CAG repeat length in RA11 is associated with age at onset variability in spinocerebellar ataxia type 2 (SCA2)

Sean Hayes1,2, Gustavo Turecki2,3, Katéri Brisebois2, Iscia Lopes-Cendes4, Claudia Gaspar2,5, Olaf Riess5, Laura P.W. Ranum7, Stefan-M. Pulst8 and Guy A. Rouleau2,+

1Department of Biology, 2Centre for Research in Neurosciences and 3Douglas Hospital Research Institute, McGill University, Montreal, Quebec H3G 1A4, Canada, 4Departamento de Genética Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, SP 13081-970, Brazil, 5Laboratorio Genética Médica, ICBAS and UnlGENe, IBMC Iniversidade do Porto, Porto, Portugal, 6Department of Molecular Human Genetics, Ruhr University, Bochum D-44780, Germany, 7Departments of Neurology and Genetics, Institute of Human Genetics, University of Minnesota, Minneapolis, MN 55455, USA and 8Division of Neurology and Rose Moss Laboratory for Parkinson’s and Neurodegenerative Disorders, Burns and Allen Research Institute, University of California, Los Angeles, CA 90048, USA

Received 11 February 2000; Revised and Accepted 9 May 2000

Spinocerebellar ataxia type 2 (SCA2) is an autosomal dominant disorder caused by the expansion of a polymorphic (CAG)n tract, which is translated into an expanded polyglutamine tract in the ataxin-2 protein. Although repeat length and age at disease onset are inversely related, ∼50% of the age at onset variance in SCA2 remains unexplained. Other familial factors have been proposed to account for at least part of this remaining variance in the polyglutamine disorders. The ability of polyglutamine tracts to interact with each other, as well as the presence of intranuclear inclusions in other polyglutamine disorders, led us to hypothesize that other CAG-containing proteins may interact with expanded ataxin-2 and affect the rate of protein accumulation, and thus influence age at onset. To test this hypothesis, we used step-wise multiple linear regression to examine 10 CAG-containing genes for possible influences on SCA2 age at onset. One locus, RA11, contributed an additional 4.1% of the variance in SCA2 age at onset after accounting for the effect of the SCA2 expanded repeat. This locus was further studied in SCA3/Machado–Joseph disease (MJD), but did not have an effect on SCA3/MJD age at onset. This result implicates RA11 as a possible contributor to SCA2 neurodegeneration and raises the possibility that other CAG-containing proteins may play a role in the pathogenesis of other polyglutamine disorders.

INTRODUCTION

The autosomal dominant cerebellar ataxias (ADCAs) are a heterogeneous group of neurodegenerative disorders typified by gait, speech and limb incoordination. The diseases are progressive in nature and generally feature degeneration of the cerebellum, brainstem and spinocerebellar tracts (1). To date, 10 spinocerebellar ataxia (SCA) loci have been implicated in ADCAs (SCA1–SCA8, SCA10 and SCA11) (2–5), while six genes have been cloned [SCA1, SCA2, SCA3/Machado–Joseph disease (MJD), SCA6, SCA7 and SCA8] (6–13) and shown to contain polymorphic (CAG/CTG)n repeats. With the exception of SCA8, these repeats are translated into polyglutamine tracts which appear to be toxic when expanded into a pathological range. For this group of ‘polyglutamine disorders’, which also includes Huntington’s disease (HD) (14), dentatorubral-pallidoluysian atrophy (DRPLA) (15,16) and spinobulbar muscular atrophy (17), a negative correlation exists between the size of the repeat tract and the age at disease onset (18), with variations in repeat size generally accounting for ∼25–80% of the variability in the age at onset (19–21), depending on the disease being examined. Several studies have suggested that the variability in age at onset not attributable to repeat size is in part due to the existence of modifying genetic factors (19,22–24). How these modifiers affect the disease process is not known, but it is an attractive possibility that these factors may interact with the disease proteins and, by virtue of their distribution, help to explain why specific subsets of neurons appear vulnerable in each disease despite widespread expression of the disease proteins themselves (25). In fact, models have been proposed where polyglutamine-containing disease proteins interact with cell-specific factors via non-covalent ‘polar-zippers’ (26) or via transglutaminase-mediated covalent cross-linking (27), producing complexes which perhaps play a role in the characteristic patterns of cell death seen in each disease.

Recently, insoluble inclusions, most frequently nuclear, have been reported in both affected patient brain material as well as transfected cells for many of the polyglutamine disorders (25,28). Although some evidence exists indicating...
that inclusions may not be necessary for initiation of SCA1 and HD pathogenesis (29,30), the existence of expanded polyglutamine tracts in these structures, as well as the finding that inclusions are commonly found in affected brain regions, suggest that inclusions are, at some level, disease-related structures (31). Much attention is now being focused on determining the protein make up of inclusions and initial results suggest that some proteins may enter these complexes in a disease-specific manner. For instance, a protein expressed predominantly in cerebellar Purkinje cells, the leucine-rich acidic nuclear protein, co-localizes to inclusions in SCA1 (32) but is not found in inclusions in SCA7 (33). Identification of the full protein complement of inclusions is important as it is possible that the spectrum of interacting proteins found in each disorder may influence the process of cell death, whereas rates at which these proteins enter aggregates may impact on the rate of neurodegeneration.

Although the list of possible interacting candidate proteins is indeed large, the ability of polyglutamine tracts to interact with one another makes proteins with even non-pathological polyglutamine stretches excellent candidates for recruitment into nuclear aggregates. In addition, the polymorphic nature of these repeats could potentially impact on the rate of aggregate formation. This in turn may influence age at disease onset, which makes polymorphic polyglutamine-coding genes strong candidates for roles as genetic modifiers of age at onset in the SCAs. Support for this hypothesis can be seen in the recent demonstration by Perez et al. (31) that ataxin-3 (the disease protein in SCA3/MJD) can interact with both ataxin-1 (the disease protein in SCA1) and the polyglutamine-containing TATA-binding protein (TBP) in nuclear aggregates. There is also evidence that an expanded repeat-containing protein can recruit the same protein with a non-expanded repeat into aggregates (34,35). However, the pathological consequences of these interactions, or their relationship to disease progression, remains unclear.

SCA2 is a typical polyglutamine disorder, with expansions of 34–59 repeats associated with disease (25). Recently, evidence of nuclear inclusions has been reported in SCA2 (36), suggesting that protein aggregation and sequestration may also play a role in this disease. Although expanded repeat sizes are relatively small in SCA2, expanded ataxin-2 still has the potential to form aggregates which interact with other polyglutamine-containing proteins, with potential pathological consequences. Therefore, as a way to begin addressing the possible role of polyglutamine-containing proteins as genetic modifiers in the SCAs, we have investigated several CAG-containing genes thought to be expressed in the brain as potential modifiers of the age at disease onset in SCA2.

### RESULTS

Table 1 lists the size ranges and the numbers of alleles found for each of the genes tested. In our initial screen of 46 SCA2 patients, regression with the small allele of each locus tested or the total of the two alleles did not yield significant results (data not shown). When the large alleles were considered, the size of the SCA2 expanded repeat accounted for 48% of the variability seen in the age at onset (r = 0.693, r² = 0.480, P < 0.0001). Of the remaining loci that were tested, only retinoic acid-induced 1 (RAI1) produced a significant improvement of fit (r = 0.737, r² = 0.543, P = 0.021). RAI1 is the human homolog of a mouse gene known by the same name that codes

### Table 1. PCR annealing temperatures, primers, allelic results and heterozygosity obtained for tested CAG-containing loci

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primer pair a</th>
<th>Map position</th>
<th>Annealing temperature (°C)</th>
<th>No. of alleles</th>
<th>Range of alleles b</th>
<th>Heterozygosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCA1</td>
<td>Rep1 and Rep2 (6)</td>
<td>6p23</td>
<td>54</td>
<td>9</td>
<td>24–33</td>
<td>0.73</td>
</tr>
<tr>
<td>SCA3/MJD</td>
<td>MJD52 and MJD25a (37)</td>
<td>14q21</td>
<td>60</td>
<td>13</td>
<td>15–41</td>
<td>0.66</td>
</tr>
<tr>
<td>SCA6</td>
<td>S-5-F1 (7) and SCA6-D (5’-CCGGGCCCCCGTGTCG-3’)</td>
<td>19p13.1</td>
<td>68</td>
<td>8</td>
<td>4–15</td>
<td>0.66</td>
</tr>
<tr>
<td>SCA7</td>
<td>7ALT and 4U716 (38)</td>
<td>3p21.1–3p12</td>
<td>56</td>
<td>5</td>
<td>10–35</td>
<td>0.52</td>
</tr>
<tr>
<td>DRPLA</td>
<td>B37 primer pair (39)</td>
<td>12p</td>
<td>62</td>
<td>16</td>
<td>11–27</td>
<td>0.78</td>
</tr>
<tr>
<td>KCNN3</td>
<td>SCZ41 (5’-GGGCTGTTGGGACTTTGATAP3-3’)</td>
<td>1q21</td>
<td>61</td>
<td>12</td>
<td>24–35</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>SCZ42 (5’-GGGCAGAAGAGCAAGGATGC-3’)</td>
<td>6p23</td>
<td>54</td>
<td>9</td>
<td>24–33</td>
<td>0.73</td>
</tr>
<tr>
<td>HD</td>
<td>HD2 (5’-TACCGCGACCTGGGAAAAGCTGAA3-3’)</td>
<td>14q21</td>
<td>60</td>
<td>13</td>
<td>15–41</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>HD1 (5’-GGGACCTGAGAAACGCTGAGG-3’)</td>
<td>14q21</td>
<td>56</td>
<td>4</td>
<td>14–38</td>
<td>0.85</td>
</tr>
<tr>
<td>TBP</td>
<td>TBP1 (5’-CTGTCTATTTTGGAGAAAGCAAAAGG-3’)</td>
<td>4p16.3</td>
<td>59</td>
<td>17</td>
<td>14–38</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>TBP2 (5’-CTGCTGGAGACCTGAGCTGAA-3’)</td>
<td>4p16.3</td>
<td>59</td>
<td>17</td>
<td>14–38</td>
<td>0.85</td>
</tr>
<tr>
<td>RAI1</td>
<td>SCZ15 (5’-GGGGCAGCGGTTCCAGAATCTTC-3’)</td>
<td>6q27</td>
<td>62</td>
<td>11</td>
<td>27–40</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SCZ16 (5’-CTGGCGTCTGACTGCATGGTACT-3’)</td>
<td>6q27</td>
<td>62</td>
<td>11</td>
<td>27–40</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SCZ91 (5’-TGCTCAGAAAACAGCGAGTTTGGTCTG-3’)</td>
<td>17p11.2</td>
<td>62</td>
<td>5</td>
<td>10–18</td>
<td>0.45</td>
</tr>
<tr>
<td>TNRC22</td>
<td>SCZ92 (5’-CCACCGCATTGGGGCCAG-3’)</td>
<td>16q</td>
<td>62</td>
<td>6</td>
<td>13–28</td>
<td>0.65</td>
</tr>
</tbody>
</table>

a Primer references shown in parentheses when available.
b Minimum and maximum sizes of CAG tracts.
for a protein of unknown function that is expressed during retinoic acid-induced differentiation of mouse embryonic carcinoma cells into neurons (40). Figure 1 shows scatter plots of the residuals (observed minus values predicted by linear regression) of age at onset and numbers of repeats for the SCA2 and RAI1 loci, which clearly illustrate the effect of these loci on age at onset. A locus (KCNN3) without any effect is also shown for the sake of comparison. A log transformation of age at onset slightly increased the fit of the regression equation for the SCA2 repeat alone ($r^2 = 0.516, P < 0.0001$), as well as for RAI1 ($r^2 = 0.576, P = 0.019$), and allowed DRPLA to enter the equation. After accounting for the effect of the SCA2 repeat and RAI1, the large allele of the DRPLA repeat improved the fit to $r^2 = 0.615 (P = 0.050)$. At no time did overall CAG load, as represented by summing both alleles of all loci or the large allele of all loci, have an effect on age at onset.

In order to address the possibility that the presence of multiple affected individuals from single nuclear families may have been producing false positive results in the initial screen, we decided to perform a second screen using a single SCA2-positive individual per nuclear family. This new panel of 26 samples was used to re-test SCA2, RAI1 and DRPLA. Using the large alleles of each of these loci in step-wise multiple regressions versus age at onset produced an increase in the fit only for RAI1 ($r^2 = 0.572, P = 0.003$, up from $r^2 = 0.362, P = 0.001$ for the SCA2 repeat alone).

To further investigate this possible association, we obtained an additional and independent sample of 47 SCA2 subjects to produce a combined panel of 73 unrelated SCA2 patients. This panel was used to test RAI1 in a final screen. Once again, the large allele of the RAI1 locus significantly increased the fit from $r^2 = 0.447 (P < 0.0001)$ for the SCA2 repeat alone to $r^2 = 0.488 (P = 0.021)$ for the two loci together. This represents an increase of 4.1% in the total explained SCA2 age at onset variability. The effect of RAI1 on age at onset seems to be linear, as shown by the ANOVA ($F = 27.39, df = 2, 41, P < 0.001$) testing the null hypothesis that none of the independent variables are predictors of the age at onset when used in the form expressed by the regression equation (linear).

Finally, the possible role that this locus may play in modifying age at onset in other SCA3 was examined by assembling a panel of 68 unrelated SCA3/MJD patients and testing this panel for the effect of the SCA3/MJD repeat, as well as RAI1. Whereas the SCA3/MJD repeat accounted for 37.3% of age at onset variability ($r = -0.610, P < 0.0001$), RAI1 did not produce a significant effect after step-wise multiple linear regression (data not shown).

**DISCUSSION**

We have identified the polymorphic (CAG)$_n$ repeat within the RAI1 gene as a potential modifier of age at disease onset for SCA2. In our SCA2 patients, the large allele of this locus accounts for an additional 4.1% of the age at onset variability after accounting for the effect of the SCA2 expanded repeat.

Interestingly, the magnitude of the effect attributable to the repeat in RAI1 is similar to that predicted by Ranum et al. (22), who found that 5% of age at onset variation in SCA1 may be due to unidentified genetic factors. A similar percentage of the total variation in age at onset in HD (4.1%) was reported by

---

**Figure 1.** Scatter plots of residuals (observed minus values predicted by linear regression) of the age at onset and of the CAG repeat length obtained in the original data set for the following loci: (a) SCA2, (b) RAI1 and (c) KCNN3. P-values refer to significance of the improvement of fit to the linear regression. KCNN3 is included as an example of a locus without effect on age at onset. The residuals are not standardized and units are years (age at onset) and number of repeats (locus). See text for additional information.
Rubinsztein et al. (23) to be associated with a polymorphic TAA repeat in the GluR6 kainate receptor locus. Although the identification of potential modifiers for the polyglutamine disorders is a significant finding, it should be noted that the modest size of these effects still makes predictive testing in these disorders impossible, as too much age at onset variation remains unexplained. As a result, the role of loci such as those reported here may be more important for the clues that they can provide about the pathways and mechanisms that may be involved in the pathogenesis of the polyglutamine disorders.

As noted by Joober et al. (41), RAI1 is the human homolog of the mouse Rail gene (previously known as GT1), which codes for a 196 kDa protein of unknown function which is expressed during retinoic acid-induced differentiation of mouse embryonic carcinoma cells into neurons (40). Imai et al. (40) showed that the protein is highly expressed in neurons of the mouse brain. This neuronal expression reinforces the possibility that RAI1 might be involved in a neurodegenerative disease. If human RAI1 is also expressed in neurons, it may be available for interaction with ataxin-2. In our study, larger RAI1 alleles were associated with earlier ages at onset (after accounting for SCA2 repeat size). Although it is impossible to establish the true nature of this association at this time, it is tempting to speculate that RAI1 protein might interact with ataxin-2 more readily as the RAI1 polyglutamine tract increases in size (thus the greatest interaction occurs with the larger gene product). This may in turn influence the rate of protein aggregation and therefore impact on disease progression, as reflected by the age at onset association. Further attempts to identify discernible protein aggregates in SCA2, as well as their constituents, will be necessary to test the validity of this model. If protein aggregation is occurring in SCA2, it would be interesting to see whether the RAI1 gene product, or other polyglutamine-containing proteins, are involved in their formation. Such interactions may simply influence the rate of aggregate formation and growth or serve to sequester important cellular components and disrupt cellular pathways. Many groups are presently attempting to address these questions.

Recently, Perez et al. (31) reported that ataxin-3 can interact with both ataxin-1 and TBP in cellular aggregates. It is interesting to note that we found no association between the TBP CAG repeat and age at onset for SCA2 patients and that similar testing with SCA3/MJD patients did not produce significant results (data not shown). This is not to say that TBP is not involved in aggregate formation in these diseases, but simply that it does not appear to influence age at onset in our panels. In addition to those noted by Perez et al. (31), ataxin-3 may interact with other polyglutamine-containing proteins, including the RAI1 gene product, although we did not find RAI1 to be associated with variation in SCA3/MJD disease onset. This is perhaps not surprising considering that it is possible that each polyglutamine disorder will have a unique spectrum of interacting proteins as well as genetic modifiers. Nevertheless, the findings of Perez et al. (31) further support the possibility that ‘normal length’ polyglutamine-containing proteins may be involved in the pathogenesis of the polyglutamine disorders. To this end, we had hypothesized that overall ‘CAG load’, i.e. the total of all the polyglutamine proteins found in the affected neurons, might have an influence on the disease process. Although the data we obtained for the total of the 10 loci tested did not support this idea, many variables existed which were not controllable. For example, the cells which are vulnerable in SCA2 potentially only express a subset of the 10 loci tested, while at the same time expressing other polyglutamine-containing proteins which were not tested. Expression levels and compartmentalization of the various proteins will also influence cellular polyglutamine levels. As a result, our ‘CAG load’ is most likely not an accurate reflection of the actual levels found within the affected cells. Nevertheless, it appears that CAG tract-containing genes, such as RAI1, may play a role in modifying the age at disease onset for at least some of the SCAs.

It should be noted that the precise magnitude of the effect that any locus has on age at onset in the polyglutamine disorders is subject to the accuracy of several variables. A certain degree of error can be expected when patients or their families are asked to identify an accurate age at disease onset. Likewise, exact sizing of polymorphic CAG repeats can be difficult, especially when somatic mosaicism produces several bands, as is seen in the expanded alleles of most SCAs. Indeed, the source tissue for DNA extraction (lymphocytes, lymphoblastoid cell lines and brain material) used in the analysis may also affect the repeat size results obtained (42). As a result, replication of results such as those obtained in this study will be important before the absolute validity of any associations can be established.

As Rubinsztein et al. (23) pointed out, the boundaries between monogenic and multifactorial diseases have become blurred somewhat in recent years by the identification of genetic modifiers. Diseases such as the polyglutamine disorders appear to be more genetically complex than originally imagined. The discovery of even modest modifier effects are important as they may serve to increase our overall understanding of the pathological mechanisms of these diseases. Hopefully, this increased understanding will ultimately lead to our ability to effectively treat, or even cure, these devastating disorders.

**MATERIALS AND METHODS**

**Patients**

A panel consisting of 46 individuals taken from 10 SCA2 pedigrees was assembled for an initial screen with potential modifier loci. Age at disease onset ranged from 16 to 60 years of age, with a mean of 34.6 years. The age at onset frequency was consistent with a normal distribution (skewness = 0.405, SE = 0.343; kurtosis = -0.358, SE = 0.674). Genomic DNA was extracted from peripheral lymphocytes or lymphoblastoid cell lines using standard protocols (43). SCA2 repeat size was determined as reported elsewhere (44).

A second panel was assembled which included 26 individuals chosen at random from SCA2 pedigrees such that a single individual was tested per pedigree (in the case of small pedigrees) or per nuclear family (in the case of large pedigrees). This panel was used to re-test any locus which yielded suggestive results in the initial screen in order to address any biases that may have been introduced by testing multiple affected individuals from single families. Any loci which still produced significant results were also tested in 47 additional unrelated SCA2 patients obtained through collaboration, yielding a total of 73 unrelated SCA2 patients for this final screen. For this
panel, age at onset ranged from 13 to 62 years of age, with a mean of 31.9 years.

Finally, a panel of 68 unrelated SCA3/MJD patients was assembled and used to test any locus which produced suggestive results with the SCA2 panels. Individuals in this panel had age at disease onset ranging from 10 to 60 years of age, with a mean of 35.9 years.

Molecular analysis

A total of 10 genes known to contain polymorphic (CAG) repeats in their coding sequences were examined as potential modifiers of SCA2 age at onset. These included six genes known to be involved in neurological disease (SCA1, SCA3/ MJD, SCA6, SCA7, DRPLA and HD), as well as four genes not unquestionably shown to be responsible for a disease phenotype: TBP, calcium-activated potassium channel 3 (KCNN3; GenBank accession no. U80757). These latter loci were selected from a panel of 50 CAG repeat-containing loci previously identified through database searches and used in an unrelated study (41) based on: (i) their high heterozygosity and/or long repeat length; and (ii) their probable expression in the brain. Repeats were amplified by PCR using conditions reported elsewhere (44) and the primer pairs and annealing temperatures listed in Table 1. SCZC41 and SCZC42 amplify both CAG tracts in KCNN3, which were added together to give total CAG length for this locus. Dimethylsulfoxide (10%) was included in all reactions except for SCA3/MJD, which used reaction conditions reported elsewhere (37). PCR products were run on 6% denaturing polyacrylamide gels and sized by comparison with an M13 sequencing ladder. For each locus, the number of CAG repeats in both alleles was recorded.

Statistical analysis

Statistical analysis was carried out using step-wise multiple linear regression with age as the dependent variable (Systat v.5.05; SPSS, Chicago, IL). This analysis was considered suitable to study the potential effect of the investigated loci as age at onset modifiers because it controls for the SCA2ted suitable to study the potential effect of the investigated loci as age at onset modifiers because it controls for the SCA2

REFERENCES


