yan J, Marr TG (2005) Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res 15:369–375

Yasuda M, Shabbeer J, Osawa M, Desnick RJ (2003) Fabry disease: novel alpha-galactosidase A 3′-terminal mutations result in multiple transcripts due to aberrant 3′-end formation. Am J Hum Genet 73:162–173

Yuregir GT, Aksoy K, Curuk MA, Dikmen N, Fei YJ, Baysal E, Huisman TH (1992) Hb H disease in a Turkish family resulting from the interaction of a deletional alpha-thalassaemia-1 and a newly discovered poly A mutation. Br J Haematol 80:527–532

Zarudnaya MI, Kolomiets IM, Potyahaylo AL, Hovorun DM (2003) Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res 31:1375–1386

Zhang MQ (1998) Statistical features of human exons and their flanking regions. Hum Mol Genet 7:919–932

Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63:405–445

Zhernakova A, Eerligh P, Barrera P, Weseloy JZ, Huizinga TW, Roep BO, Wijmenga C, Koeleman BP (2005) *CTLA4* is differentially associated with autoimmune diseases in the Dutch population. Hum Genet 118:58–66