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

L Lovrecic1, I Slavkov2, S Đzeroski2 and B Peterlin1

1Institute of Medical Genetics, Division of Gynaecology, University Medical Centre Ljubljana, Ljubljana, Slovenia; 2Department 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, 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.

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. 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. The currently used clinical rating scales, such as the Unified Huntington’s Disease Rating Scale (UHDRS) and Total Functional Capacity scale, 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. 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.

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. 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 deoxyribo-nuclease, 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 machines12 coupled with a recursive feature elimination algorithm13 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.
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–ΔΔCT method15 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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>GenBank Accession No.</th>
<th>Affymetrix microarray probe</th>
<th>Primer sequence (5′ – 3′)</th>
<th>Expression relative to control</th>
<th>Statistical significancea</th>
</tr>
</thead>
</table>
| ACTN4     | Actinin α4                                                                   | 81                    | 200601_at                   | Forward: TCTGCTCCAGACTCAGTTGC  
Reverse: TCTGCCAACTCAGCTCCTC                                                      | 1.245                      | P = 0.056                |
| ARFGEF2   | ADP-ribosylation factor guanine nucleotide-exchange factor 2                | 10564                 | 218098_at                   | Forward: CAGCAGCTTTGCAAGTTGG  
Reverse: GAGAGCAAGGATTTCCAG                                                      | 1.699                      | P = 0.011                |
| GOLGA8G   | Golgi autoantigen, golgin subfamily a, 8G                                   | 283768                | 213737_x_at                 | Forward: GCCATTTCAGTCAAAG  
Reverse: GGCCACTCTAGGGAAAATC                                                      | 1.968                      | P = 0.016                |
| PAPOLA    | Poly(A) polymerase α                                                        | 10914                 | 212718_at                   | Forward: CAAAGCTGAAAACCTGGACCT  
Reverse: CATGCGAAAGCAAAGTCA                                                      | 1.426                      | P = 0.068                |

aHuntington’s disease patients versus controls.

- **ACTN4**
  - GenBank Accession No.: 81
  - Affymetrix microarray probe: 200601_at
  - Primer sequence:
    - Forward: TCTGCTCCAGACTCAGTTGC
    - Reverse: TCTGCCAACTCAGCTCCTC
  - Expression relative to control: 1.245
  - Statistical significance: P = 0.056

- **ARFGEF2**
  - GenBank Accession No.: 10564
  - Affymetrix microarray probe: 218098_at
  - Primer sequence:
    - Forward: CAGCAGCTTTGCAAGTTGG
    - Reverse: GAGAGCAAGGATTTCCAG
  - Expression relative to control: 1.699
  - Statistical significance: P = 0.011

- **GOLGA8G**
  - GenBank Accession No.: 283768
  - Affymetrix microarray probe: 213737_x_at
  - Primer sequence:
    - Forward: GCCATTTCAGTCAAAG
    - Reverse: GGCCACTCTAGGGAAAATC
  - Expression relative to control: 1.968
  - Statistical significance: P = 0.016

- **PAPOLA**
  - GenBank Accession No.: 10914
  - Affymetrix microarray probe: 212718_at
  - Primer sequence:
    - Forward: CAAAGCTGAAAACCTGGACCT
    - Reverse: CATGCGAAAGCAAAGTCA
  - Expression relative to control: 1.426
  - Statistical significance: P = 0.068
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 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...
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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Borovecki et al.(^9) (^a)</th>
<th>Current study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANXA1</td>
<td>30</td>
<td>1621</td>
</tr>
<tr>
<td>MARCH7 (AXOT)</td>
<td>13</td>
<td>1042</td>
</tr>
<tr>
<td>CAPZA1</td>
<td>15</td>
<td>153</td>
</tr>
<tr>
<td>HIF1A</td>
<td>34</td>
<td>1416</td>
</tr>
<tr>
<td>JJAZ1</td>
<td>19</td>
<td>58</td>
</tr>
<tr>
<td>P2YS</td>
<td>54</td>
<td>2855</td>
</tr>
<tr>
<td>PCNP</td>
<td>21</td>
<td>1930</td>
</tr>
<tr>
<td>ROCK1</td>
<td>44</td>
<td>1490</td>
</tr>
<tr>
<td>SF3B1</td>
<td>12</td>
<td>123</td>
</tr>
<tr>
<td>SP3</td>
<td>28</td>
<td>2154</td>
</tr>
<tr>
<td>TAF7</td>
<td>35</td>
<td>1884</td>
</tr>
<tr>
<td>YPEL1 (YIPPEE)</td>
<td>42</td>
<td>51</td>
</tr>
</tbody>
</table>

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

ARF, alternative reading frame.
Potential biomarkers for Huntington’s disease progression

References

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