Neural correlates of gene-environment interactions in ADHD
van der Meer, Dennis

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 13-12-2019
Chapter 7

Predicting ADHD severity from psychosocial stress and stress response genes
a random forest regression approach

Abstract

Introduction: Difficulty identifying genetic variants contributing to attention-deficit/hyperactivity disorder (ADHD) may be due to the involvement of numerous common polymorphisms with small effects, interacting with each other as well as with environmental factors. ADHD has been inconsistently linked to genetic variation influencing hypothalamic-pituitary-adrenal (HPA) axis activity in response to stress. Random forest regression is well suited to explore this complexity, as it allows for the analysis of many predictors simultaneously, taking into account any higher-order interactions between them.

Methods: We predicted ADHD severity, measured by the Conners Parent rating scales, from 686 adolescents and young adults (including 281 diagnosed with ADHD) using random forest regression. The analysis included 17,374 single nucleotide polymorphisms (SNPs) across 29 genes previously linked to HPA axis activity, together with information on exposure to 24 individual long-term difficulties or stressful life events.

Results: The model explained 12.5 percent of variance in ADHD severity. The most important SNP for prediction, which also showed the strongest interaction with stress, was located in a region regulating expression of telomerase reverse transcriptase (TERT). Other high-ranking SNPs predicting ADHD severity were found in or near NPSR1, ESR1, GABRA6, PER3, NR3C2, and DRD4. Chronic stressors were more influential than single, severe, life events.

Conclusions: Random forest regression may provide an overview of how many genetic and environmental factors come together to jointly contribute to ADHD. It is able to detect novel SNPs of interest, interacting with stress exposure, and may explain prior inconsistent findings in ADHD genetics by identifying SNPs that are only relevant for specific subsets of individuals. This exploratory approach may be best combined with more focused research; top predictors, and their interactions with one another, should be replicated in independent samples.
Stress response genes and random forests

**Introduction**

Attention-deficit/hyperactivity disorder (ADHD) is thought to result primarily from numerous common genetic and environmental factors with small effects (Faraone et al., 2015). The relation of any individual risk factor with ADHD depends on other genetic polymorphisms and/or environmental factors that dampen or amplify its effect on the underlying neurobiological pathways. Failing to take interaction effects into account will lead to noisier estimates, which may contribute to the lack of consistent findings from studies on genetic variation associated with ADHD.

Individuals vary widely in their response to stressful stimuli, which can be partly attributed to differences in regulation of the hypothalamic pituitary adrenal (HPA) axis (Kudielka et al., 2009). Brain regions involved in perceiving threat, such as the prefrontal cortex, hippocampus, and amygdala, may stimulate HPA axis activity through the hypothalamus (McEwen et al., 2015). This results in the release of a range of neurotransmitters, peptides, and hormones such as cortisol that stimulate the sympathetic nervous system. The strength and duration of the stress response is determined by an intricate system of feedforward and feedback loops (Joels & Baram, 2009). HPA axis regulation is moderated by previous experiences, with stressors being particularly impactful during periods of heightened brain development, such as in early childhood (Loman & Gunnar, 2010).

ADHD has been associated with altered cortisol levels, albeit with much heterogeneity between reports. While a meta-analysis has indicated that individuals with ADHD have a blunted cortisol response to acute stressors (Scassellati et al., 2012), higher cortisol levels, both at baseline and in response to stress, have also been reported (Corominas et al., 2012). These findings may possibly relate to interindividual differences in ADHD symptom presentation, comorbidity, and duration and extent of exposure to chronic stress (Christiansen et al., 2010, Corominas et al., 2012, Freitag et al., 2009). Further suggestive evidence of HPA axis involvement in ADHD comes from findings that stimulant medication normalizes patients’ cortisol levels (Kariyawasam et al., 2002), and from the role of the HPA axis in regulation of emotion (Adam, 2012), sleep (Van Lenten & Doane, 2016), and circadian rhythm (Baird et al., 2012), which are often altered in ADHD (Cortese et al., 2013, Shaw et al., 2014).

Genetic determinants of HPA axis activity may contribute to the diversity of findings on the relation between ADHD and the stress response. ADHD has been associated with polymorphisms in the glucocorticoid and mineralocorticoid receptor genes (NR3C1 and NR3C2; Fortier et al., 2013), which provide negative feedback to the HPA axis when activated by cortisol (Buckingham, 2006). We have found that NR3C1 interacts with psychosocial stress on ADHD severity, and that this gene-environment interaction (GxE) is further moderated by the serotonin transporter gene 5-HTT Van der Meer et al., 2016). Serotonin signalling is tightly coupled to regulation of HPA axis activity (Leonard, 2005), and 5-HTT is one of several serotonergic genes that have been repeatedly linked to ADHD (Gizer et al., 2009, Oades et al., 2008). The most prominent candidate genes for ADHD, the dopamine transporter (DAT1) and dopamine receptor D4 (DRD4) are also known to influence
Chapter 7

the effect of stressors on HPA axis activity (Alexander et al., 2011, Buchmann et al., 2014). Besides reports of GxE, stress response genes have also been found to moderate each other’s effects on the HPA axis (Armbruster et al., 2009, Clasen et al., 2011), illustrating the complexity of the genetic architecture underlying the stress response pathway.

While conventional regression analyses have led to various interesting findings on ADHD genetics, they are limited in their ability to handle many predictors and interaction terms simultaneously. This undermines accurate estimation of the true effect of a risk factor on ADHD, as its contributions through interactions with other factors gets neglected.

Random forest regression (RFR) is well-suited for investigating the aetiology of complex traits using high-dimensional data (Chen & Ishwaran, 2012). It allows for inclusion of many more predictors than there are observations, and automatically incorporates all higher-order interactions between the predictors in its estimates (Breiman, 2001). RFR has been praised for its robustness and predictive accuracy, particularly for noisy data containing many factors with small effects (Fernández-Delgado et al., 2014, Scornet et al., 2015). Studies simulating complex genetic datasets have shown that it outperforms other techniques when it comes to detecting interacting single nucleotide polymorphisms (SNPs) with small marginal effects (Lunetta et al., 2004).

RFR is a non-parametric ensemble learning method, aggregating the results from many individual decision trees. Overfitting is prevented by growing each tree using a bootstrap sample, and by selecting from a random subset of variables at each split (Breiman, 2001). Observations not included in a tree’s sample due to the bootstrapping procedure, called out-of-bag (OOB; on average about 36%), serve as the tree’s test set and are used to measure prediction error. Importance of a predictor of interest can be estimated through permutation, by randomly shuffling its values in the OOB samples and comparing the resulting prediction error to the error obtained before the shuffle (Ishwaran, 2007). The so-called variable importance estimate VIMP derived in this way includes all interaction effects, as permuting a predictor will remove any influence it had on the selection of other variables further down the tree.

In this study, we build on the strengths of random forest regression by predicting ADHD severity from SNPs in genes previously implicated to influence HPA axis activity, together with long-term stress exposure. Machine-learning techniques, including tree-based techniques, have been used to predict ADHD diagnosis as accurately as possible, using neuropsychological and brain imaging data (Brown et al., 2012, Eloyan et al., 2012, Fair et al., 2012, Sato et al., 2012). Our aim was not to optimize prediction, but rather to gain a good sense of the complex genetic architecture of ADHD and how this interacts with the environment. We therefore estimated the contributions of thousands of HPA axis related SNPs plus exposure to stressors simultaneously, identifying those that are of particular interest for follow-up research. The analyses were carried out in a sample of adolescents and young adults (mean age 17.2 years) consisting of individuals with ADHD and healthy controls, as well as individuals with some symptoms of ADHD but not enough to
Stress response genes and random forests

meet the diagnostic criteria, referred to as ‘subthreshold’. This sample composition enabled analysis within a wide range of ADHD severity, in accordance with the contribution of genetic and environmental variation to the continuous distribution of ADHD traits in the general population (Larsson et al., 2012).
Methods

Participants were selected from the NeuroIMAGE study, a follow-up of the Dutch part of the International Multicentre ADHD Genetics (IMAGE) study (Von Rhein et al., 2014). NeuroIMAGE included 365 families with at least one child with ADHD and at least one biological sibling (regardless of ADHD diagnosis) and 148 control families with at least one child, without any formal or suspected ADHD diagnosis in any of the first-degree family members. ADHD families were recruited through ADHD outpatient clinics in the regions Amsterdam, Groningen, and Nijmegen (the Netherlands). Control families were recruited through primary and high schools in the same geographical regions. To be included in NeuroIMAGE, participants had to be of European Caucasian descent, between ages 5 and 30, have an IQ ≥ 70, and no diagnosis of autism, epilepsy, general learning difficulties, brain disorders, or known genetic disorders. The study was approved by the regional ethics committee (CMO Regio Arnhem – Nijmegen; 2008/163; ABR: NL23894.091.08) and the medical ethical committee of the VU University Medical Centre. All participants, and their parents (if the participant was younger than 18 years), signed informed consent; parents signed informed consent for participants under twelve years of age.

For the analyses reported in this paper, 686 participants, from 360 families, had complete data. Of these, 281 participants had an ADHD diagnosis, 88 participants had subthreshold ADHD (i.e., had elevated levels of ADHD symptoms without meeting the full criteria for an ADHD diagnosis), and 292 participants were healthy controls. ADHD diagnoses were made in accordance with DSM IV-TR criteria, on the basis of a combination of a semi-structured interview and the Conners Rating Scales (Conners et al., 1998). Participants were asked to withhold use of their stimulant medication or other psychoactive drugs for 48 hours before measurement. Mean age of this sample was 17.1 years (standard deviation (SD) 3.4) and 52.3% were males. More information on the NeuroIMAGE study, its diagnostic algorithm, and its participants, is presented in the Supplementary Information (SI) and in Von Rhein et al. (2014).

ADHD outcome measure

In order to retain as much information on ADHD as possible, we used a continuous measure of ADHD severity, the raw score on subscale N of the Conners Parent Rating Scale (CPRS; Conners et al., 1998). This score ranged from 0 to 53, with an average of 13.1 (SD 12.1). This measure was chosen because it was available for all participants, from both ADHD families and control families.

Stress exposure

Two questionnaires were used to assess exposure to psychosocial stress. Parents filled in the Long-Term Difficulties (LTD) questionnaire (Zandstra et al., 2015) containing thirteen items measuring whether their children have been exposed to chronic stressors such as a handicap, being bullied, having financial difficulties, or other persisting problems at home or school. They were asked to only report chronic, ongoing difficulties. Participants themselves filled in the Stressful Live Events (SLE) questionnaire (Bosch et al., 2012, Oldehinkel et al., 2008), containing eleven items.
on exposure to specific major stressful events in the past five years, such as death or serious illness of a loved one, physical or sexual abuse, or failure at something important to them. See the SI for the full list.

Genetics

We selected genes suspected to be involved in regulating HPA axis activity, as indicated by reports from studies into genetic moderators of stress exposure in humans. This was done through a literature search in PubMed with the following search term: ("Gene-Environment Interaction"[Mesh] OR ("Genes"[Mesh] OR "Polymorphism, Genetic"[Mesh] OR gene* OR polymorphism* OR SNP*) AND ("Stress, Psychological"[Mesh]) OR adversit* OR maltreatment OR psychosocial OR neglect OR abuse)) AND ("Hypothalamic Hormones"[Mesh] OR HPA OR hypothalamic pituitary adrenal OR cortisol OR ACTH). After filtering for English-language articles with full text available, this search generated 415 results, of which 95 were relevant original research articles using human samples investigating specific genetic polymorphisms. See Table S1 in the SI for references to the articles on each gene. Together, these studies investigated 31 unique genes. Two of these genes, MAOA and HTR2C, were excluded because they are located on the X-chromosome. All SNPs within 100 kilo base pairs (kb) of the location of the remaining 29 genes (Veyrieras et al., 2008) as found in human assembly GRCh37 were included in the study, for a total of 17374. Table 1 lists details on these genes. We used LocusZoom (http://locuszoom.sph.umich.edu) to make plots of the linkage disequilibrium (LD) and recombination rate of regions that contained one of the SNPs among the top results, which are presented in the SI.

For the IMAGE sample, DNA was extracted from blood samples or immortalized cell lines at Rutgers University Cell and DNA Repository, New Jersey, USA (Brookes et al., 2006). DNA isolation for additional samples from the NeuroIMAGE study was performed at the department of Human Genetics of the Radboud University Medical Centre in Nijmegen (Von Rhein et al., 2014).

Genome-wide genotyping was performed using the Infinitum PsychArray-24 v1.1 BeadChip, containing 265000 tag SNPs, 245000 exome markers, and 50000 additional markers associated with common psychiatric disorders (http://www.illumina.com/products/psycharray.html). Genotypes were called using Illumina GenomeStudio software, excluding samples with a call rate below 0.995. Clustering was done using GeneTrain 2.0 (no-call threshold 0.15), excluding samples with call rate below 0.98. Quality control prior to imputation included removal of SNPs with a call rate below 98% or call rate differences between cases and controls higher than 2%, removal of SNPs with a minor allele frequency of less than 1% or failing the Hardy-Weinberg equilibrium test at a threshold of $p \leq 10^{-6}$, and removal of individuals with a call rate below 98% or heterozygosity rate of more than 3 standard deviations from the mean. For imputation we used IMPUTE2 (Howie et al., 2009) with 1000 Genomes Phase 1 V3 reference data (Abecasis et al., 2010). Hard calls were made by converting to PLINK format (Purcell et al., 2007) using GCTA software (Yang et al., 2011). SNPs with low imputation quality (INFO < 0.8) were filtered out.
Table 1. List of genes based on our literature search. A total of 17374 single nucleotide polymorphisms (SNPs) spread out over these 29 genes were included in the analysis. Next to each gene is displayed its protein product, the chromosome (Chr.) it is located on, the start and end position (in base pairs, bp) of the region we included, and the number of SNPs in that region.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein product</th>
<th>Chr</th>
<th>Start bp</th>
<th>End bp</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>Angiotensin converting enzyme</td>
<td>17</td>
<td>61454422</td>
<td>61675741</td>
<td>181</td>
</tr>
<tr>
<td>ADRA2B</td>
<td>Alpha2B adrenergic receptor</td>
<td>2</td>
<td>96678623</td>
<td>96881888</td>
<td>144</td>
</tr>
<tr>
<td>APOE</td>
<td>Apolipoprotein E</td>
<td>19</td>
<td>45309039</td>
<td>45512650</td>
<td>481</td>
</tr>
<tr>
<td>AVPR1A</td>
<td>Arginine vasopressin receptor 1A</td>
<td>12</td>
<td>63436539</td>
<td>63646590</td>
<td>655</td>
</tr>
<tr>
<td>AVPR1B</td>
<td>Arginine vasopressin receptor 1B</td>
<td>1</td>
<td>206124283</td>
<td>206331482</td>
<td>164</td>
</tr>
<tr>
<td>BDNF</td>
<td>Brain-derived neurotrophic factor</td>
<td>11</td>
<td>27576442</td>
<td>27822600</td>
<td>368</td>
</tr>
<tr>
<td>CHRNA7</td>
<td>Alpha7 nicotinic acetylcholine receptor</td>
<td>15</td>
<td>32222686</td>
<td>32562384</td>
<td>394</td>
</tr>
<tr>
<td>COMT</td>
<td>Catechol-O-methyltransferase</td>
<td>22</td>
<td>19829263</td>
<td>20057498</td>
<td>795</td>
</tr>
<tr>
<td>CRHBP</td>
<td>Corticotropin-releasing hormone binding protein</td>
<td>5</td>
<td>76148680</td>
<td>76365299</td>
<td>388</td>
</tr>
<tr>
<td>CRHR1</td>
<td>Corticotropin-releasing hormone receptor 1</td>
<td>17</td>
<td>43761646</td>
<td>44013194</td>
<td>229</td>
</tr>
<tr>
<td>CRHR2</td>
<td>Corticotropin releasing hormone receptor 2</td>
<td>7</td>
<td>30591559</td>
<td>30839719</td>
<td>575</td>
</tr>
<tr>
<td>DRD4</td>
<td>Dopamine receptor D4</td>
<td>11</td>
<td>537305</td>
<td>740705</td>
<td>615</td>
</tr>
<tr>
<td>ESR1</td>
<td>Estrogen receptor alpha</td>
<td>6</td>
<td>152028454</td>
<td>152524408</td>
<td>1323</td>
</tr>
<tr>
<td>FKBP5</td>
<td>FK506 binding protein 5</td>
<td>6</td>
<td>35441362</td>
<td>35756719</td>
<td>820</td>
</tr>
<tr>
<td>GABRA6</td>
<td>Gamma-aminobutyric acid A receptor alpha 6</td>
<td>5</td>
<td>161012658</td>
<td>16122959</td>
<td>519</td>
</tr>
<tr>
<td>HTR1A</td>
<td>Serotonin receptor 1A</td>
<td>5</td>
<td>63155875</td>
<td>63358119</td>
<td>264</td>
</tr>
<tr>
<td>MC2R</td>
<td>Melanocortin 2 Receptor</td>
<td>18</td>
<td>13782043</td>
<td>14015535</td>
<td>905</td>
</tr>
<tr>
<td>NPSR1</td>
<td>Neuropeptide S receptor</td>
<td>7</td>
<td>34597897</td>
<td>35017944</td>
<td>1313</td>
</tr>
<tr>
<td>NPY</td>
<td>Neuropeptide Y</td>
<td>7</td>
<td>24223807</td>
<td>24431484</td>
<td>905</td>
</tr>
<tr>
<td>NR3C1</td>
<td>Glucocorticoid receptor</td>
<td>5</td>
<td>142557496</td>
<td>142884045</td>
<td>481</td>
</tr>
<tr>
<td>NR3C2</td>
<td>Mineralocorticoid receptor</td>
<td>4</td>
<td>148899915</td>
<td>149463672</td>
<td>1147</td>
</tr>
<tr>
<td>OPRK1</td>
<td>Kappa opioid receptor</td>
<td>8</td>
<td>54038276</td>
<td>54264194</td>
<td>904</td>
</tr>
<tr>
<td>OPRM1</td>
<td>Mu opioid receptor</td>
<td>6</td>
<td>154260443</td>
<td>154540594</td>
<td>778</td>
</tr>
<tr>
<td>OXTR</td>
<td>Oxytocin receptor</td>
<td>3</td>
<td>8692095</td>
<td>8911300</td>
<td>538</td>
</tr>
<tr>
<td>PER1</td>
<td>Period circadian protein homolog 1</td>
<td>17</td>
<td>7943788</td>
<td>8155753</td>
<td>552</td>
</tr>
<tr>
<td>PER3</td>
<td>Period circadian protein homolog 3</td>
<td>1</td>
<td>7744714</td>
<td>8005237</td>
<td>861</td>
</tr>
<tr>
<td>SLC6A3</td>
<td>Dopamine transporter</td>
<td>5</td>
<td>1292905</td>
<td>1545543</td>
<td>442</td>
</tr>
<tr>
<td>SLC6A4</td>
<td>Serotonin transporter</td>
<td>17</td>
<td>28421337</td>
<td>28662986</td>
<td>305</td>
</tr>
<tr>
<td>STMN1</td>
<td>Stathmin</td>
<td>1</td>
<td>26110677</td>
<td>26332993</td>
<td>328</td>
</tr>
</tbody>
</table>
Random forest regression analysis

All analyses were run in R v3.2.3 (R Core Team, 2015), making use of the package randomForestSRC v2.2.0 (Ishwaran et al., 2008). The 17374 SNPs were coded to reflect the participants’ number of minor alleles (‘0’, ‘1’, or ‘2’), entered as non-ordered factors to allow for all possible genetic models. The 24 stress items were coded as ‘0’ (absence) or ‘1’ (presence), and also entered in the analysis as individual predictors, rather than stress sum score. This approach ensured that all information was maintained, i.e. the marginal and interaction effects of each individual stressor. It also prevented the potential bias of RFR whereby continuous measures, or categorical ones with many levels, are more often selected than categorical factors with few levels (Strobl et al., 2007).

We grew 5000 trees fully and used the default value of $p/3$ for \textit{mtry}, the size of the random subset of available predictors at each split, in this case 5800 (17398/3 rounded up). These settings were chosen to identify important predictors while still allowing for the detection of true predictors with small effects and interactions. This is in accordance with recommendations from simulation studies on complex genetic data with interacting SNPS (Winham et al., 2012). We further checked the stability of the results by rerunning the analysis twice, with different random seeds.

As a measure of importance, we report the Breiman-Cutler permutation variable importance, referred to as VIMP. VIMP is calculated by permuting the variable of interest in each tree’s OOB sample; the resulting increase in prediction error, averaged over all trees, is expressed as percent increase in mean-squared error (MSE; Breiman, 2001, Liaw & Wiener, 2002). Further, the increase in prediction error following simultaneous permutation of two variables minus the sum of their individual VIMPs may be used as a measure of interaction. The operating definition of interaction in this context is that a split on either of the predictors influences the likelihood of a subsequent split by the other predictor. A negative numeral indicates either predictor makes it more likely that the other is selected in its subtree, and a positive numeral indicates it make this less likely, as described by Ishwaran (2007). The VIMP interaction measures reported in the results section were obtained through the ‘find.interaction’ function included in the randomForestSRC toolbox. We made use of the ‘corrplot’ package for visualization of these results for the most important predictors, shown in Figure 2. The interaction estimates are multiplied by 100 for ease of display.

Figure S2 in the SI shows the Spearman’s rank correlation coefficient between each pair of the 25 highest-ranked predictors.
Results

The model explained 12.5 percent variance in ADHD severity. Permuting all SNPs simultaneously led to an 8.3 percent increase in MSE. For all stress items together, this was 25.3 percent. The 25 most important individual predictors are listed in Table 2, containing 20 SNPs and five stress items from the Long-Term Difficulties questionnaire. Figure 1 visualizes the variable importance of every SNP individually, grouped by gene. Figure 2 displays the estimated strength of interaction between each of the top predictors. Figure 3, for illustrative purposes, depicts the interaction of the highest-ranked SNP, rs4635969, with each of the five highest-ranked stress items.

Table 2. Top 25 most important predictors, based on the increase in prediction error following permutation. The five stressors are listed first, followed by details on the 20 single nucleotide polymorphisms (SNPs).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Stressor</th>
<th>Frequency</th>
<th>VIMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Your child has a chronic illness or handicap</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>2</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>3</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.07</td>
<td>3.61</td>
</tr>
<tr>
<td>4</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.08</td>
<td>1.24</td>
</tr>
<tr>
<td>5</td>
<td>Your immediate family has financial difficulties</td>
<td>0.04</td>
<td>0.35</td>
</tr>
<tr>
<td>6</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>7</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>8</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.07</td>
<td>3.61</td>
</tr>
<tr>
<td>9</td>
<td>Your immediate family has financial difficulties</td>
<td>0.08</td>
<td>1.24</td>
</tr>
<tr>
<td>10</td>
<td>Your immediate family has financial difficulties</td>
<td>0.04</td>
<td>0.35</td>
</tr>
<tr>
<td>11</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>12</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>13</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.07</td>
<td>3.61</td>
</tr>
<tr>
<td>14</td>
<td>Your immediate family has financial difficulties</td>
<td>0.08</td>
<td>1.24</td>
</tr>
<tr>
<td>15</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>16</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>17</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.07</td>
<td>3.61</td>
</tr>
<tr>
<td>18</td>
<td>Your immediate family has financial difficulties</td>
<td>0.08</td>
<td>1.24</td>
</tr>
<tr>
<td>19</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>20</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>21</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.07</td>
<td>3.61</td>
</tr>
<tr>
<td>22</td>
<td>Your immediate family has financial difficulties</td>
<td>0.08</td>
<td>1.24</td>
</tr>
<tr>
<td>23</td>
<td>Your child has fewer friends than he/she would like</td>
<td>0.23</td>
<td>15.01</td>
</tr>
<tr>
<td>24</td>
<td>Your child is being bullied at school or in the neighbourhood</td>
<td>0.15</td>
<td>4.64</td>
</tr>
<tr>
<td>25</td>
<td>Your child cannot get along with someone in your immediate family</td>
<td>0.07</td>
<td>3.61</td>
</tr>
</tbody>
</table>

Note: For the upper part of the table, the ‘Frequency’ column indicates the proportion of individuals that have experienced the stressor. For the lower part of the table, it displays the SNP’s minor allele frequency, the ‘Location’ column represents its genomic location by chromosome and base pair count, and the ‘Region’ column denotes the SNP’s position relative to its associated gene as documented in Table 1.

VIMP = Breiman-Cutler variable importance estimate, RS ID= reference SNP identification number, kb = kilo base pairs. UTR = untranslated region.
Figure 1. Variable importance for prediction, for all single nucleotide polymorphisms (SNPs) included in the analysis. SNPs are ordered on the x-axis based on their genomic position, from chromosome 1 to 22, with the labels and alternating red and black sections marking the gene they belong to. The y-axis indicates the variable importance, as percent increase in mean-squared error of the out-of-bag predictions when the SNP was permuted. Those above the dashed blue line are part of the top 25 most important predictors, listed in Table 2.
Stress response genes and random forests

Figure 2. Interaction strengths for each pair of 25 predictors found to be important in the random forest analysis. These were calculated by subtracting the sum of the pair’s individual importance estimates from their joint importance estimate. Negative numerals indicate that one predictor made it more likely that the other was selected for a split in its subtree, positive numerals indicate this was less likely. The predictors are sorted based on the first principal component of their interaction strengths.
Figure 3. Visualization of the interaction between rs4635969 and each of the five long-term difficulties among the top predictors. Participants are grouped based on their genotype and exposure to the individual long-term difficulty shown on the x-axis. On the y-axis is the observed score on the Conners Parent Rating Scale (CPRS), subscale N. The boxes show the median, and first and third quartiles of each group. Their width is scaled by the number of participants.
Discussion

In this study, we estimated the importance of stress-related genes, in interaction with stress exposure, for predicting ADHD severity through random forest regression. The strengths of this method, namely handling high-dimensional data and taking into account all possible interactions, align well with the complexity of stress response genetics. We reasoned this would enable us to identify important contributors, and obtain a good overall sense of how a multitude of SNPs from genes involved in HPA axis activity, combined with stress exposure, relates to ADHD.

The SNP with the highest estimated importance for predicting ADHD severity in our analysis, rs4635969, also showed the strongest interaction with a stressor. Multiple genome-wide association studies (GWAS), together with a meta-analysis, have provided strong cumulative evidence that this SNP is associated with risk for several forms of cancer (Mocellin et al., 2012). While we included rs4635969 as part of the 3' end-flanking region of SLC6A3, it is possible that these findings are explained by its close proximity to micro-RNA (MIR4457) at the 5' end of the telomerase reverse transcriptase (TERT) gene, known to regulate telomere length (Diaz De Leon et al., 2010). Overexpression of TERT increases cell proliferation and resilience to oxidative stress (Armstrong et al., 2005), whereas glucocorticoid administration and chronic stress exposure have been shown to lower basal telomerase activity and shorten telomere length (Epel et al., 2004, Epel et al., 2010). Therefore, while the C-allele of rs4635969 is linked to cancer, individuals carrying the T-allele may be more vulnerable to stress exposure through inhibition of telomerase activity by the HPA axis. Our pattern of results, together with reports on children’s telomere length being related to early social deprivation (Drury et al., 2012) and hyperactivity/impulsivity (De Souza Costa et al., 2015), suggest this SNP is of interest for GxE research.

The other high-ranked SNPs were in or near NPSR1, ESR1, GABRA6, PER3, DRD4, NR3C2, and OPRK1. Besides their shared association with HPA axis activity (references listed in Table S1), polymorphisms in these genes have all been repeatedly, but inconsistently, associated with internalizing and externalizing behaviour often co-occurring with ADHD (Comings et al., 2000, Domschke & Maron, 2013, Hess et al., 2016, Jüngling et al., 2008, Laas et al., 2014, Mill et al., 2008, Smoller, 2016, Sundermann et al., 2010). This mirrors the heterogeneous pattern of findings from studies into the relation of cortisol with ADHD as well as those on internalizing and externalizing behaviour, which have indicated that low reactivity of the HPA axis is most prominent in individuals with co-occurring externalizing disorders while high HPA axis activity relates more to anxiety and depression (Corominas et al., 2012, Marsman et al., 2008). If early splits in a tree form groups more homogeneous with regard to, for instance, externalizing behaviour, they allow for detection of a SNP that impacts ADHD severity only in these individuals and not in, e.g., more internalizing individuals. These differential effects, analogous to interactions, would increase error in straightforward association studies while they get incorporated in the importance estimates produced by RFR. Gene-gene interactions may explain a considerable amount of the heritability of ADHD that remains unaccounted for (Hemani et al., 2013, Zuk et al.,
Stress response genes and random forests

2012). An interesting feature of RFR to note in this context is its applicability to group individuals based on the proximity of their end nodes, or leafs, with others (Touw et al., 2012), reflecting the importance of predictors they have in common. It may be used to study inter-individual etiological heterogeneity of ADHD by distinguishing between groups, even if they have the same ADHD severity scores.

We further found that particularly long-term difficulties, compared to stressful live events, are important for predicting ADHD severity. A stronger influence of chronic stress can be explained by the principles of the allostatic load model and its implications for psychiatric disorders (McEwen, 1998). Allostatic load refers to the detrimental consequences of repeated stress, mediated by the long-term effects of stress hormones such as cortisol. Prolonged exposure to glucocorticoids is known to be particularly damaging to the prefrontal cortex and hippocampus, likely contributing to the relation of stress with a range of psychiatric disorders (Liston et al., 2009, McEwen et al., 2015). High allostatic load is thought to result from impaired feedback to the HPA axis leading to an extended stress response, and/or from low reactivity of one component inducing hyperactivity of other components of the stress response system (McEwen, 2004). Interactions between stressors, or between stressors and genetic variants, may therefore relate to how they strengthen each other’s effects on this system, leading to dysregulation and increased allostatic load. Neuroimaging data may be used to study the relation of polymorphisms, stressors, and their interactions with brain structure and activity, providing clues on how they influence the stress system, why they interact, and what their role is in ADHD (Van der Meer et al., 2015).

We included many predictors in this analysis that are correlated with each other. Whether correlation between predictors, such as SNPs in the same region, or exposure to different concurrent stressors, helps or hinders random forests depends on the aim of the study (Nicodemus et al., 2010). Individual importance estimates of correlated predictors will be lowered because a split on one will reduce the likelihood of the other subsequently being selected and vice versa. This also influences the measures of interaction, which are calculated by subtracting the sum of the individual importance estimates from their joint importance estimate; as correlation will make it more likely that the two predictors are part of different (sub)trees, the interaction measure may become less negative or even become positive (Ishwaran, 2007). Correlation between predictors may, however, be beneficial for analysis of the type of high-dimensional data encountered in genetics; while it may lower the estimated importance of the SNP best tagging the true locus of effect, the estimates of nearby SNPs in LD will be raised and therefore may aid in its identification. This pattern is clearly visible in Figure 1 as the streak of dots below a top hit. This inflation of importance estimates for predictors surrounding the true effect does not take place under the null hypothesis of no association with the outcome (Nicodemus et al., 2010), and therefore signals the authenticity of this effect. Further, correlated SNPs will increase the odds that interacting SNPs from another region are included in the same tree, thereby increasing the ability of the forest to incorporate the impact of interactions (Winham et al., 2012). This is particularly relevant for the small effects encountered in genetics, as this lowers the number of trees that contain both SNPs and contribute to the calculation of their interaction strength. Therefore, while
correlation may lower the quantitative measure of importance for the strongest predictor, it strengthens the confidence in the findings and more accurately captures the impact of groups of predictors.

The random forest approach should be seen as complementary to the conventional statistical techniques used in ADHD etiological studies. It has great potential as an exploratory tool, given its ability to handle high-dimensional data, and to produce measures of importance. However, while improvements are continuously being proposed and incorporated since its introduction in 2001 (Breiman, 2001), debates on its applicability and true meaning of the importance measures are still ongoing (Boulesteix et al., 2012, Winham et al., 2012). The interpretability of its results has also been criticized; whereas the findings from linear regression can be relatively easily probed, e.g., by plotting their coefficients, the importance of a predictor as estimated by random forests contains its complex interaction structure with all other predictors included. Simulation studies have further shown that random forests capture interaction effects, yet that they may fail to differentiate these from marginal effects, particularly when marginal effects of the predictors are small or absent (Wright et al., 2016). This may explain the lack of noteworthy gene-gene interactions shown in Figure 2, while interactions between the strongest predictors do get identified. Here, we took a three-step approach beginning with the distribution of all individual importance estimates, followed by extracting measures of interaction between the top predictors, and subsequently visualizing the most interesting GxE. Inference on such a selection however should take place in independent samples.

To summarize, in this study we made use of the strengths of random forest regression, an ensemble learning method that may aid in the exploration of high-dimensional data for associations with ADHD. Besides illustrating how many factors with small effects come together to predict ADHD, this method enables detection of risk factors that may get overlooked due to interaction effects and that contribute to the many differences between individuals with ADHD. We identified a novel relation between ADHD and a SNP that may relate to the expression of TERT, implying an influence on telomere length in relation to stress sensitivity. The importance of other SNPs among the top predictors may reflect the ability of random forests to capture effects of polymorphisms that are relevant for only a specific subset of individuals. Such interaction effects may contribute to inconsistent association of stress response genes with ADHD, or related disorders and traits. Our results also illustrated the strong effects of chronic stress, not found for individual stressful events, in accordance with allostatic load (McEwen, 1998). It should be noted that this was a cross-sectional study, which precludes any inference on the nature of the relation between the SNPs, the stressors, and ADHD, and may include gene-environment correlations. For instance, a polymorphism may both influence the odds of experiencing a stressor such as having a chronic illness or handicap and contribute to ADHD severity, though this would not necessarily make it any less of an interesting target for further research. This explorative study may best be followed up by selecting the strongest predictors, analysing whether the effects of this selection replicate in independent samples, and investigating how and why these are dependent on each other.
Stress response genes and random forests

References


predicting aggression and cortisol levels in young adults. Psychopharmacology (Berl), 231, 3089-3097.


Christiansen, H., Oades, R.D., Psychogiou, L., Hauffa, B.P. & Sonuga-Barke, E.J. (2010) Does the cortisol response to stress mediate the link between expressed emotion and oppositional behavior in Attention-Deficit/Hyperactivity-Disorder (ADHD)? Behavioral and Brain Functions, 6, 1.


Stress response genes and random forests


Stress response genes and random forests


