The Journal of International Medical Research
2010; 38: 1653 – 1662

ADP-ribosylation Factor Guanine Nucleotide-exchange Factor 2 (ARFGEF2): a New Potential Biomarker in Huntington's Disease

L Lovrecic[1], I Slavkov[2], S Džeroski[2] and B Peterlin[1]

[1]Institute of Medical Genetics, Division of Gynaecology, University Medical Centre Ljubljana, Ljubljana, Slovenia; [2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Microarray searches have revealed potential genetic biomarkers in a wide variety of human diseases. Identification of biomarkers for disease status is particularly important in chronic neurodegenerative diseases where brain tissue cannot be sampled. A previous study identified 12 genes from microarray analysis as associated with Huntington's disease, although the relationships had not been validated. We used new machine learning approaches to reanalyse those microarray data and to rank the identified potential genetic biomarkers. We then performed quantitative reverse transcription–polymerase chain reaction analysis on a subset of the candidate genes in blood samples from an independent cohort of 23 Huntington's disease patients and 23 healthy controls. Our highest ranked genes did not overlap with the 12 previously identified, but two were significantly up-regulated in the Huntington's disease group: *ARFGEF2* and *GOLGA8G*. Little is known about the latter, but the former warrants further analysis as it is known to be associated with intracellular vesicular trafficking, disturbances of which characterize Huntington's disease.

KEY WORDS: Huntington's disease; Trinucleotide repeat diseases; Gene expression; Biomarkers; Transcriptomics; ADP-ribosylation factors; *ARFGEF2*

Introduction

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder caused by an expanded CAG tract in the *huntingtin* gene. The prevalence of this disease varies across countries; three to 10 people per 100 000 are affected in most European countries.[1] The typical age of onset being within the third to fifth decades of life and the clinical characteristics of the disease include progressive motor impairment, cognitive decline and various psychiatric symptoms. The disease is generally fatal within 15 – 20 years of diagnosis owing to progressive neurodegeneration.[2]

So far, no effective treatment has been developed to cure HD or slow its progression, as neurons of the central nervous system

cannot regenerate after cell death or damage. Although the responsible gene and mutation were identified and characterized in 1993,[3] the mechanisms underlying neurodegeneration are still not clear. More than a decade of basic research has, however, demonstrated that multiple biochemical pathways are involved, including those affecting protein degradation, apoptosis, accumulation of misfolded mutated proteins, intracellular signalling, oxidative stress, mitochondrial involvement and transcription.[4]

Biomarkers are of extreme relevance in chronic neurodegenerative diseases such as HD, since it is not possible to sample brain tissue to monitor pathophysiological processes.[5] Tremendous efforts have been made to identify neuropathological, biochemical and genetic biomarkers of these diseases, with the intention of establishing diagnosis earlier in the disease course to enable surveillance of the rate of progression and response to treatment. The mutation that encodes CAG expansion is a definitive diagnostic marker for HD, but provides no information on the clinical progression of HD. Thus, the discovery of biomarkers that can identify disease status would be useful.

In mutation carriers the period of time without clinical symptoms is rather long; neuroprotective therapy given at the right time during this stage might delay progression of neurodegeneration, or even abolish it, if started early enough.[6] The currently used clinical rating scales, such as the Unified Huntington's Disease Rating Scale (UHDRS)[7] and Total Functional Capacity scale,[8] can be very useful for long-term assessment but are insensitive to progression over short periods of time. Moreover, the scales have only limited ability to distinguish effects on disease progression from symptomatic benefits. In addition, they do not take into account differences in clinical phenotypes (burden of motor, cognitive and psychiatric symptoms) between patients, which makes direct comparative assessment difficult. An advantage of the identification of biomarkers that reflect disease-related changes in pre-symptomatic patients might be to facilitate accurate evaluation of the efficacy of new therapies and improve the safety of clinical trials.

In a previous study, Borovecki *et al*.[9] identified numerous gene expression changes by microarray analyses of blood samples from pre-symptomatic and symptomatic HD patients compared with healthy controls. More than 300 genes were differentially expressed, of which 12 were chosen for further analysis. The data were made freely available in *Gene Expression Omnibus* (http://www.ncbi.nlm.nih.gov/geo). Another study that repeated analyses of various candidate genes, including those identified by Borovecki *et al.*, was not able to confirm the observed expression changes.[10]

Machine learning can be useful in the identification of biomarkers as systems are specifically designed to analyse high-throughput data, such as generated in microarray studies. These approaches are used to identify and rank genes potentially involved in the presence and/or status of given diseases with statistically robust accuracy. We reanalysed the freely available microarray expression data reported by Borovecki *et al*.[9] to identify potential biomarkers. While the analysis they performed was based on a simple *t*-test statistic, we have used more complex machine-learning approaches. We also attempted to validate a subset of the potential biomarkers found in an independent cohort of HD patients, with comparisons against controls.

Patients and methods

PATIENTS

Microarray data were obtained for all patients reported in the study by Borovecki *et al.*,[9] which are deposited in *Gene Expression Omnibus* (GEO Accession No. GDS1331). Data were reported for blood samples from 17 HD patients (five pre-symptomatic and 12 symptomatic) and from 14 healthy controls. For the validation study an independent group of HD patients and healthy controls were recruited, matched for age and sex, from the central database at the Institute of Medical Genetics in Ljubljana and from the Clinical Department of Neurology of the University Medical Centre Ljubljana. Inclusion criteria for HD patients were the presence of HD-specific gene mutations as revealed by molecular genetic testing at the Institute of Medical Genetics in Ljubljana. For healthy controls, exclusion criteria included the presence of any acute or chronic disease state, as well as blood disease, which could interfere with gene expression in blood. This research project was approved by the National Medical Ethics Committee of the Republic of Slovenia. All participants gave written informed consent to take part.

BLOOD COLLECTION AND RNA ISOLATION

Peripheral blood was drawn in PAXgene™ blood collection tubes (PreAnalytiX [Qiagen and Becton, Dickinson and Co.], Zurich, Switzerland). For RNA isolation PAXgene™ blood RNA kits (PreAnalytiX), free from ribonuclease but containing deoxyribonuclease, were used according to the manufacturer's protocol. The quality of total RNA was analysed with the RNA 6000 Nano LabChip kit on a 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA, USA).

MACHINE LEARNING ANALYSIS

The microarray expression data were analysed with the BioDCV system, version 2.3 (MPBA, Trento, Italy).[11] The system uses support vector machines[12] coupled with a recursive feature elimination algorithm[13] to assign importance values to individual genes.

The top-ranking genes can be used to build the predictive models with the lowest errors. To determine the genes most likely to be involved in HD progression, the highest-ranking gene was first entered into a predictive support vector machine model and the calculated prediction error was recorded. The second highest gene was then added into the model and a new prediction error recorded. The addition of genes in descending rank order was continued until all genes from the ranked list were included. From all the recorded prediction errors an average testing error curve was constructed. As the number of top-ranking genes introduced increased, the error value gradually decreased, but eventually reached a point at which it levelled off, which enabled a cut-off point for importance to be determined (Fig. 1).

The following specific BioDCV parameters were used: 100 replicates with fixed cross-validation as a resampling scheme; the predictive models built were support vector machines with linear kernels (E = 0.001 and complexity parameter C = 10); and the recursive feature elimination algorithm employed the so-called entropy-based recursive feature elimination.[14]

Top-ranking genes were selected for quantitative reverse transcription–polymerase chain reaction (RT–PCR) analysis according to their original expression values, as this method is not suitable for detecting small changes in gene expression.

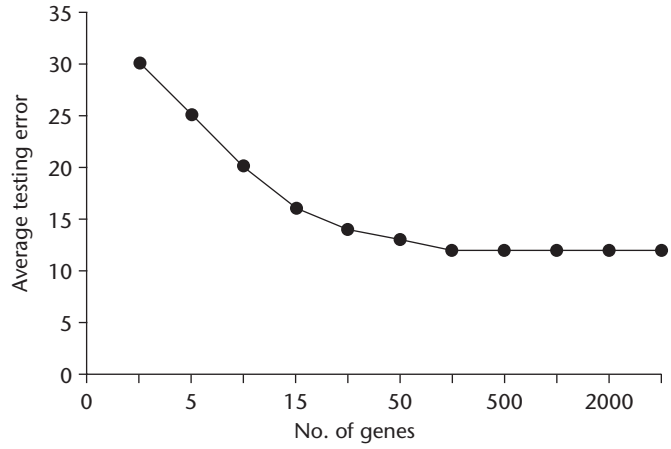


FIGURE 1: Typical example of an average testing error curve for genes entered into a support vector machines model. As genes are added to the model in descending order of importance, the error value gradually decreases, eventually levelling off and determining the cut-off point for the most important genes

QUANTITATIVE RT–PCR ANALYSIS

For RT 2 µg total RNA isolated from blood was processed with the SuperScript® first-strand synthesis system (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. The quantitative RT–PCR reactions were performed in the ABI PRISM® 7000 Sequence Detection System (Perkin-Elmer, Applied Biosystems, Norwalk, CT, USA). Primers for internal gene control and all reaction protocols remained the same as reported by Borovecki *et al.*[9]

Initial analysis was performed with ABI PRISM® system software (Perkin Elmer, Applied Biosystems). Relative gene expressions were calculated using the $2^{-\Delta\Delta Ct}$ method[15] with β-actin as an internal control. Primers were designed in the Primer3 program, version 0.4.0 (Whitehead Institute for Biomedical Research, Cambridge, MA, USA)[16] and the details of selected genes and primer sequences are shown in Table 1. Each sample was run in triplicate for each gene.

STATISTICAL ANALYSIS

For the list of genes identified as potentially interesting according to the BioDCV approach, their up- or down-regulation status was ascertained. Quantitative RT–PCR results were tested for significance using standard Student's *t*-test and calculation of the *P*-value.

Results

In total 23 symptomatic and pre-symptomatic HD patients (10 women, 13 men; mean ± SD age 48.3 ± 12.6 years) and 23 age- and gender-matched healthy controls (10 women, 13 men; mean ± SD age 43.0 ± 11.7 years) were selected for the validation study.

The 20 top-ranking genes according to the BioDCV system analysis were further investigated; this was the total considered as interesting as the average testing error (ATE) curve (Fig. 2) then levelled off. Of these top-ranked 20 genes, 14 were up-regulated in the HD group and six were down-regulated compared with those in the healthy controls (Fig. 3). Thus, a clear distinction was seen between the groups. None of the 12 genes that Borovecki *et al.*[9]

TABLE 1:
Description, GenBank details, probes, primer sequences and expression relative to healthy controls as measured by quantitative reverse transcriptase–polymerase chain reaction (RT–PCR) for the four genes with the largest up- or down-regulation from their original expression values detected by microarray analysis

Gene	Description	GenBank Accession No.	Affymetrix microarray probe	Primer sequence (5′ – 3′)	Expression relative to control	
					Fold change on quantitative RT–PCR	Statistical significance[a]
ACTN4	Actinin α4	81	200601_at	Forward: TCTGCTCCAGACTCACTTGC Reverse: TCTGCCAACTCAGCTCCTCT	1.245	$P = 0.056$
ARFGEF2	ADP-ribosylation factor guanine nucleotide-exchange factor 2	10564	218098_at	Forward: CAGCAGCTTTGCAGTTTGG Reverse: GAGAGCAAGGATTTCCAG	1.699	$P = 0.011$
GOLGA8G	Golgi autoantigen, golgin subfamily a, 8G	283768	213737_x_at	Forward: GCCAATTCCAGTCCAAAG Reverse: GGCCACTCTAGGGAAAATC	1.968	$P = 0.016$
PAPOLA	Poly(A) polymerase α	10914	212718_at	Forward: CAAGCTGGAAACTTGGACCT Reverse: CATGCGAAAGCAACAGTC	1.426	$P = 0.068$

[a]Huntington's disease patients versus controls.

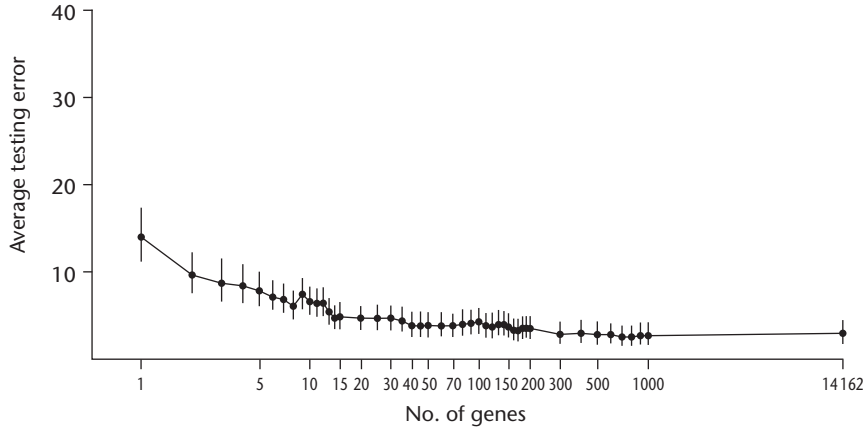


FIGURE 2: Average testing error curve for the support vector machines model designed to rank potential candidate genes in Huntington's disease identified by microarray analysis. The error value did not change notably after 20 genes had been added to the model and this value was, therefore, taken as the cut off for relevance

selected as potential biomarkers was included in the top 20 ranking; they were ranked between 51 and 2855 by the BioDCV system (Table 2).

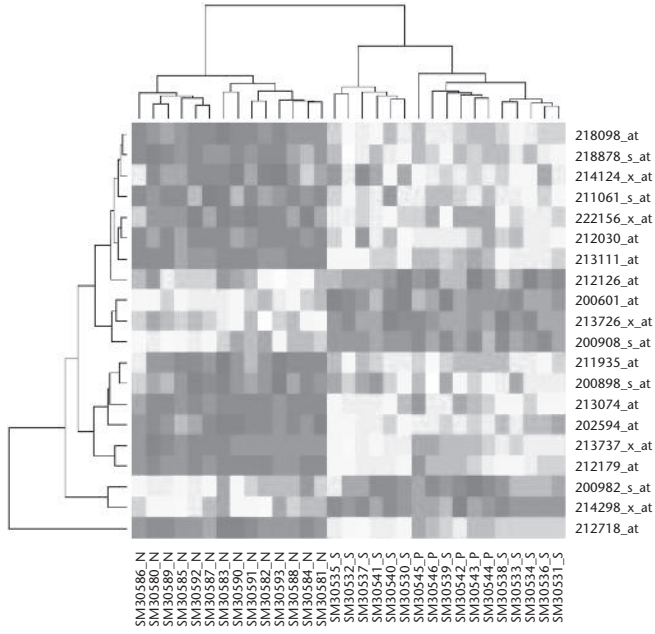
The four genes with the largest up- or down-regulation in HD patients compared to controls were selected for further analysis with quantitative RT–PCR: three of the genes (*ARFGEF2*, *GOLGA8G* and *PAPOLA*) were up-regulated and one (*ACTN4*) was down-regulated. Quantitative RT–PCR found that *ARFGEF2* and *GOLGA8G* were significantly up-regulated in HD patients compared with in controls (P = 0.011 and 0.016, respectively), whereas the difference between groups for up-regulation of *PAPOLA* did not reach statistical significance. We were unable to reproduce the down-regulation of *ACTN4* (Table 1).

Discussion

The present reanalysis of the freely available microarray gene expression data from Borovecki *et al.*[9] did not support the reported findings. A lack of overlap was also previously found by Runne *et al.*[10] when they tested the 12 genes identified by Borovecki *et al.*

Quantitative RT–PCR confirmed that *ARFGEF2* was significantly up-regulated in HD patients compared with controls. This gene encodes ADP-ribosylation factors, which play important roles in intracellular vesicular trafficking. Mutated huntingtin and its related proteins have been associated with disturbed vesicular transport and intracellular trafficking.[17,18] We believe, therefore, that *ARFGEF2* might reflect the pathological processes in HD. Although *GOLGA8G* was significantly up-regulated in HD patients, little is known about this gene and, therefore, no conclusions can be drawn on its functional role as an HD biomarker. Table 3 shows the gene ontology biological processes, molecular function and cellular component terms of the four genes that were selected for further analysis by quantitative RT–PCR in the present study.

Divergence in results is a common experience in microarray analysis, as diverse sets of potential biomarkers are generated



Probe set ID	Gene symbol	Regulation	Student's *t*-test
200601_at	*ACTN4*	Down	8.13463×10^{-10}
200898_s_at	*MGEA5*	Up	2.62652×10^{-5}
200908_s_at	*RPLP2*	Down	2.17052×10^{-5}
200982_s_at	*ANXA6*	Down	7.68081×10^{-5}
202594_at	*LEPROTL1*	Up	4.33639×10^{-10}
211061_s_at	*MGAT2*	Up	5.14006×10^{-8}
211935_at	*ARL6IP*	Up	1.20315×10^{-6}
212030_at	*RBM25*	Up	3.03373×10^{-9}
212126_at	*CBX5*	Down	0.000519726
212179_at	*SFRS18*	Up	2.27655×10^{-9}
212718_at	*PAPOLA*	Up	1.15206×10^{-10}
213074_at	*IRAK1BP1*	Up	1.43338×10^{-9}
213111_at	*PIP5K3*	Up	2.27591×10^{-10}
213726_x_at	*TUBB2C*	Down	8.10966×10^{-6}
213737_x_at	*GOLGA8G*	Up	3.7325×10^{-9}
214124_x_at	—	Up	0.000204931
214298_x_at	*SEPT6*	Down	1.87846×10^{-6}
218098_at	*ARFGEF2*	Up	1.05421×10^{-10}
218878_s_at	*SIRT1*	Up	1.00594×10^{-8}
222156_x_at	*CCPG1*	Up	5.08446×10^{-7}

FIGURE 3: Heat map and summary representation of expression of the 20 top-ranking genes for Huntington's disease as found on microarray analysis. Lighter colours indicate up-regulation and darker colours indicate down-regulation. Compared with controls, 14 genes were up-regulated and six were down-regulated in patients with Huntington's disease (data from 23 Huntington's disease patients and 23 healthy controls)

when different statistical and data mining approaches are employed. This may explain some of the difference between these results and those of Borovecki *et al*.[9] We were also unable to reproduce our initial finding of down-regulation of *ACTN4*, possibly due to the robustness of quantitative RT–PCR, which is not always able to reproduce gene expression results from microarray studies. It is, therefore, anticipated that multiple

TABLE 2:
Comparison of ranks for candidate biomarker genes previously identified by Borovecki *et al.*[9] with ranks found in the current study

Gene	Study rank	
	Borovecki *et al.*[9] [a]	Current study
ANXA1	30	1621
MARCH7 (*AXOT*)	13	1042
CAPZA1	15	153
HIF1A	34	1416
JJAZ1	19	58
P2Y5	54	2855
PCNP	21	1930
ROCK1	44	1490
SF3B1	12	123
SP3	28	2154
TAF7	35	1884
YPEL1 (*YIPPEE*)	42	51

[a]Gene expression results were first cross-referenced from both Affymetrix and Amersham platforms. Genes were then chosen according to their *P*-values, highest fold change, highest expression levels, and consistency of fold change in each individual Huntington's disease sample, compared with its age- and gender-matched control. Numbers reported here are gene ranks according to *P*-value on the Affymetrix platform.

independent studies on different HD biomarkers will be required optimally to evaluate putative biomarkers. In addition, standard operating procedures will have to be defined. Nevertheless, although studies of novel non-invasive approaches to assess gene expression in blood show encouraging results, we do not believe that gene expression microarray studies alone will be sufficient for complete biomarker identification for HD and other neurodegenerative diseases. Rather, a combined approach of using genomic, metabolomic and proteomic data will be required.

Many different statistical and machine learning methods are currently being used to analyse the vast data generated in high-throughput microarray studies depending on the aim of the study and the type of input data.[19,20] The use of support vector machines coupled with a recursive feature elimination algorithm is considered among the best methods for the ranking of features.[13] Data mining seems likely, therefore, to continue and to remain an essential and efficient part of study in all 'omic' disciplines.

In conclusion, the results of the present study did not overlap with those from previous studies and, therefore, we cannot confirm the roles of genes previously proposed as biomarkers for disease progression in HD. We have, however, identified at least one gene that warrants further analysis, namely *ARFGEF2* that is known to be associated with intracellular vesicular trafficking, disturbances of which characterize HD.

Acknowledgements

This research was supported by grant J3-7411 from the Ministry of Science and Technology, Ljubljana, Republic of Slovenia. We express special gratitude to the participants in this study.

Conflicts of interest

The authors had no conflicts of interest to declare in relation to this article.

TABLE 3:
The gene ontology (GO) biological processes, molecular function and cellular component terms of the four genes with the largest up- or down-regulation from their original expression values based on microarray analysis that were selected for further analysis by quantitative reverse transcription polymerase chain reaction

Gene	GO biological process term	GO molecular function term	GO cellular component term
ACTN4	Response to hypoxia	Nucleoside binding	Stress fibre
	Proteolysis	Actin binding	Intracellular
	Protein transport	Calcium binding	Nucleus
	Positive regulation of sodium hydrogen antiporter activity	Calcium-dependent cysteine-type endopeptidase activity	Nucleolus
	Regulation of apoptosis	Integrin binding	Cytoplasm
	Positive regulation of pinocytosis	Calcium ion binding	Ribonucleoprotein complex
	Actin filament bundle formation	Protein binding	Cortical cytoskeleton
	Negative/positive regulation of cell motion	Peptidase activity	Pseudopodium
		Hydrolase activity	Protein complex
		Protein complex binding	Perinuclear region of cytoplasm
		Protein homodimerization activity	
		Actin filament binding	
ARFGEF2	Exocytosis	Guanyl-nucleotide exchange factor activity	Golgi membrane
	Intracellular signalling cascade		Intracellular
	Regulation of ARF protein signal transduction	ARF-guanyl-nucleotide exchange factor activity	Cytoplasm
			Trans-Golgi network
		Myosin binding	Cytosol
		γ-Aminobutyric acid-receptor binding	Membrane
GOLGA8G	–	–	–
PAPOLA	Nuclear mRNA splicing, via spliceosome	Nucleotide binding	Nucleus
		RNA binding	Nucleolus
	Transcription	Polynucleotide adenylyltransferase activity	Cytoplasm
	mRNA polyadenylation		
	mRNA processing	ATP binding	
	RNA 3′-end processing	Transferase activity	
	RNA polyadenylation		

ARF, alternative reading frame.

• Received for publication 6 March 2010 • Accepted subject to revision 6 May 2010
• Revised accepted 28 September 2010

References

1 Vonsattel JP, DiFiglia M: Huntington disease. *J Neuropathol Exp Neurol* 1998; **57:** 369 – 384.
2 Imarisio S, Carmichael J, Korolchuk V: Huntington's disease: from pathology and genetics to potential therapies. *Biochem J* 2008; **412:** 191 – 209.
3 The Huntington's Disease Collaborative Research Group: A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 1993; **72:** 971 – 983.
4 Harjes P, Wanker EE: The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci* 2003; **28:** 425 – 433.
5 Rachakonda V, Pan TH, Le WD: Biomarkers of neurodegenerative disorders: how good are they? *Cell Res* 2004; **14:** 347 – 358.
6 Henley SM, Bates GP, Tabrizi SJ: Biomarkers for neurodegenerative diseases. *Curr Opin Neurol* 2005; **18:** 698 – 705.
7 Huntington Study Group: Unified Huntington's Disease Rating Scale: reliability and consistency. *Mov Disord* 1996; **11:** 136 – 142.
8 Shoulson I, Fahn S: Huntington disease: clinical care and evaluation. *Neurology* 1979; **29:** 1 – 3.
9 Borovecki F, Lovrecic L, Zhou J, *et al*: Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci USA* 2005; **102:** 11023 – 11028.
10 Runne H, Kuhn A, Wild EJ, *et al*: Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. *Proc Natl Acad Sci USA* 2007; **104:** 14424 – 14429.
11 Albanese D: *BioDCV: a Distributed Computing System for the Complete Validation of Gene Profiles*. Trento: University of Trento, 2005.
12 Vapnik V: *Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communication and Control Series)*. New York: John Wiley, 1998.
13 Guyon I, Weston J, Barnhill S, *et al*: Gene selection for cancer classification using support vector machines. *Mach Learn* 2002; **46:** 389 – 422.
14 Furlanello C, Serafini M, Merler S, *et al*: Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics* 2003; **4:** 54.
15 Livak KJ, Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* 2001; **25:** 402 – 408.
16 Rozen S, Skaletsky H: Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 2000; **132:** 365 – 386.
17 Truant R, Atwal R, Burtnik A: Hypothesis: huntingtin may function in membrane association and vesicular trafficking. *Biochem Cell Biol* 2006; **84:** 912 – 917.
18 Li XY, Li SH: HAP1 and intracellular trafficking. *Trends Pharmacol Sci* 2005; **26:** 1 – 3.
19 Lee JK, Williams PD, Cheon S: Data mining in genomics. *Clin Lab Med* 2008; **28:** 145 – 166.
20 Piroozni M, Yang JY, Yang MQ, *et al*: A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 2008; **9(suppl 1):** S13.

Author's address for correspondence:
Dr Luca Lovrecic
Institute of Medical Genetics, Division of Gynaecology, University Medical Centre Ljubjana, Šlajmerjeva 3, 1000 Ljubljana, Slovenia.
E-mail: lucalovrecic@gmail.com