OPEN

Citation: Transl Psychiatry (2017) 7, e1145; doi:10.1038/tp.2017.114

www.nature.com/tp

ORIGINAL ARTICLE

Predicting attention-deficit/hyperactivity disorder severity from psychosocial stress and stress-response genes: a random forest regression approach

D van der Meer[1,2], PJ Hoekstra[1], M van Donkelaar[3], J Bralten[3], J Oosterlaan[4], D Heslenfeld[4], SV Faraone[5,6,7], B Franke[3], JK Buitelaar[8,9] and CA Hartman[1]

Identifying genetic variants contributing to attention-deficit/hyperactivity disorder (ADHD) is complicated by the involvement of numerous common genetic variants with small effects, interacting with each other as well as with environmental factors, such as stress exposure. 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 among them. Using random forest regression, we predicted ADHD severity, measured by Conners' Parent Rating Scales, from 686 adolescents and young adults (of which 281 were diagnosed with ADHD). The analysis included 17 374 single-nucleotide polymorphisms (SNPs) across 29 genes previously linked to hypothalamic–pituitary–adrenal (HPA) axis activity, together with information on exposure to 24 individual long-term difficulties or stressful life events. The model explained 12.5% of variance in ADHD severity. The most important SNP, which also showed the strongest interaction with stress exposure, was located in a region regulating the expression of telomerase reverse transcriptase (*TERT*). Other high-ranking SNPs were found in or near *NPSR1, ESR1*, *GABRA6, PER3*, *NR3C2* and *DRD4*. Chronic stressors were more influential than single, severe, life events. Top hits were partly shared with conduct problems. We conclude that random forest regression may be used to investigate how multiple genetic and environmental factors jointly contribute to ADHD. It is able to implicate novel SNPs of interest, interacting with stress exposure, and may explain inconsistent findings in ADHD genetics. This exploratory approach may be best combined with more hypothesis-driven research; top predictors and their interactions with one another should be replicated in independent samples.

Translational Psychiatry (2017) **7,** e1145; doi:10.1038/tp.2017.114; published online 6 June 2017

INTRODUCTION

Attention-deficit/hyperactivity disorder (ADHD) results in the majority of cases from numerous common genetic and environmental factors with mostly small effects.[1] The association of any individual risk factor with ADHD will depend on other genetic polymorphisms and/or environmental factors that dampen or amplify its effect on the underlying neurobiological pathways. Their joint effect therefore shapes the clinical profile of an individual, such as number of symptoms of inattention and hyperactivity/impulsivity displayed and their persistence over time. Failing to take such interaction effects into account will lead to noisier estimates of the effect of individual polymorphisms, which may have contributed to the inconsistent findings from studies investigating the genetics of ADHD.

We and others have shown that stress exposure has a role in ADHD.[2,3] 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.[4] Brain regions involved in perceiving threat, such as the prefrontal cortex, hippocampus and amygdala may stimulate HPA axis activity through the hypothalamus.[5] 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.[6] HPA axis regulation is moderated by previous experiences, with stress exposure being particularly impactful during periods of heightened brain development, such as in adolescence.[7]

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,[8] higher cortisol levels, both at baseline and in response to stress, have also been reported repeatedly.[9] These findings may possibly be linked to the duration and extent of exposure to chronic stress.[10] They may also relate to differences in ADHD symptom presentation and comorbidity;

[1]Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [2]K.G. Jebsen Centre for Psychosis Research/Norwegian Centre for Mental Disorder Research (NORMENT), Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [3]Department of Human Genetics and Psychiatry, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands; [4]Department of Clinical Neuropsychology, VU University Amsterdam, Amsterdam, The Netherlands; [5]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA; [6]Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, NY, USA; [7]K.G. Jebsen Centre for Psychiatric Disorders, Department of Biomedicine, University of Bergen, Bergen, Norway; [8]Karakter Child and Adolescent Psychiatry University Centre, Nijmegen, The Netherlands and [9]Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands. Correspondence: Dr D van der Meer, Department of Psychiatry, University of Groningen, University Medical Center Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands.
E-mail: dvdm09@gmail.com
Received 28 November 2016; revised 24 April 2017; accepted 28 April 2017

particularly the heightened levels of conduct problems seen in individuals with ADHD has been coupled to low levels of cortisol, which has been suggested to be causally related to this behavior by reflecting underarousal.[9–11] Further indirect indication of HPA axis involvement in ADHD comes from findings that stimulant medication normalizes patients' cortisol levels,[12] and from the role of the HPA axis in the regulation of emotion,[13] sleep[14] and circadian rhythms,[15] which are often altered in ADHD.[16,17]

Genetic determinants of HPA axis activity may contribute to the diversity of findings on the relationship between ADHD and the stress response. ADHD has been associated with polymorphisms in the glucocorticoid and mineralocorticoid receptor genes *NR3C1* and *NR3C2*,[18] which provide negative feedback to the HPA axis when activated by cortisol.[19] We have found that *NR3C1* interacts with psychosocial stress on ADHD severity, and that this gene–environment interaction (G × E) is further moderated by the serotonin transporter gene *5-HTT*.[20] Serotonin signaling is tightly coupled to the regulation of HPA axis activity,[21] and *5-HTT* is one of several serotonergic genes that have been repeatedly linked to ADHD.[22,23] The most extensively studied candidate genes for ADHD, the dopamine transporter (*DAT1*) and dopamine receptor D4 (*DRD4*) are also known to influence the effect of stressors on HPA axis activity.[24,25] Besides reports of G × E, stress-response genes have also been found to moderate each other's effects on the HPA axis,[26,27] illustrating the complexity of the genetic architecture underlying the stress-response pathway.

Although 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 contribution 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 etiology of complex traits using high-dimensional data.[28] It allows for inclusion of many more predictors than there are respondents, and automatically incorporates all higher-order interactions between the predictors in its estimates.[29] RFR has been praised for its robustness and predictive accuracy, particularly for noisy data containing many predictors with small effects.[30,31] Studies simulating complex genetic data sets have shown that it outperforms other techniques when it comes to detecting interacting single-nucleotide polymorphisms (SNPs) with small marginal effects.[32]

In this study, we utilized random forest regression to predict ADHD severity from SNPs in genes previously implicated to influence HPA axis activity, together with measures of 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.[33–36] Our aim was not to optimize prediction *per se*, but to improve our understanding of the complicated relation between stress-response genetics and ADHD, by estimating the contributions of thousands of HPA-axis-related SNPs plus exposure to stressors simultaneously. We thereby sought to illustrate the complex genetic architecture of this disorder and to identify those factors that are of particular interest for follow-up research. Given the intricacy of the stress response,[6] and the heterogeneity of findings in the literature regarding the relation between the HPA axis and ADHD,[9] we hypothesized that many factors with small effects are involved. The strengths of random forest regression, particularly its ability to take into account higher-order interactions between many predictors, may therefore make it particularly well suited for this task. In addition, based on the same literature, we suspected that co-occurring conduct problems may be an important influence on the relation between ADHD and the HPA axis; we therefore also sought to investigate its role in our findings. 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 subthreshold ADHD. This sample composition thus 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.[37]

MATERIALS AND METHODS

The participants were selected from the NeuroIMAGE study, a follow-up of the Dutch part of the International Multicenter ADHD Genetics (IMAGE) study.[38] NeuroIMAGE includes 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. The ADHD families were recruited through ADHD outpatient clinics in the regions Amsterdam, Groningen and Nijmegen (The Netherlands). The control families were recruited through primary and high schools in the same geographical regions. To be included in NeuroIMAGE, the participants had to be of European Caucasian descent, between the ages 5 and 30, have an intelligence quotient $\geqslant 70$ and no diagnosis of autism, epilepsy, a general learning difficulty, a brain disorder or a known genetic disorder. 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 Center. All the participants and their parents (if the participant was younger than 18 years) signed informed consent; parents signed informed consent for participants under 12 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 (that is, 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 5 criteria on the basis of a combination of a semi-structured interview and the Conners' Rating Scales.[39] The participants were asked to withhold the use of their stimulant medication or other psychoactive drugs for 48 h before measurement. The mean age of this sample was 17.1 years (s.d. 3.4) and 52.3% were males. In this sample, 95 participants had an oppositional defiant disorder or conduct disorder, 22 had an internalizing disorder and 79 had a reading disorder. More information on the NeuroIMAGE study, its diagnostic algorithm and its participants is presented in the Supplementary Information and in ref. 38.

ADHD outcome measure

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 CPRS (Conners' Parent Rating Scale), which has been shown to have high test–retest reliability and strong discriminatory power.[39] This measure consists of 18 items asking about the 18 DSM symptoms of inattention and hyperactivity impulsivity, each rated on a four-point Likert scale (0: not at all true, to 3: very much true). In this sample, the score ranged from 0 to 53, with an average of 13.1 (s.d. 12.1). This measure was available for all the participants, from both ADHD families and control families.

Given the family design of NeuroIMAGE, we calculated the intraclass correlation for our outcome measure to estimate the degree of non-independence of the data.[40] Using Searle's exact confidence limit equation, we found a nonsignificant intraclass correlation of 0.088 with a 95% confidence interval ranging from −0.023 to 0.196, with an average cluster (family) size of 1.90, indicating the non-independence is rather low.

Stress exposure

Two questionnaires were used to assess exposure to psychosocial stress. Parents filled in the long-term difficulties questionnaire,[41] 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 questionnaire,[42,43] containing 11 items on exposure to specific major stressful events in the past 5 years, such as death or serious illness of a loved one, physical or sexual abuse, or failure at something important to them. Scores on the long-term difficulties and stressful live events questionnaires have been shown to correlate with cortisol and other biological measures of stress, as well as to be predictive

Table 1. List of genes based on our literature search

Gene	*Protein product*	*Chr.*	*Start bp*	*End bp*	*SNPs*
ACE	Angiotensin-converting enzyme	17	61454422	61675741	181
ADRA2B	Alpha2B adrenergic receptor	2	96678623	96881888	144
APOE	Apolipoprotein E	19	45309039	45512650	481
AVPR1A	Arginine vasopressin receptor 1A	12	63436539	63646590	655
AVPR1B	Arginine vasopressin receptor 1B	1	206124283	206331482	164
BDNF	Brain-derived neurotrophic factor	11	27576442	27822600	368
CHRNA7	Alpha7 nicotinic acetylcholine receptor	15	32222686	32562384	394
COMT	Catechol-O-methyltransferase	22	19829263	20057498	795
CRHBP	Corticotropin-releasing hormone binding protein	5	76148680	76365299	388
CRHR1	Corticotropin-releasing hormone receptor 1	17	43761646	44013194	229
CRHR2	Corticotropin-releasing hormone receptor 2	7	30591559	30839719	575
DRD4	Dopamine receptor D4	11	537305	740705	615
ESR1	Estrogen receptor alpha	6	152028454	152524408	1323
FKBP5	FK506 binding protein 5	6	35441362	35756719	820
GABRA6	Gamma-aminobutyric acid A receptor alpha 6	5	161012658	161229598	519
HTR1A	Serotonin receptor 1A	5	63155875	63358119	264
MC2R	Melanocortin 2 Receptor	18	13782043	14015535	905
NPSR1	Neuropeptide S receptor	7	34597897	35017944	1313
NPY	Neuropeptide Y	7	24223807	24431484	905
NR3C1	Glucocorticoid receptor	5	142557496	142884045	481
NR3C2	Mineralocorticoid receptor	4	148899915	149463672	1147
OPRK1	Kappa opioid receptor	8	54038276	54264194	904
OPRM1	Mu opioid receptor	6	154260443	154540594	778
OXTR	Oxytocin receptor	3	8692095	8911300	538
PER1	Period circadian protein homolog 1	17	7943788	8155753	552
PER3	Period circadian protein homolog 3	1	7744714	8005237	861
SLC6A3	Dopamine transporter	5	1292905	1545543	442
SLC6A4	Serotonin transporter	17	28421337	28662986	305
STMN1	Stathmin	1	26110677	26332993	328

A total of 17 374 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.

of later mental health problems, in large longitudinal cohort studies of child development.[41–43] See the Supplementary Information for the full list of items, and van der Meer *et al.*[2] for a more extensive description of its use in the NeuroIMAGE cohort.

Genetics

Given our hypothesis that many factors are involved, based on the intricacy of the stress response and the inconsistencies in the literature on the relation between ADHD and HPA-axis-related genes, we took an inclusive approach regarding the selection of SNPs. We included all the available SNPs in all genes coupled to the regulation of the HPA axis activity, as indicated by the 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). We made use of the wildcard symbol * and PubMed's mesh terms to find as many relevant articles as possible. 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 Supplementary Table S1 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 were located on the X-chromosome, for which no genotyping data were available. All SNPs within 100 kilo base pairs (kb) of the location of the remaining 29 genes,[44] as found in human assembly GRCh37 were included in the study, for a total of 17 374 SNPs. 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 Supplementary Information.

For the IMAGE sample, DNA was extracted from the blood samples or immortalized cell lines at Rutgers University Cell and DNA Repository, New Jersey, USA.[45] DNA isolation for additional samples from the NeuroIMAGE study was performed at the department of Human Genetics of the Radboud University Medical Center in Nijmegen.[38]

Genome-wide genotyping was performed using the Infinitum PsychArray-24 v1.1 BeadChip, containing 265 000 tag SNPs, 245 000 exome markers and 50 000 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 < 0.994. Clustering was done using GeneTrain 2.0 (no-call threshold 0.15), excluding samples with call rate < 0.98. Before quality control, the data set contained 594 663 SNPs. Basic quality control steps included checks for sex mismatches, visualization of sample relatedness and assessment of genetic homogeneity using multidimensional scaling. No individuals were removed based on sex mismatches or population structure. Four individuals were removed based on identity by descent estimation (two identical twin pairs and two duplicate sample pairs were detected). Further quality control included removal of SNPs with a call rate below 98% or call rate differences between cases and controls higher than 2%, or failing the Hardy–Weinberg equilibrium test at a threshold of $P \leqslant 10^{-6}$. Individuals with a call rate below 98% or heterozygosity rate of more than three standard deviations from the mean ($n = 33$) were removed as well. After quality control, the data set contained 584 262 SNPs. A further 221 865 SNPs with a minor allele frequency of less than 1% were removed from the set before imputation. Imputation was carried out according to the protocol supplied by ENIGMA (http://enigma.ini.usc.edu/), using MaCH[46] for haplotype phasing and minimac[47] for imputation, with 1000 Genomes Phase 1 V3 reference data.[48] We reasoned that imputation makes more genetic information available for the analysis[49] and therefore allows for a more comprehensive assessment of the true relation between ADHD and variation in genes

influencing the stress response. SNPs with low imputation quality ($R^2 < 0.8$) were filtered out. Subsequently, hard calls, needed as input for the analysis, were made by converting to PLINK format,[50] using GCTA software.[51]

Random forest regression analysis

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.[29] Observations not included in a tree's sample due to the bootstrapping procedure, called out-of-bag (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 out-of-bag samples and comparing the resulting prediction error to the error obtained before the shuffle.[52] 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 deeper in the tree.

All analyses were run in R v3.2.3,[53] making use of the package randomForestSRC v2.2.0.[54] The code used is available upon request from the corresponding author. The 17 374 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. This approach ensured that all information was maintained, that is, the marginal and interaction effects of each 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.[55]

We grew 5000 trees fully and used the default value of $p/3$ for *mtry*, the size of the random subset of available predictors at each split, in this case 5800 (17 398/3 rounded up). These settings were chosen to identify important predictors while still allowing for the detection of true predictors with small effects and interactions, and in accordance with recommendations from simulation studies on complex genetic data with interacting SNPs.[56] We further checked the stability of the results by rerunning the analysis twice, with different random seeds.

The reported percent variance explained is calculated as 1 − (mean-squared error/variance of y), with mean-squared error calculated from the difference between the observed score and the predicted score, averaged over all trees where the observation was 'out-of-bag'.[57] 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 out-of-bag sample; the resulting increase in prediction error, averaged over all trees, is expressed as percent increase in mean-squared error.[29,57] 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 by either of the predictors influences the likelihood of a subsequent split by the other predictor, with a negative numeral indicating an increased likelihood that one is selected in the subtree of the other and a positive numeral indicating a reduced likelihood, as explained fully elsewhere.[52] 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. The interaction estimates are multiplied by 100 for ease of display.

Supplementary Figure S1 shows the Spearman's rank correlation coefficient between each pair of the 25 highest-ranked predictors.

Supplementary analyses

Many studies on the HPA axis and related genes in ADHD have shown that especially co-occurring conduct problems drives HPA-axis-related differences with typically developing controls. Given conduct disorder was also among the most common comorbidities in this sample, we ran two additional RFR analyses aimed at providing an indication of the role of co-occurring conduct problems in our findings. We used the score on the CPRS subscale A, which has been found to specifically measure conduct problems rather than externalizing behaviors associated with ADHD in general.[39] We ran one analysis where we added this measure as a predictor to the original model, with ADHD severity as outcome, and a second one where we used the score on the CPRS subscale A as outcome, adding ADHD severity to the set of predictors from the main analysis. See the Supplementary Information for more information on these analyses.

RESULTS

The model explained 12.5% variance in ADHD severity. Permuting all SNPs simultaneously led to an 8.3% increase in mean-squared error compared with the intact model. For all stress items together, this was 25.3%. 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.

For both additional analyses into the role of conduct problems, we found the same long-term difficulties and the same SNPs in *PER3*, *ESR1* and *NR3C2* that were also among the top predictors in the main analysis. SNPs in *SLC6A3*, *NPSR1*, *DRD4* and *GABRA6* remained among the top hits when CPRS subscale A was included as a predictor, but not when it was used as the outcome. Detailed output can be found in the Supplementary Information.

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 the ability to handle high-dimensional data and to take into account all possible interactions, align well with the complexity of stress-response genetics. We reasoned that this would enable us to identify important contributors to ADHD severity, and to document 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, together with a meta-analysis, have provided strong cumulative evidence that this SNP is also associated with risk for several forms of cancer.[58] Although we included rs4635969 as part of the 3′ end-flanking region of *SLC6A3*, it is possible that this finding is explained by its close proximity to other genes, such as micro-RNA (MIR4457) at the 5′ end of the telomerase reverse transcriptase (*TERT*) gene, known to regulate telomere length.[59] Overexpression of *TERT* increases cell proliferation and resilience to oxidative stress,[60] whereas glucocorticoid administration and chronic stress exposure have been shown to lower basal telomerase activity and shorten telomere length.[61,62] 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 finding, together with reports on children's telomere length being related to early social deprivation[63] and hyperactivity/impulsivity,[64] suggests this SNP is of interest for ADHD and G × E research.

The other high-ranked SNPs were in or near *NPSR1*, *ESR1*, *GABRA6*, *PER3*, *DRD4*, *NR3C2* and *OPRK1*. Besides their associations with HPA axis activity (references listed in Supplementary Table S1), polymorphisms in these genes have all been repeatedly, but inconsistently, associated with internalizing and externalizing behavior often co-occurring with ADHD.[65–72] This inconsistency mirrors the heterogeneity of findings on the relation of cortisol with ADHD as well as with internalizing and externalizing behavior, which have indicated that low reactivity of the HPA axis is most prominent in individuals with ADHD and co-occurring externalizing disorders while high HPA axis activity relates more to

Table 2. Top 25 most important predictors, based on the increase in prediction error following permutation

Rank	*Stressor*	*Frequency*	*VIMP*
1	Your child has a chronic illness or handicap	0.23	15.01
2	Your child has fewer friends than he/she would like	0.15	4.64
3	Your child is being bullied at school or in the neighborhood	0.07	3.61
4	Your child can't get along with someone in your immediate family	0.08	1.24
6	Your immediate family has financial difficulties	0.04	0.35

	RS ID	*Location*	*Gene*	*Region*		
5	rs4635969	5:1308552	*SLC6A3*	84 kb from 3′ end	0.20	0.44
7	rs35311906	7:34873557	*NPSR1*	Intron	0.15	0.23
8	rs985191	6:152283458	*ESR1*	Intron	0.11	0.21
9	rs10035808	5:161189729	*GABRA6*	60 kb from 3′ end	0.46	0.19
10	rs11587880	1:7822957	*PER3*	22 kb from 5′ end	0.08	0.17
11	rs7932167	11:620599	*DRD4*	17 kb from 5′ end	0.19	0.16
12	rs77714417	6:152274190	*ESR1*	Intron	0.06	0.13
13	rs56821207	1:7951115	*PER3*	46 kb from 3′ end	0.18	0.13
14	rs179265	1:7942692	*PER3*	37 kb from 3′ end	0.45	0.12
15	rs11587479	1:7823085	*PER3*	22 kb from 5′ end	0.08	0.12
16	rs74325817	6:152271827	*ESR1*	Intron	0.11	0.12
17	rs35365822	6:152270364	*ESR1*	Intron	0.11	0.12
18	rs2530547	7:34697922	*NPSR1*	5′-UTR	0.36	0.11
19	rs9340910	6:152272233	*ESR1*	Intron	0.11	0.10
20	rs77595592	5:161034192	*GABRA6*	78 kb from 5′ end	0.17	0.10
21	rs35953391	5:1312329	*SLC6A3*	81 kb from 3′ end	0.20	0.10
22	rs6930114	6:152268250	*ESR1*	Intron	0.11	0.09
23	rs35527038	4:148970403	*NR3C2*	30 kb from 3′ end	0.15	0.09
24	rs10002896	4:149028802	*NR3C2*	Intron	0.24	0.09
25	rs143748464	8:54146601	*OPRK1*	Intron	0.03	0.09

Abbreviations: kb, kilo base pair; RS ID, reference SNP identification number; UTR, untranslated region; VIMP, Breiman–Cutler variable importance estimate. The five stressors are listed first, followed by details on the 20 single-nucleotide polymorphisms (SNPs). 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.

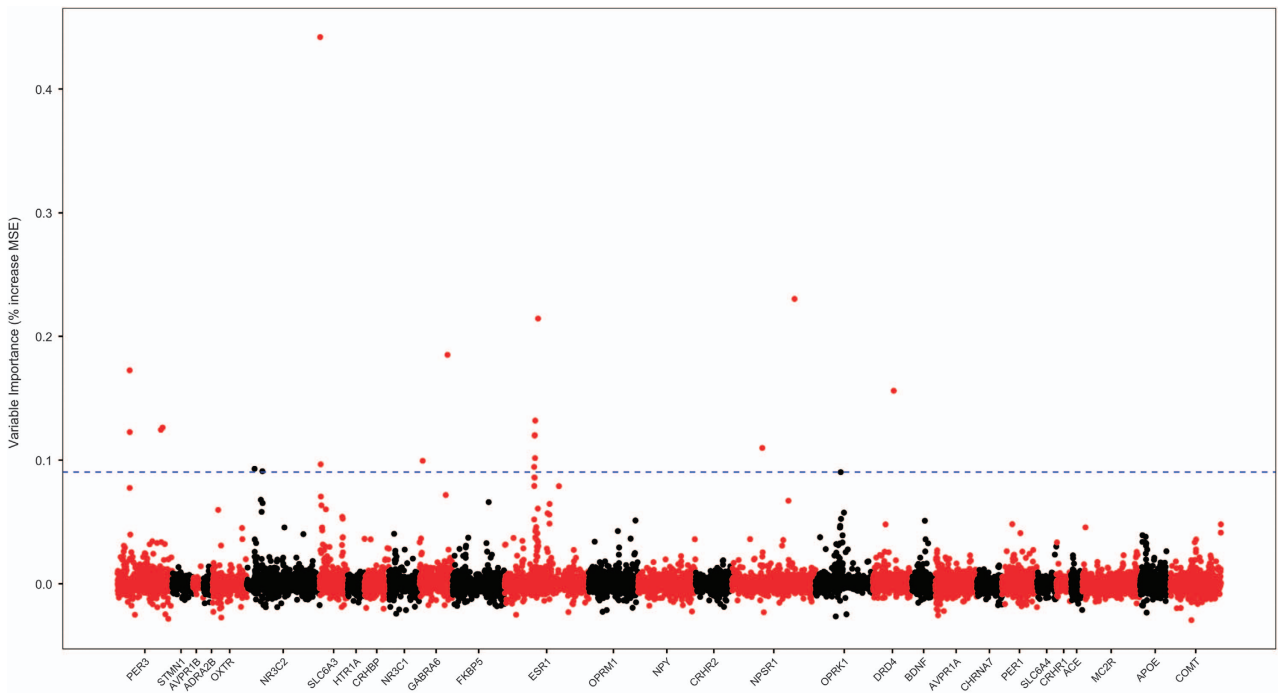

Figure 1. Variable importance for prediction, for all single-nucleotide polymorphisms (SNPs) included in the analysis. SNPs are ordered on the *x* axis based on their genomic position, from chromosome 1 to 22, with the labels and alternating red and black sections marking the gene they belong to. The *y* axis indicates the variable importance, as percent increase in mean-squared error (MSE) 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.

	Few friends	Bullied	Family argues	PER3 rs56821207	OPRK1 rs143748464	ESR1 rs985191	ESR1 rs77714417	Financial difficulties	NPSR1 rs2530547	GABRA6 rs77595592	NR3C2 rs10002896	ESR1 rs6930114	ESR1 rs9340910	ESR1 rs74325817	ESR1 rs35365822	PER3 rs11587479	GABRA6 rs10035808	PER3 rs179265	PER3 rs11587880	NPSR1 rs35311906	NR3C2 rs35527038	DRD4 rs7932167	SLC6A3 rs35953391	SLC6A3 rs4635969
Illness or handicap	36	9	6	3	2	0	0	1	0	−1	−1	−3	−3	−3	−4	−4	−3	−3	−5	−6	−9	−1	−16	−46
Few friends	9	−2	0	1	1	0	−1	0	0	−1	1	1	0	0	−1	−2	−2	0	1	0	−16	−1	0	
Bullied	−3	1	1	0	1	−1	1	1	−1	0	1	−1	0	0	0	0	0	0	0	0	−1	0	0	
Family argues	1	0	0	−1	1	0	1	0	0	0	0	0	0	0	−1	1	0	0	0	0	0	0	0	
PER3 rs56821207	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	
OPRK1 rs143748464	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ESR1 rs985191	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ESR1 rs77714417	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
Financial difficulties	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	
NPSR1 rs2530547	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
GABRA6 rs77595592	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
NR3C2 rs10002896	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	0	0	
ESR1 rs6930114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ESR1 rs9340910	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ESR1 rs74325817	−1	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ESR1 rs35365822	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
PER3 rs11587479	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
GABRA6 rs10035808	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
PER3 rs179265	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
PER3 rs11587880	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
NPSR1 rs35311906	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
NR3C2 rs35527038	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
DRD4 rs7932167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
SLC6A3 rs35953391	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 2. Interaction strengths for each pair of 25 top predictors from 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 on the basis of the first principal component of their interaction strengths.

anxiety and depression.[9,73] If early splits in a tree form groups more homogeneous with regard to, for instance, externalizing behavior, they allow for detection of other SNPs that impact ADHD severity only in these individuals and not in, for example, 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. The ability of this technique to capture shared genetics of psychiatric disorders[74] is corroborated by our additional analyses, showing that the polymorphisms in *ESR1, NR3C2* and *PER3* were also among the top predictors for our measure of conduct problems in this sample. The other top hits appeared to be more specific to ADHD. The reported associations may still be influenced by any of the range of co-occurring problems seen in ADHD, which would contribute to inconsistent findings across studies. Follow-up studies investigating the relation between ADHD and HPA-axis-related factors should therefore carefully consider comorbid conditions.

We further found that particularly long-term difficulties, compared with stressful live events, are important for predicting ADHD severity as well as co-occurring conduct problems. This stronger influence of chronic stress may be explained by the principles of the allostatic load model and its implications for psychiatric disorders.[75] 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, thought to contribute to the relation of stress with a range of psychiatric disorders.[5,76] High allostatic load may 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.[77] 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.[78,79]

We included many predictors in this analysis that are correlated with each other. Whether correlation between predictors, such as SNPs in regions of high LD, or exposure to different concurrent stressors, helps or hinders random forests depends on the aim of the study.[80] 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.[52] Correlation between predictors may, however, be beneficial for the analysis of the type of high-

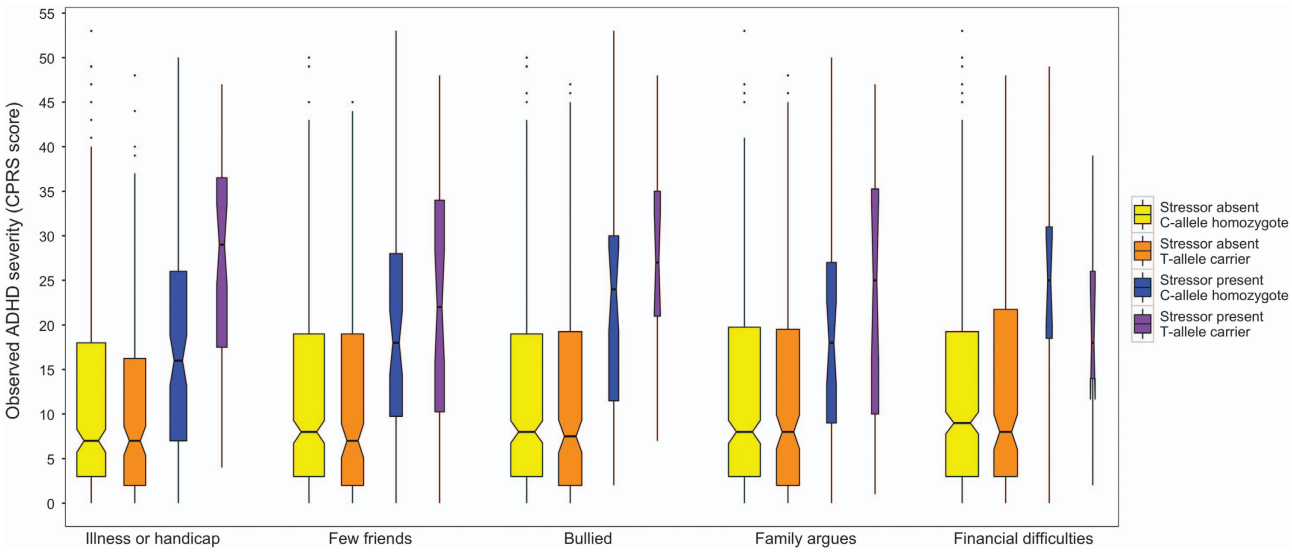

Figure 3. Visualization of the interaction between SLC6A3 rs4635969 and each of the five long-term difficulties among the top predictors. The participants are grouped based on their genotype and exposure to the individual long-term difficulty shown on the *x* axis. On the *y* axis is the observed score on the Conners' Parent Rating Scale (CPRS), subscale N. The boxes show the median, and the first and third quantiles of each group. Their width is scaled by the number of participants. ADHD, attention-deficit/hyperactivity disorder.

dimensional data encountered in genetic studies; 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 the top hits. 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,[80] 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.[56] 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 approach taken in this study should be seen as complementary to the conventional statistical techniques used in ADHD etiological studies. Random forest regression has great potential as an exploratory tool, given its ability to handle high-dimensional data, and to produce measures of importance. However, the interpretability of its results has been criticized; whereas the findings from conventional regression analyses can be relatively easily probed, for example, by plotting the association on the basis of the regression 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 small interaction effects contribute to the overall predictive accuracy, but that current measures are unable to identify them.[81] While gene–gene interactions may explain a considerable amount of the heritability of ADHD that currently remains unaccounted for,[82,83] the effects of individual SNPs and their interactions are likely to be small, and their estimated size is further diminished by the LD between SNPs with the current inclusive approach. This may explain the lack of noteworthy gene–gene interactions shown in Figure 2, whereas interactions between the strongest predictors, predominantly the long-term difficulties, do get identified.

It should be noted that this was a cross-sectional study, which precludes any statements on the nature of the relation between the SNPs, the stressors and ADHD, and therefore 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, although this does not make it any less of an interesting target for further research. Other stressors, such as having few friends, may partly result from ADHD-related behavior; the direction of effects may be teased apart by longitudinal studies. We further were unable to correct for the presence of siblings in the sample. Although we showed that the degree of non-independence was low and we did not perform any inferential statistics, we cannot rule out that the family design influenced the pattern of the results. We also chose not to add an additional, external, round of validation because of the relatively small sample size for a genetics study, which limits confidence in the findings. Depending on the goal of the study and the available sample size, future studies may choose other approaches, such as a discovery-replication approach and/or LD pruning of the SNP selection.

To summarize, in this exploratory study, we aimed to illustrate the strengths of random forest regression, an ensemble learning method that may be useful for exploring high-dimensional data to discover associations with ADHD. Besides documenting 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 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 G × E. Inference on such a selection, however, should take place in independent samples. We identified a novel association between ADHD severity and a SNP that may relate to *TERT,* suggesting 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 as those with conduct problems, thereby contributing to inconsistent association of stress-response genes with ADHD. Our results also illustrated the strong effects of chronic

stress, not found for individual stressful events, in accordance with allostatic load models.[75] This explorative study may best be followed up by selecting the strongest predictors, analyzing whether the effects of this selection replicate in independent samples, and investigating how and why these are dependent on each other.

CONFLICT OF INTEREST

BF has received a speaker fee from Merz. PJH has received an unrestricted research grant from Shire and has been member of the advisory boards of Shire and Eli Lilly. JO has received an unrestricted investigator initiated research grant from Shire pharmaceuticals. JKB has been in the past 3 years a consultant to/member of advisory board of/and/or speaker for Janssen Cilag BV, Eli Lilly, Shire, Novartis, Roche and Servier. He is not an employee of any of these companies and not a stock shareholder of any of these companies. In the past year, SVF received income, travel expenses and/or research support from Pfizer, Ironshore, Shire, Akili Interactive Labs, Alcobra, VAYA Pharma and SynapDx, and research support from the National Institutes of Health (NIH). His institution is seeking a patent for the use of sodium–hydrogen exchange inhibitors in the treatment of ADHD. In previous years, he received consulting fees or was on Advisory Boards or participated in continuing medical education programs sponsored by Shire, Alcobra, Otsuka, McNeil, Janssen, Novartis, Pfizer and Eli Lilly. SVF receives royalties from books published by Guilford Press: *Straight Talk about Your Child's Mental Health* and Oxford University Press: *Schizophrenia: The Facts*. The remaining authors declare no conflict of interest.

ACKNOWLEDGMENTS

We acknowledge the Department of Pediatrics of the VU University Medical Center for having the opportunity to use the mock scanner for preparation of our participants. This work was supported by NIH Grant R01MH62873 (to SVF), NWO Large Investment Grant 1750102007010 and NWO Brain and Cognition an Integrative Approach grant (433-09-242) (to JKB), and grants from Radboud University Nijmegen Medical Center, University Medical Center Groningen and Accare, and VU University Amsterdam. The research leading to these results also received funding from the European Community's Seventh Framework Programme (FP7/2007– 2013) under grant agreement numbers 278948 (TACTICS), 602450 (IMAGEMEND) and number 602805 (Aggressotype), and from the European Community's Horizon 2020 Programme (H2020/2014–2020) under grant agreement number 643051 (MiND). BF is supported by a Vici grant from NWO (grant number 016-130-669). In addition, JKB and BF are supported by a grant for the ENIGMA Consortium (grant number U54 EB020403) from the BD2K Initiative of a cross-NIH partnership.

REFERENCES

1 Faraone SV, Asherson P, Banaschewski T, Biederman J, Buitelaar JK, Ramos-Quiroga JA *et al.* Attention-deficit/hyperactivity disorder. *Nat Rev Dis Primers* 2015; **1**: 15020.
2 van der Meer D, Hartman CA, Richards J, Bralten JB, Franke B, Oosterlaan J *et al.* The serotonin transporter gene polymorphism 5-HTTLPR moderates the effects of stress on attention-deficit/hyperactivity disorder. *J Child Psychol Psychiatry* 2014; **55**: 1363–1371.
3 Banerjee TD, Middleton F, Faraone SV. Environmental risk factors for attention-deficit hyperactivity disorder. *Acta Paediatr* 2007; **96**: 1269–1274.
4 Kudielka BM, Hellhammer DH, Wust S. Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoneuroendocrinology* 2009; **34**: 2–18.
5 McEwen BS, Bowles NP, Gray JD, Hill MN, Hunter RG, Karatsoreos IN *et al.* Mechanisms of stress in the brain. *Nat Neurosci* 2015; **18**: 1353–1363.
6 Joels M, Baram TZ. The neuro-symphony of stress. *Nat Rev Neurosci* 2009; **10**: 459–466.
7 Lupien SJ, McEwen BS, Gunnar MR, Heim C. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nat Rev Neurosci* 2009; **10**: 434–445.
8 Scassellati C, Bonvicini C, Faraone SV, Gennarelli M. Biomarkers and attention-deficit/hyperactivity disorder: a systematic review and meta-analyses. *J Am Acad Child Adolesc Psychiatry* 2012. *51* **10**: 1003–1019, e1020.
9 Corominas M, Ramos-Quiroga JA, Ferrer M, Sáez-Francàs N, Palomar G, Bosch R *et al.* Cortisol responses in children and adults with attention deficit hyperactivity disorder (ADHD): a possible marker of inhibition deficits. *Atten Defic Hyperact Disord* 2012; **4**: 63–75.
10 Freitag CM, Hänig S, Palmason H, Meyer J, Wüst S, Seitz C. Cortisol awakening response in healthy children and children with ADHD: impact of comorbid disorders and psychosocial risk factors. *Psychoneuroendocrinology* 2009; **34**: 1019–1028.
11 Christiansen H, Oades RD, Psychogiou L, Hauffa BP, Sonuga-Barke EJ. Does the cortisol response to stress mediate the link between expressed emotion and oppositional behavior in attention-deficit/hyperactivity-disorder (ADHD)? *Behav Brain Funct* 2010; **6**: 1.
12 Kariyawasam SH, Zaw F, Handley SL. Reduced salivary cortisol in children with comorbid attention deficit hyperactivity disorder and oppositional defiant disorder. *Neuroendocrinol Lett* 2002; **23**: 45–48.
13 Adam EK. Emotion—cortisol transactions occur over multiple time scales in development: implications for research on emotion and the development of emotional disorders. *Monogr Soc Res Child Dev* 2012; **77**: 17–27.
14 Van Lenten SA, Doane LD. Examining multiple sleep behaviors and diurnal salivary cortisol and alpha-amylase: within- and between-person associations. *Psychoneuroendocrinology* 2016; **68**: 100–110.
15 Baird AL, Coogan AN, Siddiqui A, Donev RM, Thome J. Adult attention-deficit hyperactivity disorder is associated with alterations in circadian rhythms at the behavioural, endocrine and molecular levels. *Mol Psychiatry* 2012; **17**: 988–995.
16 Shaw P, Stringaris A, Nigg J, Leibenluft E. Emotion dysregulation in attention deficit hyperactivity disorder. *Am J Psychiatry* 2014; **171**: 276–293.
17 Cortese S, Brown TE, Corkum P, Gruber R, O'Brien LM, Stein M *et al.* Assessment and management of sleep problems in youths with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry* 2013; **52**: 784–796.
18 Fortier ME, Sengupta SM, Grizenko N, Choudhry Z, Thakur G, Joober R. Genetic evidence for the association of the hypothalamic-pituitary-adrenal (HPA) axis with ADHD and methylphenidate treatment response. *Neuromol Med* 2013; **15**: 122–132.
19 Buckingham JC. Glucocorticoids: exemplars of multi-tasking. *Br J Pharmacol* 2006; **147**(Suppl 1): S258–S268.
20 van der Meer D, Hoekstra PJ, Bralten J, van Donkelaar M, Heslenfeld DJ, Oosterlaan J *et al.* Interplay between stress response genes associated with attention deficit-hyperactivity disorder and brain volume. *Genes Brain Behav* 2016; **15**: 627–636.
21 Leonard BE. The HPA and immune axes in stress: the involvement of the serotonergic system. *Eur Psychiatry* 2005; **20**: S302–S306.
22 Gizer IR, Ficks C, Waldman ID. Candidate gene studies of ADHD: a meta-analytic review. *Hum Genet* 2009; **126**: 51–90.
23 Oades RD, Lasky-Su J, Christiansen H, Faraone SV, Sonuga-Barke EJ, Banaschewski T *et al.* The influence of serotonin- and other genes on impulsive behavioral aggression and cognitive impulsivity in children with attention-deficit/hyperactivity disorder (ADHD): findings from a family-based association test (FBAT) analysis. *Behav Brain Funct* 2008; **4**: 1–14.
24 Alexander N, Osinsky R, Mueller E, Schmitz A, Guenthert S, Kuepper Y *et al.* Genetic variants within the dopaminergic system interact to modulate endocrine stress reactivity and recovery. *Behav Brain Res* 2011; **216**: 53–58.
25 Buchmann AF, Zohsel K, Blomeyer D, Hohm E, Hohmann S, Jennen-Steinmetz C *et al.* Interaction between prenatal stress and dopamine D4 receptor genotype in predicting aggression and cortisol levels in young adults. *Psychopharmacology (Berl)* 2014; **231**: 3089–3097.
26 Armbruster D, Mueller A, Moser DA, Lesch KP, Brocke B, Kirschbaum C. Interaction effect of D4 dopamine receptor gene and serotonin transporter promoter polymorphism on the cortisol stress response. *Behav Neurosci* 2009; **123**: 1288–1295.
27 Clasen PC, Wells TT, Knopik VS, McGeary JE, Beevers CG. 5-HTTLPR and BDNF Val66Met polymorphisms moderate effects of stress on rumination. *Genes Brain Behav* 2011; **10**: 740–746.
28 Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics* 2012; **99**: 323–329.
29 Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
30 Scornet E, Biau G, Vert J-P. Consistency of random forests. *Ann Stat* 2015; **43**: 1716–1741.
31 Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* 2014; **15**: 3133–3181.
32 Lunetta KL, Hayward LB, Segal J, Eerdewegh PV. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* 2004; **5**: 32.
33 Sato JR, Hoexter MQ, Fujita A, Rohde LA. Evaluation of pattern recognition and feature extraction methods in ADHD prediction. *Front Syst Neurosci* 2012; **6**: 68.
34 Brown MRG, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH *et al.* ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front Syst Neurosci* 2012; **6**: 69.
35 Fair DA, Bathula D, Nikolas MA, Nigg JT. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *Proc Natl Acad Sci USA* 2012; **109**: 6769–6774.
36 Eloyan A, Muschelli J, Nebel MB, Liu H, Han F, Zhao T *et al.* Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Front Syst Neurosci* 2012; **6**: 61.
37 Larsson H, Anckarsater H, Råstam M, Chang Z, Lichtenstein P. Childhood attention-deficit hyperactivity disorder as an extreme of a continuous trait: a quantitative genetic study of 8,500 twin pairs. *J Child Psychol Psychiatry* 2012; **53**: 73–80.

38 von Rhein D, Mennes M, van Ewijk H, Groenman AP, Zwiers MP, Oosterlaan J *et al.* The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *Eur Child Adolesc Psychiatry* 2014; **24**: 265–281.
39 Conners CK, Sitarenios G, Parker JD, Epstein JN. The revised Conners' Parent Rating Scale (CPRS-R): factor structure, reliability, and criterion validity. *J Abnorm Child Psychol* 1998; **26**: 257–268.
40 Hox JJ, Moerbeek M, van de Schoot R. *Multilevel Analysis: Techniques and Applications*. Routledge: New York, NY, 2010.
41 Zandstra AR, Hartman CA, Nederhof E, van den Heuvel ER, Dietrich A, Hoekstra PJ *et al.* Chronic stress and adolescents' mental health: modifying effects of basal cortisol and parental psychiatric history. The TRAILS Study. *J Abnorm Child Psychol* 2015; **43**: 1119–1130.
42 Bosch NM, Riese H, Reijneveld SA, Bakker MP, Verhulst FC, Ormel J *et al.* Timing matters: long term effects of adversities from prenatal period up to adolescence on adolescents' cortisol stress response. The TRAILS study. *Psychoneuroendocrinology* 2012; **37**: 1439–1447.
43 Oldehinkel AJ, Verhulst FC, Ormel J. Low heart rate: a marker of stress resilience. The TRAILS study. *Biol Psychiatry* 2008; **63**: 1141–1146.
44 Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008; **4**: e1000214.
45 Brookes K, Xu X, Chen W, Zhou K, Neale B, Lowe N *et al.* The analysis of 51 genes in DSM-IV combined type attention deficit hyperactivity disorder: association signals in DRD4, DAT1 and 16 other genes. *Mol Psychiatry* 2006; **11**: 934–953.
46 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
47 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012; **44**: 955–959.
48 Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
49 Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
50 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
51 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**: 76–82.
52 Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat* 2007; **1**: 519–537.
53 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2015.
54 Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; **2**: 841–860.
55 Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007; **8**: 1.
56 Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M *et al.* SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics* 2012; **13**: 1–13.
57 Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; **2**: 18–22.
58 Mocellin S, Verdi D, Pooley KA, Landi MT, Egan KM, Baird DM *et al.* Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis. *J Natl Cancer Inst* 2012; **104**: 840–854.
59 Diaz de Leon A, Cronkhite JT, Katzenstein A-LA, Godwin JD, Raghu G, Glazer CS *et al.* Telomere lengths, pulmonary fibrosis and telomerase (*TERT*) mutations. *PLoS ONE* 2010; **5**: e10680.
60 Armstrong L, Saretzki G, Peters H, Wappler I, Evans J, Hole N *et al.* Overexpression of telomerase confers growth advantage, stress resistance, and enhanced differentiation of ESCs toward the hematopoietic lineage. *Stem Cells* 2005; **23**: 516–529.
61 Epel ES, Blackburn EH, Lin J, Dhabhar FS, Adler NE, Morrow JD *et al.* Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci USA* 2004; **101**: 17312–17315.
62 Epel ES, Lin J, Dhabhar FS, Wolkowitz OM, Puterman E, Karan L *et al.* Dynamics of telomerase activity in response to acute psychological stress. *Brain Behav Immun* 2010; **24**: 531–539.
63 Drury SS, Theall K, Gleason MM, Smyke AT, De Vivo I, Wong JY *et al.* Telomere length and early severe social deprivation: linking early adversity and cellular aging. *Mol Psychiatry* 2012; **17**: 719–727.
64 de Souza Costa D, Rosa DVF, Barros AGA, Romano-Silva MA, Malloy-Diniz LF, Mattos P *et al.* Telomere length is highly inherited and associated with hyperactivity-impulsivity in children with attention deficit/hyperactivity disorder. *Front Mol Neurosci* 2015; **8**: 28.
65 Laas K, Reif A, Kiive E, Domschke K, Lesch K-P, Veidebaum T *et al.* A functional NPSR1 gene variant and environment shape personality and impulsive action: a longitudinal study. *J Psychopharmacol* 2014; **28**: 227–236.
66 Comings DE, Gade-Andavolu R, Gonzalez N, Wu S, Muhleman D, Blake H *et al.* Multivariate analysis of associations of 42 genes in ADHD, ODD and conduct disorder. *Clin Genet* 2000; **58**: 31–40.
67 Mill J, Kiss E, Baji I, Kapornai K, Daroczy G, Vetro A *et al.* Association study of the estrogen receptor alpha gene (ESR1) and childhood-onset mood disorders. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147b**: 1323–1326.
68 Sundermann EE, Maki PM, Bishop JR. A review of estrogen receptor α gene (ESR1) polymorphisms, mood, and cognition. *Menopause* 2010; **17**: 874–886.
69 Hess JL, Kawaguchi DM, Wagner KE, Faraone SV, Glatt SJ. The influence of genes on "positive valence systems" constructs: a systematic review. *Am J Med Genet B Neuropsychiatr Genet* 2016; **171**: 92–110.
70 Jüngling K, Seidenbecher T, Sosulina L, Lesting J, Sangha S, Clark SD *et al.* Neuropeptide S-mediated control of fear expression and extinction: role of intercalated GABAergic neurons in the amygdala. *Neuron* 2008; **59**: 298–310.
71 Domschke K, Maron E. Genetic factors in anxiety disorders. *Mod Trends Pharmacopsychiatry* 2013; **29**: 24–46.
72 Smoller JW. The genetics of stress-related disorders: PTSD, depression, and anxiety disorders. *Neuropsychopharmacology* 2016; **41**: 297–319.
73 Marsman R, Swinkels SHN, Rosmalen JGM, Oldehinkel AJ, Ormel J, Buitelaar JK. HPA-axis activity and externalizing behavior problems in early adolescents from the general population: the role of comorbidity and gender: the TRAILS study. *Psychoneuroendocrinology* 2008; **33**: 789–798.
74 O'Donovan MC, Owen MJ. The implications of the shared genetics of psychiatric disorders. *Nat Med* 2016; **22**: 1214–1219.
75 McEwen BS. Stress, adaptation, and disease: allostasis and allostatic load. *Ann N Y Acad Sci* 1998; **840**: 33–44.
76 Liston C, McEwen BS, Casey BJ. Psychosocial stress reversibly disrupts prefrontal processing and attentional control. *Proc Natl Acad Sci USA* 2009; **106**: 912–917.
77 McEwen BS. Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. *Ann N Y Acad Sci* 2004; **1032**: 1–7.
78 van der Meer D, Hoekstra PJ, Zwiers M, Mennes M, Schweren LJ, Franke B *et al.* Brain correlates of the interaction between 5-HTTLPR and psychosocial stress mediating attention deficit hyperactivity disorder severity. *Am J Psychiatry* 2015; **172**: 768–775.
79 Gerritsen L, Tendolkar I, Franke B, Vasquez AA, Kooijman S, Buitelaar J *et al.* BDNF Val66Met genotype modulates the effect of childhood adversity on subgenual anterior cingulate cortex volume in healthy subjects. *Mol Psychiatry* 2012; **17**: 597–603.
80 Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* 2010; **11**: 1.
81 Wright MN, Ziegler A, König IR. Do little interactions get lost in dark random forests? *BMC Bioinformatics* 2016; **17**: 1.
82 Hemani G, Knott S, Haley C. An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet* 2013; **9**: e1003295.
83 Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012; **109**: 1193–1198.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)