Association of intronic repetition of *SLC26A4* gene with Hashimoto thyroiditis disease

SALIMA BELGUITH-MAALEJ[1]*, RIHAB KALLEL[1,2], MOUNA MNIF[3],
MOHAMED ABID[3], HAMMADI AYADI[1] AND HASSEN HADJ KACEM[1,2]

[1] *Unité Cibles pour le Diagnostic et la Thérapie, Centre de Biotechnologie de Sfax, Tunisie*
[2] *Laboratoire de Microorganismes et Biomolécules, équipe des Procédés de Criblage Moléculaires et Cellulaires, Centre de Biotechnologie de Sfax, Tunisie*
[3] *Service d'Endocrinologie, CHU, Hédi Chaker, Sfax, Tunisie*

Summary

Intronic microsatellites repeats were implicated in the pathogenic mechanisms of several diseases. *SLC26A4* gene, involved in the genetic susceptibility of autoimmune thyroid disease (AITD), harbours large non-coding introns. Using the tandem repeat finder (TRF) Software, two new polymorphic microsatellite markers, *rs59736472* and *rs57250751*, located at introns 10 and 20, respectively, were identified. A case-control design including 308 patients affected with AITD (146 GD, 90 HT and 72 PIM) and 212 unmatched healthy controls were performed for each marker (*rs59736472*, *D7S2459* and *rs57250751*). Furthermore, we used PHASE 2.0 version to reconstruct haplotypes, Kolmogorov–Smirnov (KS) and the Clump analysis program for multivariate analysis. The fluorescent genotyping revealed three alleles (106,112 and 115 bp) for *rs57250751* and 12 alleles for both *D7S2459* and *rs59736472* ranging from 134 to 156 bp and from 144 to 168 bp, respectively. The case-control analysis confirmed the positive association of *D7S2459* with Hashimoto thyroiditis (HT) disease previously reported. Moreover, a significant association was found only with *rs59736472* and HT disease. Haplotype-specific analysis showed that the 140-148-115 haplotype may increase the risk of HT ($\chi^2 = 9.8$, 1 DF, $P = 0.0017$, OR = 2.07, IC [1.27–3.36]). Consequently, considering the number of repetitions of both *D7S2459* and *rs59736472*, we found 15 alleles ranging from 45 to 59 repetitions. The case-control analysis showed a significant association of the 55 repetition with HT disease ($\chi^2 = 6.32$, 1 DF, $p_c = 0.012$, OR = 1.74, IC [1.1–2.76]). In conclusion, we suggest the association of longer alleles of intron 10 of *SLC26A4* gene with HT disease.

1. Introduction

The Pendred Syndrome gene, *PDS*, also named *SLC26A4* gene (MIM #605646), located on chromosome 7q31, encodes pendrin, a hydrophobic trans-membrane protein composed of 780 amino acids with 11–12 trans-membrane domains (Everett *et al.*, 1997; Royaux *et al.*, 2000). Pendrin is expressed in various tissues including principally thyroid, inner ear and kidney. Functional analysis showed that pendrin is an apical transporter of iodide in thyroid cells (Scott *et al.*, 1999; Bidart *et al.*, 2000; Belguith-Maalej *et al.*, 2011), and it mediates secretion of HCO_3^- into the endolymph in the cochlea (Wangemann *et al.*, 2007).

Using Tunisian families affected with autoimmune thyroid disease (AITD), Hadj Kacem *et al.* have shown the involvement of *SLC26A4* gene in the genetic susceptibility of these diseases (Hadj Kacem *et al.*, 2003).

Recently, we have shown a differential expression pattern of *SLC26A4* gene in AITD tissues. In fact, the *SLC26A4* gene expression follows the physiological state of thyroid: the expression pattern shift from high level in Grave's disease (GD) to reduced level in Hashimoto thyroiditis (HT) (Belguith-Maalej *et al.*, 2011). This pathogenic *SLC26A4* expression variation could be due either to *SLC26A4* sequence variations

* Corresponding author: Laboratoire de Microorganismes et Biomolécules, équipe des Procédés de criblage moléculaires et cellulaires, Centre de Biotechnologie de Sfax. Route Sidi Mansour Km 6. BP "1177", 3018, Sfax, Tunisia. Tel.: +216 74 871 816. Fax: +216 74 875 818. E-mail: belguithsalima@gmail.com

Table 1. *Primer pairs sequences used in SLC26A4 genotyping*

Marker name	Primer	Primer sequence	Product length (bp)
rs59736472	F (Fam)	CTGGGCGACAAGAGCTAGAC	144–168
	R	GGGCAAGTCACTGTTGGAGT	
D7S2459	F(Hex)	AAGAAGTGCATTGAGACTCC	134–156
	R	CCGCCTTAGTAAAACCC	
rs57250751	F(Hex)	GCTGGATTTCATCTTGTTGG	106–115
	R	CACTCCAGTCCAGGTGACAG	

and/or molecular environment (transcription factor, hormones and iodide level). Exploring the genomic sequence of *SLC26A4* gene seems to be primordial. Therefore, recently, the *SLC26A4* exonic sequences were explored in patients affected with GD (Hadj-Kacem *et al.*, 2010). In this study, Hadj Kacem *et al.* suggested an eventual defective effect of the identified polymorphism. On the other hand, investigating *SLC26A4* non-coding region is also important since large intronic sequence may include functional polymorphism, microRNA and may harbour the sequence of other genes (Zhou & Lin, 2008). Intronic microsatellite repeats constitute widespread regulatory elements of alternative splicing (Hui *et al.*, 2005), may affect the gene expression level by modulating the movement of RNA polymerase (Peck & Wang, 1985). Indeed, it has been reported that shorter CA-repeat length is associated with an increased expression of epidermal growth factor receptor (EGFR) in lung carcinoma (Sueoka-Aragane *et al.*, 2008; Suzuki *et al.*, 2008).

In this study, we identified two new polymorphic microsatellite markers located in intron 10 (*rs59736472*) close to *D7S2459* and in intron 20 (*rs57250751*). Our case-control association study showed a significant association of the *D7S2459* and *rs59736472* markers with autoimmune hypo-thyroid-itis. Moreover, we suggested the association of longer alleles of intron 10 with HT diseases.

2. Subjects and methods

(i) Subjects

One hundred and forty-six unrelated Tunisian patients with GD (32 men and 114 women), 90 unrelated Tunisian patients with HT (7 men and 83 women) and 72 unrelated Tunisian patients with primary idiopathic myxidoma PIM (8 men and 64 women) were included in this study. Data gathered from the patient group were compared with those obtained from a control population of 212 unrelated healthy subjects (110 men and 102 women) originating from the same area with no clinical evidence or family history of AITD and inflammatory joint disease. Informed consent was obtained from all participants.

(ii) Genetic studies: genotyping of newly identified microsatellite markers

Tandem Repeat Finder (TRF) Software was used to identify two new polymorphic microsatellite markers located at introns 10 ((AT)n(AC)nAAAC(AT)-nTCTC(AT)n) and 20 ((TTA)n). DNA from all patients was genotyped with these two markers, in addition to *D7S2459* previously explored (Hadj Kacem *et al.*, 2003).

Genetic analysis was performed on genomic DNA extracted from blood leucocytes using a standard phenol–chloroform procedure (Kawasaki, 1990). Fluorescent primers designed to amplify the microsatellite markers are listed in Table 1. Microsatellite markers were amplified in 96-well Falcon flexi plates by the PCR using Applied Biosystems thermocycler (GeneAmp® PCR system 2700). PCRs were performed in a final volume of 15 μl, containing 50 ng of genomic DNA, 0·16 μM of each primer and 9 μl of True Allele™ PCR Premix kit (Applied Biosystems, Foster City, CA).

Each PCR was performed using a hot-start procedure, and amplification was carried out using 38 cycles of denaturation at 94 °C for 30 s, with annealing at 55 °C for 30 s, and elongation at 72 °C for 45 s, followed by a final elongation at 72 °C for 7 min. Then 2 μl of each sample were denatured at 96 °C for 3 min in presence of 7·5 μl of formamide and 0·5 of GeneScan™–500 ROX™ Size Standard (Applied Biosystems, Warrington, UK). Fluorescently labelled alleles were analysed on an ABI PRISM® 3100-Avant genetic analyser (Applied Biosystems, HITACHI High-Technologies Corporation, Japan). For each microsatellite marker, four homozygotic DNAs were sequenced to verify allele sequences and sizes using Applied Biosystem ABI PRISM® 3100 Avant.

(iii) Statistical analysis

The allele distribution among unrelated patients versus controls was compared using the Chi-square (χ^2) test, Kolmogorov–Smirnov (KS) and the Clump analysis program (Sham & Curtis, 1995). The KS test was ensured by comparing the observed divergence (D_{obs}) to the critical divergence (D_{crit}) at 5% of

Table 2. *Allele frequencies of rs57250751 microsatellite marker in GD, HT and PIM patients and controls*

rs 57250751 alleles		Controls (n = 212)		GD[a] (n = 144)		HT[b] (n = 90)		PIM[c] (n = 66)	
N	Size (bp)	n	%	n	%	n	%	n	%
1	115	181	42·7	127	44·1	78	43·3	59	44·7
2	112	243	57·3	160	55·55	102	56·7	72	54·55
3	106	0	0	1	0·35	0	0	1	0·75

[a] Grave's disease.
[b] Hashimoto thyroiditis.
[c] Primary idiopathic myxidoma.

significance, calculated according to the following formula:

$$D_{\mathrm{crit}}(a = 0{\cdot}05) = 1{\cdot}36\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

The Clump analysis program is suited for the analysis of multi-allelic markers and has been designed to overcome problems occurring in sparse contingency tables. Statistical significance was reached when $P < 0{\cdot}05$, Mantel–Haenszel test and Fisher's exact test was used when necessary. Odds ratios (ORs) were calculated using the Epi-Info software (version 3.5.3).

Estimated heterozygosity (HETEst) and polymorphisms information content (PIC) were calculated by HET program provided by J. Ott (Utility Programs for Analysis of Genetic Linkage). The observed heterozygosity (HETObs) was calculated manually.

Haplotypes were inferred using PHASE 2.1.1, a computational tool based on Bayesian methods, developed by Stephens and associates (Stephens et al., 2001; Stephens & Donnelly, 2003; Stephens & Scheet, 2005). Linkage disequilibrium and Hardy–Weinberg disequilibrium were performed with GDA program by Paul O. Lewis and Dmitri Zaykin designed to accompany the book 'Genetic Data Analysis' by Bruce S. Weir (1996, Sinauer Associates).

3. Results and discussion

SLC26A4 gene plays a crucial role in thyroid hormonogenesis. Since 2003, we reported the involvement of *SLC26A4* gene in the genetic susceptibility of AITD using family and population-based designs (Hadj Kacem et al., 2003). In that study, the case-control analysis showed a significant association of the intragenic marker *D7S2459* with HT and/or PIM.

In the present work, we identified new microsatellite markers into the *SLC26A4* gene intronic sequences using TRF Software (Benson, 1999). Two new polymorphic microsatellite markers, *rs59736472* and *rs57250751*, located at introns 10 and 20, respectively, were found polymorphic (build129; http://

Table 3. *Association of rs 59736472 and D7S2459 microsatellite makers with HTA subgroup*

P values	rs 59736472[a]	D7S2459[b]	rs 57250751[c]
AITD-C*	0·031	0·728	0·44
GD-C	0·378	0·841	0·43
HT-C	0·144	0·003	0·88
PIM-C	0·156	0·29	0·17
HTA-C	0·019	0·025	0·47

[a] rs59736472: a significant association was observed with all AITD patients and with HTA subgroup.
[b] D7S2459: no significant association was observed with AITD, however, a positive association was found with HTA and especially with HT.
[c] rs57250751: no significant difference in allelic distribution between patients affected with AITD and control was observed.
*: Control.

www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs = 59736472 and http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs = 57250751, respectively).

A case-control study, based on 308 unrelated AITD patients and 212 controls, was carried out with the *rs59736472* and *rs57250751*, in addition to *D7S2459* previously explored (Hadj Kacem et al., 2003).

Three alleles with sizes 106, 112 and 115 bp were observed for the *rs57250751* (HETEst = 0·49, HETObs = 0·467, PIC = 0·37) (Table 2). The comparison of the allelic distribution between patients affected with AITD and control data showed no significant difference ($\chi^2 = 1{\cdot}63$, DF = 2, P = 44). This result was maintained with different subgroups (GD, HT and PIM) (Table 3). The *D7S2459* microsatellite marker polymorphism analysis revealed 12 alleles with size ranging from 134 to 156 bp. At 350 bp downstream of *D7S2459*, the *rs59736472* microsatellite marker polymorphism in *SLC26A4* gene revealed 12 alleles ranging from 29 to 41 repetitions (HETEst = 0·814, HETObs = 0·744, PIC = 0·788) (Table 4). While the D7S2459 was not significantly associated with AITD ($\chi^2 = 4{\cdot}44$, DF = 7, P = 0·728), a

Table 4. *Allele frequencies of rs59736472 microsatellite marker in GD, HT and PIM cases and controls*

rs 59736472 alleles		Controls (n=211)		GD (n=142)		HT (n=90)		Alleles PIM (n=70)	
N	Size (bp)	n	%	n	%	n	%	n	%
1	168	0	0	0	0	0	0	1	0·7
2	164	0	0	0	0	1	0·56	0	0
3	162	3	0·71	0	0	2	1·1	1	0·7
4	160	8	1·89	11	3·87	8	4·45	4	2·85
5	158	50	11·85	29	10·2	24	13·33	15	10·7
6	156	118	28	79	27·8	41	22·8	41	29·3
7	154	45	10·66	25	8·8	11	6·1	7	5
8	152	14	3·31	11	3·87	6	3·33	10	7·15
9	150	61	14·45	51	18	29	16·1	19	13·6
10	148	103	24·4	67	23·6	51	28·33	41	29·3
11	146	19	4·5	6	2·1	5	2·8	0	0
12	144	1	0·23	5	1·76	2	1·1	1	0·7

Table 5. *Haplotype distribution between controls and AITD patients*

	Control	HT[a]	TH+HTA[b]	PIM[c]	GD[d]
>5%					
138-148-115	24	2	8	6	15
140-148-115	50	39	57	19	27
144-156-112	56	20	36	17	39
146-150-115*	18	14	18	0	26
146-158-112	27	17	23	4	7
148-156-112	24	14	24	9	17
<5%					
	225	74	158	89	161

A case-control study was performed considering haplotypes with frequencies >5%. The remained haplotypes (<5%) were pooled together.
[a] $\chi^2 = 24\cdot04$, DF=6, $P=0\cdot001$.
[b] $\chi^2 = 11\cdot69$, DF=6, $P=0\cdot069$.
[c] $\chi^2 = 3\cdot77$, DF=5, $P=0\cdot58$.
[d] $\chi^2 = 13\cdot13$, DF=6, $P<0\cdot041$.
*: χ^2 analysis was performed considering haplotype 146-150-115 ranged with haplotypes with frequencies <5%

Fig. 1. *D7S2459* and *rs59736472* microsatellite markers located in intron 10 used in this study. In10 marker is the sum of the number of repeats observed in *D7S2459* and *rs59736472* microsatellite markers.

Case-control analysis in different subgroups, provided by PHASE, was in accordance with results mentioned above. In fact, a significant association was maintained with HTA ($P=0\cdot01$) and HT ($P=0\cdot03$). However, there are no difference in haplotype distribution between GD patients and control group ($P>0\cdot05$). In HT subgroup, the gender effect was excluded ($P>0\cdot05$).

To refine the analysis, a case-control study was performed considering haplotypes with frequencies >5%. The remained haplotypes (<5%) were pooled together (Table 5). As result, the statistical significance of the association with HTA was lost ($P>0\cdot05$); however, it was confirmed with HT patients ($\chi^2 = 24\cdot04$, DF=6, $P=0\cdot001$) (Table 5).

Haplotype specific analysis showed that the 140-148-115 haplotype may increase the risk of HT ($\chi^2 = 9\cdot8$, 1 DF, $P=0\cdot0017$, OR=2·07, IC [1·27–3·36]). The predispositional effect of this haplotype could be related either to a functional polymorphism in LD with the 'risky' haplotype, or to a pathogenic length of intron 10, generated by the two closed markers *D7S2459* and *rs59736472*.

In fact, several studies suggested that dinucleotide repeat inducing intron-length variability have influence on gene expression (mRNA splice, stability and half-life) (Croston *et al.*, 1991; Lu *et al.*, 1993; Beutler *et al.*, 1998; Bharaj *et al.*, 2000; Sharma *et al.*, 2007).

positive association was observed with HTA ($\chi^2 = 15\cdot95$, DF=7, $P=0\cdot025$) and particularly with HT ($\chi^2 = 19\cdot77$, DF=6, $P=0\cdot003$) (Table 3), confirming previously reported data gathered from 54 HT patients and 154 controls (Hadj Kacem *et al.*, 2003).

On the other hand, a significant association was observed with *rs59736472* and all AITD patients ($\chi^2 = 16\cdot939$, DF=8, $P=0\cdot031$) and HTA subgroup ($\chi^2 = 18\cdot34$; DF=8; $P=0\cdot019$). This association was not maintained with different subgroups (GD, HT or PIM) ($P>0\cdot05$) (Table 3).

Considering the three intronic markers and using PHASE program, we were able to determine 94 haplotypes with 60 in healthy control group, 55 in both patients affected with GD or HTA, 42 in HT patients and 45 in PIM patients.
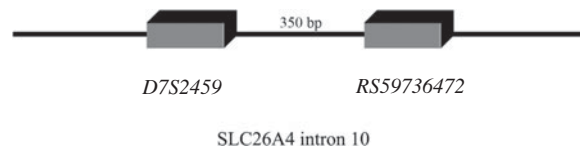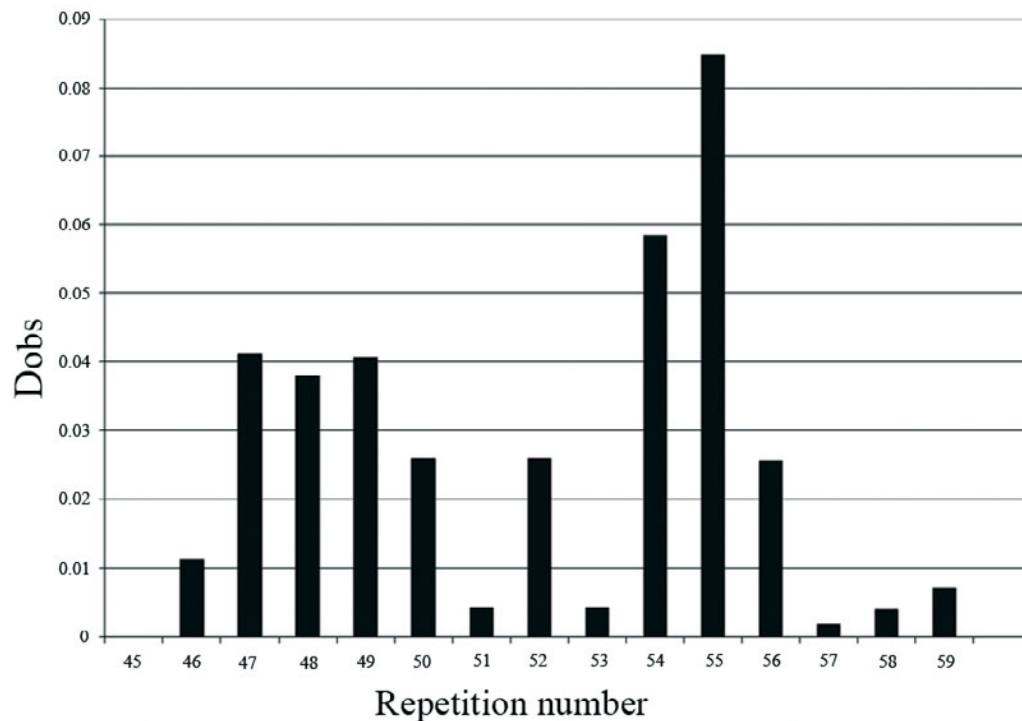
Fig. 2. KS representation: distribution of the repeat size in intron 10 of *SLC26A4* gene.

In this regard, our previous studies showed variability in *SLC26A4* gene expression among AITD patients (Belguith-Maalej *et al.*, 2011). Thus, on the assumption that the intron-length polymorphism could be pathogenic, identifying the risk length seems to be exciting.

In line with this suggestion, we assumed that In10 marker is the sum of the number of repeats observed in *D7S2459* and *rs59736472* microsatellite markers (Fig. 1). Considering the Best pairs option outputed by PHASE, we calculated the number of repetitions for both markers for each individual. In total, 15 alleles were observed with size ranging from 45 to 59 repeats (Table 6). Three major alleles were identified, the 47, 53 and 55, with a frequency more than 50% in patients and controls. GDA result indicated that the In10 marker is in Hardy–Weinberg equilibrium ($P = 0.04$).

The comparison of the allelic distribution (alleles >56 repeats were pooled together) between HT patients and controls data showed a significant difference ($\chi^2 = 24.52$; 11 DF; $P = 0.011$). However, no significant statistical result was found using the KS test ($D_{obs} = 0.08 < D_{crit} = 0.120$). The highest deviation of KS test was reached at allele 55 (with respect to the cumulative increasing order from the smallest to the longest allele) (Fig. 2). In addition, using clump analysis program, all Chi-squared obtained values (T1, T2, T3 and T4) are significant with 10 000 sets of simulations. We repeated the statistical analysis in clumping the major alleles whose lengths are ⩾55 in the same case of χ^2 test. Our result shows a

Table 6. *In10 dinucleotide repetition distribution between controls and HT subjects**

Alleles	Repetition number	Control	HT
1	59	7	5
2	58	34	5
3	57	63	41
4	56	13	6
5	55	11	2
6	54	28	8
7	53	45	23
8	52	34	9
9	51	96	31
10	50	30	8
11	49	55	34
12	48	4	6
13	47	0	1
14	46	1	1
15	45	3	0

*: In10 marker is the sum of the number of repeats observed in *D7S2459* and *rs59736472* microsatellite markers

predispositional effect imposed by the 55 allele ($\chi^2 = 6.32$, 1 DF, $P = 0.011$, $p_c = 0.012$, OR = 1.74, IC [1.1–2.76]). Thus, HT disease was significantly associated with longer dinucleotide repeat among intron 10. Several studies have shown that intronic long dinucleotide repeats ($n \geqslant 12$) are more likely to display length polymorphisms and may act as functional-regulator of transcription (Dib *et al.*, 1996; Wren *et al.*, 2000).

In fact, the change in *SLC26A4* intron 10 lengths could give rise to an aberrant splice and/or gene expression defect. In this regard, the length of EGFR CA repeat polymorphism in lung carcinoma is inversely related with level of EGFR protein expression in the carcinoma (Suzuki *et al.*, 2008). Moreover, increasing the length of intron may substantially delay mRNA maturation (Kandul & Noor, 2009; Farlow *et al.*, 2012). Given this interest, in future research we plan to explore the correlation between *SLC26A4* gene expression defect in AITD tissues (Belguith-Maalej *et al.*, 2011) and intron 10 length polymorphisms. Although *SLC26A4* gene intron 10 length displays a significant positive association with HT disease, it remains unclear whether the contribution of this variation in the genetic susceptibility to HT arises from its functional role or because it is in linkage disequilibrium with other polymorphisms. Thus, we also intend to examine the relationship between *SLC26A4* gene variation previously reported (Hadj-Kacem *et al.*, 2010) and this variation in intron 10.

In conclusion, our work highlights the association of polymorphisms of intron 10 length with HT disease. By its variable length, the intron 10 polymorphisms could be an additional functional candidate to explain the *SLC26A4* involvement in AITD diseases.

4. Declaration of interest

There are no conflicts of interest.

References

Belguith-Maalej, S., Rebuffat, S. A., Charfeddine, I., Mnif, M., Nadir, R. F., Abid, M., Ghorbel, A., Peraldi-Roux, S., Ayadi, H & Hadj-Kacem, H. (2011). SLC26A4 expression among autoimmune thyroid tissues. *Immunobiology* **216**, 571–578.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580.

Beutler, E., Gelbart, T. & Demina, A. (1998). Racial variability in the UDP-glucuronosyltransferase 1 (UGT1A1) promoter: a balanced polymorphism for regulation of bilirubin metabolism? *Proceedings of the National Academy of Sciences USA* **95**, 8170–8174.

Bharaj, B., Scorilas, A., Giai, M. & Diamandis, E. P. (2000). TA repeat polymorphism of the 5-alpha-reductase gene and breast cancer. *Cancer Epidemiology, Biomarkers and Prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **9**, 387–393.

Bidart, J. M., Mian, C., Lazar, V., Russo, D., Filetti, S., Caillou, B. & Schlumberger, M. (2000). Expression of pendrin and the Pendred syndrome (SLC26A4) gene in human thyroid tissues. *Journal of Clinical Endocrinology and Metabolism* **85**, 2028–2033.

Croston, G. E., Kerrigan, L. A., Lira, L. M., Marshak, D. R. & Kadonaga, J. T. (1991). Sequence-specific anti-repression of histone H1-mediated inhibition of basal RNA polymerase II transcription. *Science* **251**, 643–649.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154.

Everett, L. A., Glaser, B., Beck, J. C., Idol, J. R., Buchs, A., Heyman, M., Adawi, F., Hazani, E., Nassir, E., Baxevanis, A. D., Sheffield, V. C. & Green, E. D. (1997). Pendred syndrome is caused by mutations in a putative sulphate transporter gene (SLC26A4). *Nature Genetics* **17**, 411–422.

Farlow, A., Dolezal, M., Hua, L. & Schlotterer, C. (2012). The genomic signature of splicing-coupled selection differs between long and short introns. *Molecular Biology and Evolution* **29**, 21–24.

Hadj-Kacem, H., Kallel, R., Belguith-Maalej, S., Mnif, M., Charfeddine, I., Ghorbel, A., Abid, M., Ayadi, H. & Masmoudi, S. (2010). SLC26A4 variations among Graves' hyper-functioning thyroid gland. *Disease Markers* **29**, 63–69.

Hadj Kacem, H., Rebai, A., Kaffel, N., Masmoudi, S., Abid, M. & Ayadi, H. (2003). SLC26A4 is a new susceptibility gene to autoimmune thyroid diseases: association and linkage study. *Journal of Clinical Endocrinology and Metabolism* **88**, 2274–2280.

Hui, J., Hung, L. H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S. A. & Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO Journal* **24**, 1988–1998.

Kandul, N. P. & Noor, M. A. (2009). Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC Genetics* (*Electronic resource*) **10**, 67.

Kawasaki, E. S. (1990). Sample preparation from blood, cells, and other fluids. In *PCR Protocols: a Guide to Methods and Applications* (ed. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J.), pp. 146–152. New York: Academic Press.

Lu, Q., Wallrath, L. L., Granok, H. & Elgin, S. C. (1993). (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. *Molecular and Cellular Biology* **13**, 2802–2814.

Peck, L. J. & Wang, J. C. (1985). Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. *Cell* **40**, 129–137.

Royaux, I. E., Suzuki, K., Mori, A., Katoh, R., Everett, L. A., Kohn, L. D. & Green, E. D. (2000). Pendrin, the protein encoded by the Pendred syndrome gene (SLC26A4), is an apical porter of iodide in the thyroid and is regulated by thyroglobulin in FRTL-5 cells. *Endocrinology* **141**, 839–845.

Scott, D. A., Wang, R., Kreman, T. M., Sheffield, V. C. & Karniski, L. P. (1999). The Pendred syndrome gene encodes a chloride–iodide transport protein. *Nature Genetics* **21**, 440–443.

Sham, P. C. & Curtis, D. (1995). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Annals of Human Genetics* **59**, 97–105.

Sharma, V. K., Kumar, N., Brahmachari, S. K. & Ramachandran, S. (2007). Abundance of dinucleotide repeats and gene expression are inversely correlated: a

role for gene function in addition to intron length. *Physiological Genomics* **31**, 96–103.

Stephens, M. & Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genet* **73**, 1162–1169.

Stephens, M. & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**, 449–462.

Stephens, M., Smith, N. J. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

Sueoka-Aragane, N., Imai, K., Komiya, K., Sato, A., Tomimasu, R., Hisatomi, T., Sakuragi, T., Mitsuoka, M., Hayashi, S., Nakachi, K. & Sueoka, E. (2008). Exon 19 of EGFR mutation in relation to the CA-repeat polymorphism in intron 1. *Cancer Science* **99**, 1180–1187.

Suzuki, M., Kageyama, S., Shinmura, K., Okudela, K., Bunai, T., Nagura, K., Igarashi, H., Kiyose, S., Mori, H., Tao, H., Goto, M., Takamochi, K., Mochizuki, T.,

Suzuki, K., Ohashi, R., Ogawa, H., Yamada, T., Niwa, H., Tsuneyoshi, T. & Sugimura, H. (2008). Inverse relationship between the length of the EGFR CA repeat polymorphism in lung carcinoma and protein expression of EGFR in the carcinoma. *Journal of Surgical Oncology* **98**, 457–461.

Wangemann, P., Nakaya, K., Wu, T., Maganti, R. J., Itza, E. M., Sanneman, J. D., Harbidge, D. G., Billings, S. & Marcus, D. C. (2007). Loss of cochlear HCO_3^- secretion causes deafness via endolymphatic acidification and inhibition of Ca^{2+} reabsorption in a Pendred syndrome mouse model. *American Journal of Physiology*. *Renal Physiology* **292**, F1345–F1353.

Wren, J. D., Forgacs, E., Fondon, J. W. 3rd, Pertsemlidis, A., Cheng, S. Y., Gallardo, T., Williams, R. S., Shohet, R. V., Minna, J. D. & Garner, H. R. (2000). Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *American Journal of Human Genetics* **67**, 345–356.

Zhou, H. & Lin, K. (2008). Excess of microRNAs in large and very 5′ biased introns. *Biochemical and Biophysical Research Communications* **368**, 709–715.