Genome-wide association study identifies novel restless legs syndrome susceptibility loci on 2p14 and 16q12.1.

Juliane Winkelmann, Darina Czamara, Barbara Schormair, Franziska Knauf, Eva Schulte, Claudia Trenkwalder, Yves Dauvilliers, Olli Polo, Birgit Högl, Klaus Berger, et al.

To cite this version:

Juliane Winkelmann, Darina Czamara, Barbara Schormair, Franziska Knauf, Eva Schulte, et al.. Genome-wide association study identifies novel restless legs syndrome susceptibility loci on 2p14 and 16q12.1.. PLoS Genetics, Public Library of Science, 2011, 7 (7), pp.e1002171. <10.1371/journal.pgen.1002171>. <inserm-00711782>

HAL Id: inserm-00711782
http://www.hal.inserm.fr/inserm-00711782
Submitted on 25 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

Restless legs syndrome (RLS) is a sensorimotor disorder with an age-dependent prevalence of up to 10% in the general population above 65 years of age. Affected individuals suffer from uncomfortable sensations and an urge to move in the lower limbs that occurs mainly in resting situations during the evening or at night. Moving the legs or walking leads to an improvement of symptoms. Concomitantly, patients report sleep disturbances with consequences such as reduced daytime functioning. We conducted a genome-wide association study (GWA) for RLS in 922 cases and 1,526 controls (using 301,406 SNPs) followed by a replication of 76 candidate SNPs in 3,935 cases and 5,754 controls, all of European ancestry. Herein, we identified six RLS susceptibility loci of genome-wide significance, two of them novel: an intergenic region on chromosome 2p14 (rs6747972, \( P = 9.03 \times 10^{-11}, \text{OR} = 1.23 \)) and a locus on 16q12.1 (rs3104767, \( P = 9.4 \times 10^{-19}, \text{OR} = 1.35 \)) in a linkage disequilibrium block of 140 kb containing the 5′-end of TOX3 and the adjacent non-coding RNA BC034767.
Novel Restless Legs Syndrome Susceptibility Loci

Introduction

Restless legs syndrome (RLS) is a common neurological disorder with a prevalence of up to 10%, which increases with age [1]. Affected individuals suffer from an urge to move due to uncomfortable sensations in the lower limbs present in the evening or at night. The symptoms occur during rest and relaxation, with walking or moving the extremity leading to prompt relief. Consequently, initiation and maintenance of sleep become defective [1]. RLS has been associated with iron deficiency, and is pharmacologically responsive to dopaminergic substitution. Increased cardiovascular events, depression, and anxiety count among the known co-morbidities [1].

Genome-wide association studies (GWAs) identified generic risk factors within MEIS1, BTBD9, PTPRD, and a locus encompassing MAP2K5 and SKOR1 [2–4]. To identify additional RLS susceptibility loci, we undertook an enlarged GWA in a German case-control population, followed by replication in independent case-control samples originating from Europe, the United States of America, and Canada. In doing so, we identified six RLS susceptibility loci with genome-wide significance in the joint analysis, two of them novel: an intergenic region on chromosome 2p14 and a locus on 16q12.1 in close proximity to TOX3 and the adjacent non-coding RNA BC034767.

Results/Discussion

We enlarged our previously reported [2,4] GWA sample to 954 German RLS cases and 1,814 German population-based controls. From the KORA-S3/F3 survey and genotyped them on Affymetrix 5.0 (cases) and 6.0 (controls) arrays. To correct for population factors within the KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. The collection of sociodemographic and clinical data in the Dortmund Health Study was supported by the German Migraine & Headache Society (DMHG) and by unrestricted grants of equal share from Astra Zeneica, Berlin Chemie, Boots Healthcare, Glaxo-Smith-Kline, McNeil Pharma (former Woelm Pharma), MSD Sharp & Dohme, and Pfizer to the University of Muenster. Blood collection in the Dortmund Health Study was done through funds from the Institute of Epidemiology and Social Medicine, University of Muenster. Data collection in the COR-Study was supported by unrestricted grants of the German RLS Society (Deutsche Restless Legs Vereinigung e.V) and Axonris Pharma, Boehringer Ingelheim Pharma, Mundipharma Research, Roche Pharma, and UCB to the University of Muenster. CG Bachmann was supported by grants of the German RLS Society. Deutsche Restless Legs Vereinigung, e.V. H Prokisch and T Metlinger were supported by the German Federal Ministry of Education and Research (BMBF) project Systems Biology of Metabotypes (SysMBio #0315494A). RP Allen and CJ Earley were supported by the grant PO1-A21190 National Institute of Health, USA; the Canadian part of the study was supported by a Canadian Institutes of Health Research (CIHR) grant to GA Rouleau and J Montplaisir. D Kemlnik and S Nevesimalova were supported by an ESRF grant MSM0021620849; Javonv and K Sonka were supported by grant MSM0021620816. Recruitment of Czech controls was funded by grant IGA NR 8563-5, Ministry of Health of the Czech Republic. B Frauscher, I Courto-Rebeix, M Francavilla, and C Fontenille are co-authors on behalf of BRC-RFGENSEP, which is supported by INSERM, AFM (Généthon), ARSEP, and GIS-IBISA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: winkelmann@lrz.tu-muenchen.de

These authors contributed equally to this work.


Editor: Mark I. McCarthy, University of Oxford, United Kingdom

Received December 7, 2010; Accepted May 24, 2011; Published July 14, 2011

Copyright: © 2011 Winkelmann et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The replication phase was supported by a grant from the US RLS Foundation. Part of this work was financed by the National Genome Research Network (NGFN). The KORA study group consists of H-E Wichmann (speaker), R Holle, J John, T Illig, C Metzinger, A Peters, and their coworkers, who are responsible for the design and conduction of the KORA studies. The KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. The collection of sociodemographic and clinical data in the Dortmund Health Study was supported by the German Migraine & Headache Society (DMHG) and by unrestricted grants of equal share from Astra Zeneica, Berlin Chemie, Boots Healthcare, Glaxo-Smith-Kline, McNeil Pharma (former Woelm Pharma), MSD Sharp & Dohme, and Pfizer to the University of Muenster. Blood collection in the Dortmund Health Study was done through funds from the Institute of Epidemiology and Social Medicine, University of Muenster. Data collection in the COR-Study was supported by unrestricted grants of the German RLS Society (Deutsche Restless Legs Vereinigung e.V) and Axonris Pharma, Boehringer Ingelheim Pharma, Mundipharma Research, Roche Pharma, and UCB to the University of Muenster. CG Bachmann was supported by grants of the German RLS Society. Deutsche Restless Legs Vereinigung, e.V. H Prokisch and T Metlinger were supported by the German Federal Ministry of Education and Research (BMBF) project Systems Biology of Metabotypes (SysMBio #0315494A). RP Allen and CJ Earley were supported by the grant PO1-A21190 National Institute of Health, USA; the Canadian part of the study was supported by a Canadian Institutes of Health Research (CIHR) grant to GA Rouleau and J Montplaisir. D Kemlnik and S Nevesimalova were supported by an ESRF grant MSM0021620849; Javonv and K Sonka were supported by grant MSM0021620816. Recruitment of Czech controls was funded by grant IGA NR 8563-5, Ministry of Health of the Czech Republic. B Frauscher, I Courto-Rebeix, M Francavilla, and C Fontenille are co-authors on behalf of BRC-RFGENSEP, which is supported by INSERM, AFM (Généthon), ARSEP, and GIS-IBISA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: winkelmann@lrz.tu-muenchen.de

These authors contributed equally to this work.

* E-mail: winkelmann@lrz.tu-muenchen.de

These authors contributed equally to this work.
forest plots (Figure 2). There was no evidence of epistasis between any of the six risk loci ($P_{\text{Bonferroni}} = 0.45$).

The association signal on 2p14 (rs6747972: nominal $\lambda$-corrected $P_{\text{JOINT}} = 9.03 \times 10^{-11}$, odds ratio (OR) = 1.23) is located in an LD block of 120 kb within an intergenic region 1.3 Mb downstream of MEIS1 (Figure 3). Assuming a long-range regulatory function of the SNP-containing region, in silico analysis for clusters of highly conserved non-coding elements using the ANCORA browser (http://ancora.genereg.net) identified MEIS1 as well as ETAA1 as potential target genes [5,6].

The second locus on chromosome 16q12.1 (rs3104767: nominal $\lambda$-corrected $P_{\text{JOINT}} = 9.4 \times 10^{-19}$, OR = 1.35) is located within an LD block of 140 kb (Figure 3), which contains the 5’UTR of TOX3 (synonyms TNRC9 and CAGF9) and the non-coding RNA BC034767 (synonym LOC643714). TOX3 is a member of the high mobility group family of non-histone chromatin proteins which interacts with CREB and CBP and plays a critical role in mediating calcium-dependent transcription in neurons [7]. GWAs have identified susceptibility variants for breast cancer in the identical region [8]. The best-associated breast cancer SNP, rs3803662, is in low LD ($r^2$, 0.1, HapMap CEU data) with rs3104767, but showed association to RLS ($\lambda$-corrected nominal $P_{\text{GWA}} = 7.29 \times 10^{-7}$). However, logistic regression analysis conditioned on rs3104767 demonstrated that this association is dependent on rs3104767 (rs3803662: $P_{\text{GWA/conditioned}} = 0.2883$).

BC034767 is represented in GenBank by two identical mRNA transcripts, BC034767 and BC029912. According to the gene model information of the UCSC and Ensembl genome browsers (http://genome.ucsc.edu and http://www.ensembl.org/index.html), these mRNAs are predicted to be non-coding. Additional in silico analysis using the Coding Potential Calculator (http://cpc.cbi.pku.edu.cn) supported this by attributing only a weak coding potential to this RNA, suggesting a regulatory function instead [9].

We also searched for rare alleles with strong effects and performed a mutation screening by sequencing all coding and non-coding...
exons of TOX3 and BC034767 in 188 German RLS cases (Table S4). In TOX3, a total of nine variants not listed in dbSNP (Build 130) were found, three of which are non-synonymous. Only one of these is also annotated in the 1000 Genomes project (November 2010 data release). Three additional new variants were located in putative exons 1 and 2 of BC034767. Analysis of the frequency of these variants as well as all known non-synonymous, frameshift, and splice-site coding SNPs in TOX3 in a subset of one of the replication samples (726 cases and 735 controls from the GER1 sample) did not reveal any association to RLS. For a power of >80%, however, variants with an OR above 4.5 and a MAF ≥0.01 would be required. For even lower MAFs, ORs ≥4.5 would be necessary for sufficient power. Furthermore, the described CAG repeat within exon 7 of TOX3 was not polymorphic as shown by fragment analysis in 100 population-based controls.

According to publicly available expression data (http://genome.ucsc.edu), in humans, BC034767 is expressed in the testes only, while TOX3 expression has been shown in the salivary glands, the trachea, and in the CNS. Detailed in-depth real time PCR profiling of TOX3 showed high expression levels in the frontal and occipital cortex, the cerebellum, and the retina [10]. To assess a putative eQTL function of rs6747972 or rs3104767, we studied the SNP-genotype-dependent expression of TOX3 and BC034767 as well as of genes known to directly interact with TOX3 (CREB-1/CREBBP/CITED1) and potential target genes of long-range regulatory elements at the locus on chromosome 2 (MEIS1/ETAA1) in RNA expression microarray data from peripheral blood in 323 general population controls [11]. No differential genotype-dependent expression variation was found.

To assess the potential for genetic risk prediction, we split our GWA sample in a training and a test set and determined classifiers for case-control status in the training set to predict case-control status in the test set. Training and test set were independent of each other—not only with respect to included individuals but also with respect to the genotyping procedure as we used genotypes generated on different genotyping platforms. As training set, we used those cases of the current GWA which had been genotyped on 500K arrays in a previous GWA and the corresponding control set [2], in total, 326 cases and 1,490 controls. The test set comprised 583 cases and 1,526 controls, genotyped on 5.0/6.0 arrays as part of the current study. Prior to the analysis, we removed the six known risk loci and performed LD-pruning to limit the analysis to SNPs not in LD with each other. In the end, a total of 76,332 SNPs were included in the pruned dataset. We conducted logistic regression with age and sex as covariates. Based on these association results, the sum score of SNPs showing the most significant effects (i.e. the number of risk alleles over all SNPs) weighted by the ln(OR) of these effects was chosen as predictor variable in the test set. We then varied the P-value threshold for SNPs included in the sum score. For a P-value <0.6, we observed a maximum area under the curve (AUC) of 63.9% and an explained genetic variance of 6.6% (Nagelkerke’s R), values comparable to estimates obtained for other complex diseases such as breast cancer or diabetes (Table S5) [12–14]. Inclusion of the six known risk loci in this analysis resulted in a maximum AUC of 64.2% and an explained genetic variance of 6.8%.

Additionally, we performed risk prediction in the combined GWA and replication sample including only the six established RLS risk loci. For this purpose, we used the weighted risk allele score resulting in ORs of up to 8.6 (95% CI: 2.46–46.25) and an AUC of 63.1% (Figures S3 and S4).

By increasing the size of our discovery sample, we have identified two new RLS susceptibility loci. The top six loci show effect sizes between 1.22 and 1.77 and risk allele frequencies between 19 and 82%, and reveal genes in neuronal transcription pathways not previously suspected to be involved in the disorder.

Materials and Methods

Study population and phenotype assessment

Ethics statement. Written informed consent was obtained from each participant in the respective language. The study has

Table 1. Association results of GWA and joint analysis of GWA and replication.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Locus</th>
<th>LD block (Mb)</th>
<th>SNP</th>
<th>Position (bp)</th>
<th>Risk allele</th>
<th>Risk allele frequency cases/controls</th>
<th>P_{GWA}</th>
<th>P_{REPLICATION}</th>
<th>P_{JOINT}</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>MEIS1</td>
<td>66.57–66.64</td>
<td>rs2300478</td>
<td>66634957</td>
<td>G</td>
<td>0.35/0.24</td>
<td>4.39×10^{-16}</td>
<td>3.40×10^{-9}</td>
<td>6.80 (1.57–181)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BTBD9</td>
<td>37.82–38.79</td>
<td>rs9357271</td>
<td>38473851</td>
<td>T</td>
<td>0.82/0.76</td>
<td>2.01×10^{-16}</td>
<td>7.75×10^{-12}</td>
<td>1.47 (1.35–1.47)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PTPRD</td>
<td>8.80–8.88</td>
<td>rs1767710</td>
<td>8836955</td>
<td>T</td>
<td>0.19/0.16</td>
<td>1.07×10^{-6}</td>
<td>3.49×10^{-10}</td>
<td>1.29 (1.19–1.40)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>MAP2K5/ SKOR1</td>
<td>65.25–65.94</td>
<td>rs12593813</td>
<td>65823906</td>
<td>G</td>
<td>0.75/0.68</td>
<td>1.49×10^{-6}</td>
<td>1.54×10^{-17}</td>
<td>1.37×10^{-22}</td>
<td>1.41 (1.32–1.52)</td>
</tr>
</tbody>
</table>

New genome-wide significant loci (P_{JOINT} <5.2×10^{-8})

<table>
<thead>
<tr>
<th>Chr</th>
<th>Locus</th>
<th>LD block (Mb)</th>
<th>SNP</th>
<th>Position (bp)</th>
<th>Risk allele</th>
<th>Risk allele frequency cases/controls</th>
<th>P_{GWA}</th>
<th>P_{REPLICATION}</th>
<th>P_{JOINT}</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Intergenic region</td>
<td>67.68–68.00</td>
<td>rs6747972</td>
<td>67923729</td>
<td>A</td>
<td>0.47/0.44</td>
<td>1.37×10^{-6}</td>
<td>3.73×10^{-6}</td>
<td>9.03×10^{-11}</td>
<td>1.23 (1.16–1.31)</td>
</tr>
<tr>
<td>16</td>
<td>TOX3/ BC034767</td>
<td>51.07–51.21</td>
<td>rs3104786</td>
<td>51182239</td>
<td>G</td>
<td>0.65/0.58</td>
<td>7.38×10^{-7}</td>
<td>2.16×10^{-13}</td>
<td>9.40×10^{-19}</td>
<td>1.35 (1.27–1.43)</td>
</tr>
</tbody>
</table>

RLS-associated SNPs with genome-wide significance. P_{GWA}, λ-corrected nominal P-value of GWA stage. P_{REPLICATION}, nominal P-value obtained from meta-analysis of the replication stage samples. P_{JOINT}, nominal P-value of the joint meta-analysis of GWA and replication stage, λ-corrected in samples where λ-values were available. Nominal P-values in GWA were calculated using logistic regression with sex, age, and the first four components from the MDS analysis of the IBS matrix as covariates. For nominal P-values, a fixed-effects inverse-variance meta-analysis was performed. Risk allele frequencies and odds ratios were calculated in the joint sample. LD blocks were defined by D’ using Haploview 4.2 based on HapMap CEU population data from HapMap release r27. CI, 95% confidence interval. Genome positions refer to the Human March 2006 (hg18) assembly. doi:10.1371/journal.pgen.1002171.001
been approved by the institutional review boards of the contributing authors. The primary review board was located in Munich, Bayerische Arztekammer und Technische Universität München.

RLS patients (GWA and replication phase). A total of 2,944 cases (GWA = 954, replication = 1,990) of European descent were recruited in two cycles via specialized outpatient clinics for RLS. German and Austrian cases for the GWA (GWA) and the replication sample (GER1) were recruited in Munich, Marburg, Kassel, Göttingen, Berlin (Germany, n in GWA = 830, n in GER1 = 1,028), Vienna, and Innsbruck (Austria, n in GWA = 124, n in GER1 = 288). The additional replication samples originated from Prague (Czech Republic (CZ), n = 351), Montpellier (France (FR), n = 182), and Turku (Finland (FIN), n = 141). In all patients, diagnosis was based on the diagnostic criteria of the International RLS Study Group [1] as assessed in a personal interview conducted by an RLS expert. A positive family history was based on the report of at least one additional family member affected by RLS. We excluded patients with secondary RLS due to uremia, dialysis, or anemia due to iron deficiency. The presence of secondary RLS was determined by clinical interview, physical and neurological examination, blood chemistry, and nerve conduction studies whenever deemed clinically necessary.

In addition, 1,104 participants (GER2) of the “Course of RLS (COR-) Study”, a prospective cohort study on the natural course of disease in members of the German RLS patient organizations, were included as an additional replication sample. After providing informed consent, study participants sent their blood for DNA extraction to the Institute of Human Genetics, Munich, Germany. A limited validation of the RLS diagnosis among the majority of members was achieved through a diagnostic questionnaire. Five percent had also received a standardized physical examination and interview in one of the specialized RLS centers in Germany prior to recruitment. To avoid doublets, we checked these subjects against those recruited through other German RLS centers and excluded samples with identical birth date and sex.

556 cases (US) were recruited in the United States at Departments of Neurology at Universities in Baltimore, Miami, Houston, and Palo Alto. Diagnosis of RLS was made as mentioned above.

285 cases (CA) were recruited and diagnosed as above in Montréal, Canada. All subjects were exclusively of French-Canadian ancestry as defined by having four grandparents of French-Canadian origin.

Detailed demographic data of all samples are provided in Table S6.

Control populations (GWA and replication phase). Controls for German and Austrian cases were of European descent and recruited from the KORA S3/F3 and S4 surveys, general population-based controls from southern Germany. KORA procedures and samples have been described [15]. For the GWA phase, we included 1,814 subjects from S3/F3, and, for the replication stage, 1,471 subjects from S4.

For replication of the GER2 sample, we used controls from the Dortmund Health Study (DHS), a population-based survey conducted in the city of Dortmund with the aim of determining the prevalence of chronic diseases and their risk factors in the general population. Sampling for the study was done randomly from the city’s population register stratified by five-year age group and gender [16]. 597 subjects selected at random from the Czech blood and bone marrow donor registry served as Czech controls [17]. French controls included 768 parents of multiple sclerosis patients recruited from the French Group of Multiple Sclerosis Genetics Study [REFGENSEP] [18]. Finnish controls comprised 360 participants of the National FINRISK Study, a cross-sectional population survey on coronary risk factors collected every five years. The current study contains individuals recruited in 2002. Detailed description of the FINRISK cohorts can be found at www.nationalbiobanks.fi.

French-Canadian controls were 285 unrelated individuals recruited at the same hospital as the cases.

1,200 participants of the Wisconsin Sleep Cohort (WSC), an ongoing longitudinal study on the causes, consequences, and natural course of disease of sleep disorders, functioned as US controls [19].

None of the controls were genotyped for RLS. All studies were approved by the institutional review boards in Germany, Austria, Czech Republic, France, Finland, the US, and Canada. Written informed consent was obtained from each participant. Detailed demographic data of all samples are provided in Table S6.

Genotyping

GWA. Genotyping was performed on Affymetrix Genome-Wide Human SNP Arrays 5.0 (cases) and 6.0 (controls) following the manufacturer’s protocol. The case sample included 628 cases from previous GWAs [2,4] and 326 new cases. After genotype-calling using the BRLMM-P clustering algorithm [20], a total of 475,976 overlapping SNPs on both Affymetrix arrays were subjected to quality control. We added 655 cases of a different phenotype unrelated to RLS, genotyped on 5.0 arrays, to the analysis and excluded those SNPs which showed a significant difference of allele frequencies in cases (RLS and unrelated phenotype on 5.0) and controls (6.0) (n = 92). Thereby, we filtered out SNPs likely to be false-positive associations. We excluded SNPs with a minor allele frequency (MAF) < 5% (n = 88,582), a call rate < 98% (n = 65,906) or a significant deviation from Hardy-Weinberg Equilibrium (HWE) in controls (P < 0.00001) (n = 20,060). Cluster plots of the GWA genotyping data for the best-associated SNPs in Table 1 are shown in Figure S5. Genotypes of these SNPs are available in Table S7.

Replication. We selected all SNPs with a χ²-corrected Pnominal<10⁻⁴ in the GWA for replication. These SNPs clustered in 26 loci (defined as the best associated SNP ± 150 kb of flanking sequence). We genotyped a total of three SNPs in each of the 26 regions. These were either further associated neighbouring SNPs with a χ²-corrected Pnominal<10⁻³ or, in case of singleton SNPs, additional neighbouring SNPs from HapMap with the highest possible r² (at least >0.5) with the best-associated SNP. We also genotyped the best-associated SNPs identified in the previous GWAs [2,4].

Genotyping was performed on the MassARRAY system using MALDI-TOF mass spectrometry with the iPLEX Gold chemistry (Sequenom Inc, San Diego, CA, USA). Primers were designed using AssayDesign 3.1.2.2 with iPLEX Gold default parameters. Automated genotype calling was done with SpectroTYPEr 3.4. Genotype clustering was visually checked by an experienced evaluator.
Figure 3. New genome-wide significant RLS loci. a) Risk locus on chromosome 2p14, showing the best-associated SNP rs6747972 and ±200 kb of surrounding sequence. b) Risk locus on chromosome 16p21, showing the best-associated SNP rs3104767 and ±200 kb of surrounding sequence. The left-hand x-axis shows the negative log10 of the nominal $\lambda$-corrected $P$-values of the GWA stage for all SNPs genotyped in the respective region. The right-hand x-axis shows the recombination frequency in cM/Mb. The y-axis shows the genomic position in Mb based on the hg18 assembly. The $r^2$-based LD between SNPs is colour-coded, ranging from red ($r^2>0.8$) to dark blue ($r^2<0.2$) and uses the best-associated SNP as reference. This SNP is depicted as a violet diamond. Recombination frequency and $r^2$ values are calculated from the HapMap II (release 22) CEU population. Plots were generated with LocusZoom 1.1 (http://csg.sph.umich.edu/locuszoom/).

doi:10.1371/journal.pgen.1002171.g003
SNPs with a call rate <95%, MAF <5%, and P < 0.00001 for deviations from HWE in controls were excluded. DNA samples with a call rate <90% were also excluded.

Population stratification analysis

**GWA.** To identify and correct for population stratification, we performed an MDS analysis as implemented in PLINK 1.07 (http://pngu.mgh.harvard.edu/~purcell/plink, [21]) on the IBS matrix of our discovery sample. After excluding outliers by plotting the main axes of variation against each other, we performed logistic regression with age, sex, and the values of the MDS components as covariates. Using the Genomic Control approach [22], we obtained an inflation factor λ of 1.11.

Additionally, we performed a variance components analysis using the EMMAX software (http://genetics.cs.ucla.edu/emmax, [23]) and, again, calculated the inflation factor with Genomic Control, now resulting in a λ of 1.025. EMMAX uses a mixed linear model and does not only correct for population stratification but also for hidden relatedness. We, therefore, decided to base correction for population substructure on the EMMAX results.

**Replication.** Correction for population stratification was performed for the German, Czech, and the Canadian subsamples. The λ-values of 1.1032, 1.2286, and 1.2637 were derived from a previous Genomic Control experiment within the same samples using 176 intergenic or intronic SNPs [4]. Here, we had applied the expanded Genomic Control method GCF developed by Devlin and Roeder [24]. In the meta-analysis of all replication samples, the λ-corrected standard errors were included for the German, Czech, and Canadian samples. For the other replication samples from France, Finland, and the USA, no such data was available and, therefore, no correction factor was included in the analysis.

Statistical analysis

Statistical analysis was performed using PLINK 1.07 (http://pngu.mgh.harvard.edu/~purcell/plink, [21]). In the GWA sample, we applied logistic regression with age, sex, and the first four axes of variation resulting from an MDS analysis as covariates.

P-values were λ-corrected with the λ of 1.025 from the EMMAX analysis. In the individual analysis of the single replication samples, we tested for association using logistic regression and correcting for gender and age as well as for population stratification where possible (see Population Stratification). Each replication sample was Bonferroni-corrected using the number of SNPs which passed quality control for the respective sample.

For the combined analysis of all replication samples, we performed a fixed-effects inverse-variance meta-analysis. Where available, we used λ-corrected standard errors in this analysis. Bonferroni-correction was performed for 74 SNPs, i.e. the number of SNPs which passed quality control in at least one replication sample.

For the joint analysis of the GWA and the replication samples, we also used a fixed-effects inverse-variance meta-analysis and again included λ-corrected values as far as possible. For the conditioned analysis, the SNP to be conditioned on was included as an additional covariate in the logistic regression analysis as implemented in PLINK.

Interaction analysis was performed using the epistasis option in PLINK. Significance was determined via Bonferroni-correction (i.e. 0.05/28, as 28 SNP combinations were tested for interaction).

Power calculation

Power calculation was performed using the GaTS power calculator [25] using a prevalence set of 0.08 and an additive genetic model (Table S3). The significance level was set at 0.05/74 for replication stage analysis and at 0.05/301,406 for genome-wide significance in the joint analysis of GWA and replication. For the rare variants association study, the significance level was set at 0.05/12.

Mutation screening of TOX3 and BC034767

All coding and non-coding exons including adjacent splice sites of TOX3 (reference sequence NM_001146188) and BC034767 (reference sequence IMAGE 5172237) were screened for mutations in 186 German RLS cases.

Mutation screening was performed with high resolution melting curve analysis using the LightScanner technology and standard protocols (IDADO Technology Inc.). DNAs were analyzed in doublets. Samples with aberrant melting pattern were sequenced using BigDye Terminator chemistry 3.1 [ABI] on an ABI 3730 sequencer. Sequence analysis was performed with the Staden package [26]. Primers were designed using Exon Primer (http://ihg.gsf.de) or Primer3plus (www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). All identified variants were then genotyped in 735 RLS cases and 735 controls of the general population (KORA cohort) on the MassARRAY system, as described above.

In addition, fragment analysis of exon 7 of TOX3 was performed to screen for polymorphic CAG trinucleotide repeats. DNA of 100 controls (50 females, 50 males) was pooled and analyzed on an ABI 3730 sequencer with LIZ-500 [ABI] as a standard. Primers were designed using Primer3plus, the forward Primer contains FAM for detection. Analysis was performed using GeneMapper v.3.5.

Expression analyses

Associations between MEIS1/ETA11 RNA expression and rs6747972 and between TOX3/BC034767/CREB-1/CREBBP/CITED1 expression and rs3104767 were assessed using genome-wide SNP data [Affymetrix 6.0 chip] in conjunction with microarray data for human blood samples (n = 323) general population controls from the KORA cohort, Illumina Human WG6 v2 Expression BeadChip [11]. A linear regression model conditioned on expression and controlling for age and sex was used to test for association.

Prediction of genetic risk

**Based on the performance of P-value-threshold selected SNPs in a training and a test sample.** As training sample, we used those GWA-cases which had also been genotyped for our previous study [2]. We also included the control samples from this study. As a first quality control step, we carried out an association analysis comparing the Affymetrix 500K genotypes of these GWA-cases to the Affymetrix 5.0 genotypes of the same cases. Significant P-values would indicate systematic differences in the genotyping between the different chips. For further analysis, we only used those 259,302 SNPs with P-values >0.10. We performed a second quality control step in which IDs with a callrate below 98% and SNPs with a callrate below 98%, a MAF lower than 5%, or a P-value for deviation from HWE <0.00001 were removed.

Further, we excluded the four already known risk loci as well as the two newly identified loci and performed LD-pruning to limit the analysis to SNPs not in LD with each other. This was performed using a window-size of 50 SNPs. In each step, this window was shifted 5 SNPs. We used a threshold of 2 for the VIF (variance inflation factor). 76,532 SNPs, 326 cases, and 1,498 controls were included in the final training dataset. We conducted logistic regression with age and sex as covariates. Based on these association results, the sum score of SNPs showing the most
significant effects (i.e. the number of risk alleles over all SNPs) weighted by the ln(OR) of these effects was chosen as predictor variable in the test set, comprising the remaining 583 cases of the GWA sample and 1,526 controls. None of these cases/controls were included in the training-sample, i.e. the test-sample constitutes a completely independent sample. Based on this sum score, we calculated the ROC curve and Nagelkerke’s R to measure the explained variance.

**Based on a weighted risk allele score.** To evaluate the predictive value in our sample, we calculated a weighted sum score of risk alleles in the combined GWA and replication sample. To this end, we used one SNP from each RLS risk region and also included markers from the two newly identified regions on chromosome 16q12 and 2p14 (included markers from the two newly identified regions on chromosome 16q12 and 2p14 (described in: Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and Application of European Genetic Substructure Using 300 K SNP Information. PLoS Genet 4: e4. doi:10.1371/journal.pgen.0040004). For the best-associated SNPs at each risk locus, clusterplots were generated for cases and controls. Intensities of the A and B allele (based on the Affymetrix annotation of the SNPs) are given on the x- and y-axes and the respective genotypes are indicated in blue, green, and orange. (PDF)

### Supporting Information

#### Figure S1
MDS analysis plot for GWA. Distribution of cases (red) and controls (black) along the two main axes of variation identified in the MDS analysis. The three visible clouds are due to a common 3.8 Mb inversion polymorphism on chromosome 8 (described in: Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and Application of European Genetic Substructure Using 300 K SNP Information. PLoS Genet 4: e4. doi:10.1371/journal.pgen.0040004). (TIFF)

#### Figure S2
QQ-plot of GWA results. QQ-plot showing the P-value distribution before (red) and after (blue) correction for population stratification using Genomic Control. (TIFF)

#### Figure S3
Weighted risk allele score analysis. Histogram of the weighted risk allele scores for cases and controls. The corresponding OR and CI for each category against the median category is depicted in green. The left y-axis refers to the number of individuals (in %), the right-axis refers to the OR values. (TIFF)

#### Figure S4
ROC curve for weighted risk score analysis. Receiver operating characteristic (ROC) curve for the weighted risk allele score approach of risk prediction. The area under the curve (AUC) is 65.1%. (TIFF)

#### Figure S5
Cluster plots of GWA genotyping for the six risk loci. For the best-associated SNPs at each risk locus, clusterplots were generated for cases and controls. Intensities of the A and B allele (based on the Affymetrix annotation of the SNPs) are given on the x- and y-axes and the respective genotypes are indicated in blue, green, and orange. (PDF)

#### Table S1
GWA results for SNPs with $\lambda$-corrected $P_{\text{GWA}}<10^{-4}$ and additional SNPs selected for replication. A star (*) indicates SNPs which had been identified in previous RLS GWAs [2–4]. P-values of the GWA phase are given as $\lambda$-corrected nominal P-values. Two different methods for $\lambda$ correction were applied, multi-dimensional-scaling (MDS)-analysis using PLINK and variance components (VC)-analysis using the EMMAX software with the P-values listed in the respective columns “MDS $\lambda$-corrected $P_{\text{GWA}}$” and “VC $\lambda$-corrected $P_{\text{GWA}}$”. The selection of SNPs for replication was based on the MDS $\lambda$-corrected P-values. $r^2$-values based on Hapmap CEU data are given for those SNPs which were selected for replication based on their LD with the best-associated SNP in each region. Genomic position and gene annotation refer to the hg18 genome. (DOC)

#### Table S2
Replication stage association results for individual replication samples. P-values are derived from logistic regression and correcting for gender and age as well as for population stratification where possible (see Materials and Methods). Each replication sample was Bonferroni-corrected using the number of SNPs which passed quality control for the respective sample. The OR refers to the minor allele. NA: SNP could not be analysed due to failing quality control in the respective sample. (DOC)

#### Table S3
Power analysis for GWA, replication and joint analysis of GWA and replication. Power calculation was performed using the CaTS power calculator [25] using a prevalence set of 0.08 and an additive genetic model. The significance level $\alpha$ was set at 0.05/74 for replication stage analysis and at 0.05/301,406 for genome-wide significance in the joint analysis of GWA and replication. (DOC)

#### Table S4
Results of TOX3 and BC034767 mutation screening. * “A” refers to the mutant allele, “B” to the reference allele. Position refers to hg18 genome annotation. Codon numbering refers to the reference sequence NM_001146188. Data of the 1000 genomes project was obtained from the November 2010 release via the 1000 genomes browser (http://browser.1000genomes.org/index.html). (DOC)

#### Table S5
Prediction of genetic risk; training- and test-set approach. Inclusion threshold P-values were derived from a logistic regression with age and sex as covariates in the training sample. # SNPs indicates the number of SNPs passing the inclusion threshold. Based on these association results, the sum score of SNPs showing the most significant effects (i.e. the number of risk alleles over all SNPs) weighted by the ln(OR) of these effects was chosen as predictor variable in the test set. Based on this sum score, an AUC and Nagelkerke’s R were calculated. (DOC)

#### Table S6
Demographic data of GWA and replication samples. Mean age, mean age of onset and respective standard deviations and ranges are given in years. N: number of individuals; SD: standard deviation; AAO: age of onset. GWA: Genome-wide association study; CZ: Czechia; FR: France; FIN: Finland; CA: Canada; US: United States. * indicates that this information is not applicable for the respective sample. (DOC)

#### Table S7
Genotype data of GWA samples. Genotypes of the GWA samples are given for the eight best-associated SNPs (see Table 1). SNP alleles are ACGT-coded. Phenotype information includes gender (1 = male, 2 = female) and disease status (1 = unaffected, 2 = affected). (XLS)

### Acknowledgments
We are grateful to all patients who participated in this study. We thank Jelena Golic, Regina Feldmann, Sibylle Frischholz, Susanne Lindhof,
Katja Junghans, Milena Radivojkov-Blagojevic, and Bianca Schmick for excellent technical assistance.

Author Contributions


References