Identification of Restless Legs Syndrome Genes by Mutational Load Analysis

Running head: RLS genes by mutational load

Erik Tilch PhD*[,1], Barbara Schormair PhD*[,1], Chen Zhao PhD[1], Aaro V. Salminen PhD[1], Ana Antic Nikolic MD[1], Evi Holzknecht MD[2], Prof Birgit Högl MD[2], Prof Werner Poewe MD[2], Cornelius G. Bachmann MD[3], Prof Walter Paulus MD[4], Prof Claudia Trenkwalder MD[5,6], Prof Wolfgang H. Oertel MD[1], Prof Magdolna Hornyak MD[7], Prof Ingo Fietze MD[8], Prof Klaus Berger MD[9], Peter Lichtner PhD[10], Christian Gieger PhD[11], Prof Annette Peters PhD[11], Prof Bertram Müller-Myhsok MD[12,13,14], Prof Alexander Hoischen PhD[15], Prof Juliane Winkelmann MD*[,1,12,16], Prof Konrad Oexle MD*[,1]

*These authors contributed equally.

1. Institute of Neurogenomics, Helmholtz Center München, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany
2. Department of Neurology, Medical University of Innsbruck, Anichstraße 35, 6020 Innsbruck, Austria
3. Department of Neurology, Paracelsus Klinik, Osnabrueck, Germany
4. Department of Clinical Neurophysiology, University Medical Centre, Georg August University Göttingen, Göttingen, German
5. Clinic for Neurosurgery, University Medical Centre, Georg August University Göttingen, Göttingen, Germany
6. Paracelsus-Elena Hospital, Centre of Parkinsonism and Movement Disorders, Kassel, Germany
7. Neuropsychiatry Centre Erding/München, Erding, Germany

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ana.25658

8. Department of Cardiology and Angiology, Centre of Sleep Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany
9. Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany
10. Institute of Human Genetics, Helmholtz Centre München, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany
11. Institute of Epidemiology II, Helmholtz Centre München, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany
12. Munich Cluster for Systems Neurology (SyNergy), Feodor-Lynen-Str. 17, 81377 Munich, Germany
13. Max Planck Institute of Psychiatry, Kraepelinstr. 2-10, 80804 Munich, Germany
14. Institute of Translational Medicine, University of Liverpool, Liverpool, UK
15. Department of Human Genetics, Radboud University Medical Centre, Nijmegen, The Netherlands
16. Chair Neurogenetics and Institute of Human Genetics, Technische Universität München, Trogerstr. 32, 81675 Munich, Germany

Corresponding authors:
Prof Juliane Winkelmann MD
Chair Neurogenetics, Technical University of Munich, Trogerstr. 32, D-81675 München
Director, Institute of Neurogenomics, Helmholtz Zentrum München - German Research Center for Environmental Health, Ingolstaedter Landstr. 1, D-85764 Neuherberg
Phone: +49 89 3187-1883, Fax: -3297, Email: juliane.winkelmann@tum.de

Prof Konrad Oexle MD
Institute of Neurogenomics, Group Leader, Neurogenetic Systems Analysis, Helmholtz Zentrum München - German Research Center for Environmental Health, Ingolstaedter Landstr. 1, D-85764 Neuherberg
Phone: +49 89 3187-1882, Fax: -3297, Email: konrad.oexle@helmholtz-muenchen.de

Number of characters in the title: 74
Number of characters in the running head: 28

Number of words in the Abstract: 250
Number of words in the Introduction: 495
Number of words in the Discussion: 1,249
Number of words in the body of the manuscript: 4,500
Number of figures: 1
Number of color figures: 0
Number of tables: 3

Abstract

Objective

Restless legs syndrome is a frequent neurological disorder with substantial burden on individual wellbeing and public health. Genetic risk loci have been identified but the causatives genes at these loci are largely unknown, so that functional investigation and clinical translation of molecular research data are still inhibited. To identify putatively causative genes we searched for highly significant mutational burden in candidate genes.

Methods

We analyzed 84 candidate genes in 4,649 patients and 4,982 controls by next generation sequencing using molecular inversion probes that targeted mainly coding regions. The burden of low-frequency and rare variants was assessed, and in addition, an algorithm (binomial performance deviation analysis) was established to estimate independently the sequence variation in the probe binding regions from the variation in sequencing depth.

Results

Highly significant results (considering the number of genes in the genome) of the conventional burden test and the binomial performance deviation analysis overlapped significantly. Fourteen genes were highly significant by one method and confirmed with nominal significance by the other to show a differential burden of low-frequency and rare variants in restless legs syndrome. Nine of them (*AAGAB*, *ATP2C1*, *CNTN4*, *COL6A6*, *CRBN*, *GLO1*, *NTNG1*, *STEAP4*, *VAV3*) resided in the vicinity of known restless legs syndrome loci while five (*BBS7*, *CADM1*, *CREB5*, *NRG3*, *SUN1*) have not previously been associated with restless legs syndrome. Burden test and binomial performance deviation

analysis also converged significantly in fine-mapping potentially causative domains within these genes.

Interpretation

Differential burden with intragenic low-frequency variants reveals putatively causative genes in restless legs syndrome.

Introduction

The contribution of low frequency variants to the genetic architecture of common diseases has gained considerable attention recently[1-3]. However, the required sample size in association analysis of single rare variants is inversely proportional to their allele frequency, necessitating strategies to overcome this power problem[1, 4-6]. Increasing the sample size while minimizing genotyping/sequencing costs, is one such strategy. Class-wise analysis of variants as in gene-wise burden testing, is another.

We applied both strategies when investigating the contribution of rare (minor allele frequency MAF $\leq$ 1%) and low frequency variation (1% < MAF $\leq$ 5%) to restless legs syndrome (RLS), one of the most prevalent neurological disorders in Western countries. RLS affects 5-10% of the population, presenting with unpleasant symptoms in the legs, an urge to move at rest predominantly in the evening, and sleep disturbances[7, 8]. Previous genome-wide association studies of common variants revealed 19 RLS loci[9-11]. For the detection of low frequency variants in a set of 84 candidate genes we now applied tiling molecular inversion probes (MIPs) which allow for targeted enrichment and re-sequencing of large sample sizes[12].

MIP technology is based on two linked probes that bind to the same DNA target strand. Upon binding, the MIP is circularized enzymatically. Only then the captured DNA between the probes can be sequenced. To do so, next generation sequencing (NGS) adapters and multiplex tags are attached to the captured DNA by PCR. Sequencing reads are mapped to a reference so that polymorphisms can be detected. This and similar MIPs methods are widely used for resequencing of various traits [13-15].

The standard MIP analysis focuses on variation found in the captured sequences. However, the MIP probes' hybridization binding capacities are influenced by genetic variation within the probe binding sequences. While high reading coverage and double tiling of the MIPs may largely neutralize the potential impact of such variation on the sequencing results, which is

desirable especially in diagnostic applications on a limited number of genes[16], it is also possible to regard the deviation of the yield of MIP sequencing reads as independent information on the genetic variation within the MIP probe binding sequences. Devoting our limited resources to the scientific analysis of a considerable number of genes in a large number of patients and controls, we followed the latter option and developed a tool called binomial performance deviation analysis (BIPEDAL) which assessed the differential variant burden directly from the difference in MIP read number. Incidentally, BIPEDAL is sensitive to copy number variations that may escape sequencing. BIPEDAL results were compared to standard burden of rare variant testing (BRVT)[17] of the MIP target sequences. Examining 11,214 MIPs at 84 RLS candidate genes we detected a large and significant overlap in the results of these two independent methods. The significant overlap of BRVT and BIPEDAL also holds in fine-mapping of intragenic domains. Beyond the mere calling of variants, enrichment-based NGS technologies provide quantitative hybridization data that can thus be made useful for the genetic study of complex traits.

Methods

Targeted sequencing, variant calling and quality control

We designed molecular inversion probes (MIPs)[12, 18] for extended exons (extending by 20 bp over the exon borders) and promotors (500 bp upstream from any transcription start site) of 84 putative RLS genes. 65 genes were selected from loci detected in a previous GWAS[11] using eQTL and additional functional annotations as the major selection criteria. The remaining 19 genes were the top genes from a case-control Illumina ExomeChip study, where no gene had reached significance after multiple testing correction of the burden test[19]. These 19 genes were included in the MIPseq analysis because it covered more variants than the ExomeChip analysis and thus was potentially providing the burden test with sufficient power. The 84 candidate genes and their selection criteria are provided in Supplementary Table 1. MIP binding sites were selected such as to avoid variants with minor allele frequency (MAF) greater 1% listed in dbSNP Build 141. Alternatively, degenerated MIPs were used so as to neutralize the effect of the common single nucleotide polymorphisms (SNPs). MIPs were pooled and calibrated as described previously[12, 15, 16, 19]. We paired-end sequenced in batches à 186 samples on 6 Illumina HiSeq 4000 lanes[19]. Together, we sequenced 4,649 RLS cases collected from Germany/Austria and 4,982 controls from the

region of Augsburg (KORA) in the South-East of Germany[20] (Table 1).. The study was approved by the respective IRB and participants provided informed written consent. RLS cases have been diagnosed by experienced neurologists based on the IRLSSG criteria[21].

Paired-end reads were merged using BBMerge[22], trimmed at head/tail for the sequencing adapters, and mapped to the reference genome (GRCh37/hg19, hs37d5, downloaded from https://software.broadinstitute.org/gatk/download/bundle) using Bowtie2[23]. We filtered mapped reads (mapping uniquely to target region, 108 < length < 113, Mapq > 30), counted the read numbers as input for BIPEDAL (see below), and called variants using GATK[24] v3.5 (in target region, Average base quality for reads supporting alleles > 30, Root Mean Square of the mapping quality of reads across all samples> 30, primary alignments only, insertion-deletion depth > 50, insertion-deletion fraction > 0.1, contamination = 0.01) as input for BRVT. We normalized variants using Bcftools[25] v1.2 and removed those of low quality (genotype quality < 50, coverage depth < 10, and also SNPs with Root Mean Square of the mapping quality of reads across all samples ≤ 30, call rate < 0.5) using Vcftools[26] v0.1.12b. Using Plink[27, 28] (v1.07, v2.00aLM), we further removed individuals with low call rate (< 0.5) and excess heterozygosity ± 5 standard deviations (SD). We filtered variants on Hardy-Weinberg equilibrium (HWE) in controls ($p < 0.00001$). Individuals were excluded if data on age or sex, or data on common SNP as assessed in our latest GWAS[11] were missing. Moreover, we excluded duplicated individuals, related individuals ($\hat{\pi} > 0.09375$), and population outliers based on 10 principal components (PCs) and ± 6 SD (derived from common SNP array data[11]). We filtered variants on minor allele frequency ($0 < MAF \leq 0.05$ in either cases or controls). For subsequent analysis we used R[29] v3.0.2.

Burden analysis (BRVT)

We grouped variants by gene and ran a burden of rare variant test (BRVT), that is, a modified Morris-Zeggini test as described by Auer et al. 2013 including a correction for differential missing genotypes and covariates age, sex, PC1, PC2, batch, and sub-batch. The number of PCs was determined based on a Scree plot analysis. Multiple testing was corrected on a stringent scale (0.05/25,000, corresponding to the number of genes in the genome) in line with the previous ExomeChip analysis, whose results were used for candidate gene selection.

BIPEDAL

To test for an association between disease and variant burden as revealed by target sequencing depth, we developed BIPEDAL (binomial performance deviation analysis). Data processing by BIPEDAL can be subdivided into a first step which identifies MIP targets that show Bonferroni significant difference in sequencing depth between cases and controls and comprises a calibration, and a second step which determines a gene-wise or segment-wise burden with these disease-associated MIPs.

In the first step, the sequencing depth of each MIP target is compared between cases and controls. Therefore, for n MIP targets and m individuals, consider an n x m matrix **H** of MIP target depths. To normalize by total read number over all MIPs in an individual, we transform **H** to **Z**, so that $z_{i,j} = h_{i,j}/(h_{1,j} + h_{2,j} + \ldots + h_{n,j})$. The $z_{i,j}$ are limited to the interval [0,1]. Therefore, their logit $k_{i,j} = \log(z_{i,j}/(1 - z_{i,j}))$ is used in the basic regression model of BIPEDAL

(equation 1) $k_{i,j} \sim g_{i,j} + f_{i,j} + d_j + b_i$,

where $g_{i,j}$ and $f_{i,j}$ measure genetic variation in the probe and target regions of MIP target i in individual j, respectively, d_j quantifies individual bias, and b_i is the baseline for MIP target i. The individual bias covariate is needed because the read depths of the MIPs of an individual are not independent of each other. They share the same chemical resources during the library preparation and they become interrelated via the normalization procedure (see above). Thus, for instance, if one particular MIP cannot bind due of a CNV, the read depths of all other MIPs in that individual will increase, especially if the drop-out affects a MIP with a high baseline.

Baseline parameter b_i may be estimated from a submatrix **K*** of **K**, comprising c controls. Vector $\mathbf{b} = c^{-1} \cdot \mathbf{K^*} \cdot \mathbf{J}_{c,1}$ contains the baselines of all MIP targets. The baseline is only needed for the next step of the calibration, the estimation of d_j. This is done by regression on the subset of MIP targets for which $g_{i,j}$ and $f_{i,j}$ are available in all individuals, $k_{.j} \sim \mathbf{b} + g_{.j} + f_{.j} + d_j$. The intercept d_j captures the individual bias.

The genetic variation of individual j in MIP target i may be correlated with disease probability a_j, so that $g_{i,j} + f_{i,j} \sim \text{logit}(a_j)$, which is the model for the burden of rare variant test. Hence, we rearrange equation 1 to

(equation 2) logit(**a**) ~ $k_{i.}$ + **d** + b_i

which models the association between disease probability and the calibrated sequencing depth of each MIP. The regression parameters β_i of the k_i are determined by logistic regression and subsequently tested for significant association using the Wald test. Individual covariates such as age and sex may be added to the equation, while b_i may be dropped as it would be captured by the intercept.

In the worst case, there may be an overall bias between cases and controls, causing an imbalance of the effect directions sign(β_i). For quantification and eventual correction of this bias, we estimate the probability $p_{\text{bin}} = P(\beta_i < 0)$ of randomly drawing a MIP target with lower depth in cases under the null hypothesis of no association between MIP target depth and disease status. As an estimator of p_{bin} we choose the proportion of betas below zero among all those betas with significance above the nominal threshold (0.05), thus excluding the true associations from the estimation.

The second step of BIPEDAL addresses all those MIP targets that showed significant association (Bonferroni-corrected for the number of MIPs) to the disease. We group them by gene and, for each gene, perform a one-sided binomial test to see whether there are significantly more MIPs with negative betas in that gene than expected, yielding

(equation 3) $p_{BIPEDAL} = \sum_{s=r}^{n} B(s|p_{bin},n)$,

where p_{bin}, as indicated above, is the probability for negative beta under the null hypothesis, and n and r are the numbers of all MIPs and all negative MIPs, respectively, within the examined gene. As such, BIPDEAL is alike a burden test, operating in the spirit of a meta-analysis, but allowing for correction for residual bias by setting the probability of success to p_{bin}.

Application of BIPEDAL

In our data, we considered only MIPs loci with at least one read in the whole experiment and only individuals with at least 200,000 reads of all MIPs combined. We calibrated BIPEDAL based on variants and for individuals from a more stringent QC (i.e., QC as before, but heterozygosity filter at ± 3 SD, HWE filter at $p < 0.0001$, population outliers from 10 PCs

± 4 SD). We added age, sex, batch, sub-batch, and 10 PCs from common variants as covariates. Multiple testing was corrected on the same stringent scale (0.05/25,000) as in case of the BRVT.

To assess the overlap between significant genes in BIPEDAL and in BRVT a two-sided Fisher exact test was applied.

Independent tests for case-control bias

In addition to the described methods to adjust for a putative case-control bias in BRVT or BIPEDAL, we applied a positive control. A general bias (null hypothesis) would create a false uniform genetic architecture across all genes. We sampled by permutations of gene-variant assignments (100x) or gene-MIP assignments (1,000x), each followed by BRVT and BIPEDAL, respectively, and then correlated (Spearman) the logit-transformed *p*-values with the gene sizes (as measured by the number of variants or MIP targets). We thus obtained an empirical null distribution of correlations to which we compared the observed correlation to get an empirical *p*-value for testing the null hypothesis.

As an additional control approach we applied BRVT to variant subclasses binned by their predicted consequences. For each gene, we calculated the BRVT *p*-values 2,500 times in each subclass after jackknife-sampling of equal numbers of variants in each subclass. For each subclass we thus obtained an empirical distribution of λ, which is the median deviation of the genes' *p*-values from a random distribution. The subclasses' λ distributions were empirically compared by ANOVA.

Fine-mapping

Fine-mapping was applied to the 14 genes with significance above the stringent threshold in BIPEDAL or BRVT and cross-wise Bonferroni-corrected confirmation (Table 2). In BIPEDAL fine-mapping we identified and excluded the subsegments within these genes that did not show a strong burden in cases, retaining those subsegments which carried the burden. To test subsegment *z* binomially (in analogy to eq. 3), we use:

(equation 4) $p_z = \sum_{s=0}^{x_z} B(s|p_{binH0}, n_z)$,

where x_z of the n_z MIPs have a differentially lower target depth, and p_{binH0} is at least as large as the average proportion of MIP targets that have a differentially lower target depth (to reach high specificity, we chose $p_{binH0} = 0.99$). In a sliding window approach, n_z is set to n (= window size) for all subsegments which determines the α and β errors:

(equation 5) $\alpha_n = 0.05/(m/n)$,

(equation 6) $\beta_n = \sum_{s=k_n+1}^{n} B(s|p_{binH1},n)$,

where m is the number of all considered MIP targets and p_{binH1} is expected to be 0.5. Both errors are linked by the critical binomial quantile

(equation 7) $k_n = \text{qbinom}(\alpha_n, n, p_{binH0})$.

We chose the smallest n for which $1 - \beta_n \geq 0.99$.

For BRVT fine-mapping, we merged overlapping promotors and exons from all considered transcripts, and subjected the variants within the merged segments to BRVT with covariates as described above, one-sided significance testing for a greater burden in cases, and Bonferroni correction for the number of tested segments. BRVT might be underpowered if applied to small segments such as a single exon. Therefore, if possible, we extended BRVT fine-mapping to the segments that had been identified in BIPEDAL fine-mapping before.

Fine-mapping interpretation

To test for an overlap between the results of BIPEDAL and BRVT fine-mappings (excluding the results of extended BRVT fine-mapping), we performed 500 rounds of permutation. Within the set of genes subjected to fine-mapping, each round of permutation randomly reassigned the positions of BIPEDAL fine-mapped segments and randomly resampled the same number of BRVT subsegments. The significance level was determined by comparing the resulting empirical distribution of overlapping base pairs with the observed overlap.

For segments that were detected in both the BIPEDAL fine-mapping and the extended BRVT fine-mapping, we determined the called variants and MIP probes that indicated a higher burden of genetic variation in cases. We annotated these probes/variants using Variant

Effect Predictor (VEP)[30] variant consequences as well as functional regions downloaded from the UCSC table browser[31] for ENCODE transcription factor binding sites (TFBS) and conserved TFBS[32], DNase clusters V3 and UMass brain H3K4Me3 peaks[33], miRNA regulatory sites[34] and UNIPROT protein annotations.

Results

Targeted sequencing, variant calling and quality control

We designed 11,214 molecular inversion probes (MIPs) for targeted sequencing of 84 positional and functional RLS candidate genes. After quality control, 8,379 individuals (4,001 cases) and 31,445 variants with MAF $\leq$ 5% in either cases or controls (Supplementary Tables 2 and 3) remained. DNA aliquots from two individuals were processed on each PCR plate, serving as replicates for quality control. Their median pairwise concordance was 0.9992 (95%CIl: 0.9984 to 0.9998) and 0.9994 (95%CI 0.9981 to 0.9999), respectively.

BRVT

We performed burden of rare variant testing (BRVT)[17]. Highly significant burden was found for 14 genes, i.e. *COL20A1*, *CREB5*, *AAGAB*, *DMPK*, *CNTN4*, *MICALL2*, *VAV3*, *CADM1*, *BBS7*, *CRBN*, *ATP2C1*, *GLO1*, *NTNG1*, and *ASTN2* ($p \leq 0.05/25{,}000$). Conditioning BRVT of each gene on the respective GWAS lead SNP or on variant burden of neighboring genes indicated the independence of the signals (Supplementary Table 1). Variant burdens in *MEIS1* which comprises the strongest RLS GWAS signal[11] and its paralog *MEIS2* were significant (p = 5E-04 and p = 4E-05, respectively) but did not overcome the stringent threshold. The BRVT p-values were not a sole effect of a potential genotyping bias between cases and controls (p = 0.03, 100 permutations of gene-to-variant assignments). All significant genes showed a higher burden of minor alleles in the cases, suggesting that their effects in RLS are detrimental. When we binned variants by predicted consequence, performed BRVT in each bin subclass, and determined the deviation of the genes' p-values from a random distribution (λ), we observed a significant difference between the classes (p = 0.001). In subclasses with likely low effect sizes, λ was low. Intronic variants outside of TFBS, for instance, implied λ = 1.9 which was not significantly different from 1 (95%CI 0.7 to 3.9), while missense and 5'-UTR variants implied λ of 2.7 and 3.2, respectively, that were significantly larger than 1 (95%CIs 1.2 to 5.0 and 1.4 to 4.8, respectively).

BIPEDAL

We applied BIPEDAL to MIPseq data of RLS 4,649 cases and 4,982 controls. For estimation of their individual biases (accounting for individual leanings of the MIP targets' depths) in the first step of BIPEDAL we selected a subset of 880 MIPs with 9,193 variants mapping to respective probe and target regions. The burden of variants and the MIPs' general baselines explained a major proportion of the targets' depths variance in each individual (95%CI of R^2: 0.72 to 0.89). The median effect sizes of the burden in probe and target regions were -0.37 and 0.02 (95%CIs -80 to 15 and -1.34 to 1.52), respectively, suggesting that variants in the probe binding regions had a stronger effect on target depth. This negative effect was also observed for the individual variants (Fig 1A to C). As expected, the individual bias was slightly larger in cases (linear regression of individual bias on disease status, $R^2 = 0.064$, $p < 2E\text{-}16$, $\beta = 0.40$): Some MIPs usually generate many reads. If the target depth of such a MIP is substantially lowered in an individual, e.g. due to a copy number variation in the probe binding regions, then the other MIPs' relative proportion of reads in that individual may increase, resulting in a relevant individual bias. This scenario is more likely in cases since they tend to have more genetic variation. We succeeded to calibrate BIPEDAL for 9,434 MIP targets, of which 5,664 showed a significantly different target depth between cases and controls after multiple testing correction. Among the non-significant MIP targets ($p > 0.05$), effect directions were balanced ($p_{bin} = 0.501$).

We then performed BIPEDAL analysis by gene. 16 genes showed stringent significance, i.e. *ATP2C1*, *NTNG1*, *LAMA1*, *STEAP4*, *PTPRM*, *VAV3*, *ADAM22*, *CNTN4*, *PTPRD*, *SUN1*, *OSBP*, *RIMS2*, *BBS7*, *COL6A6*, *CREB5*, and *NRG3*.

Comparison of BRVT and BIPEDAL

BIPEDAL confirmed 10 genes that had stringent significance in BRVT and, vice versa, BRVT confirmed 10 genes detected by BIPEDAL with stringent significance, resulting in 14 genes cross-wise confirmed with Bonferroni-corrected significance of $p < 0.05/14$ and $p < 0.05/16$, respectively (Table 2). Six genes showed stringent significance in both BRVT and BIPEDAL analysis, i.e. *ATP2C1*, *BBS7*, *CNTN4*, *CREB5*, *NTNG1*, *VAV3*. A set of that size was highly unlikely to occur by chance (two-sided Fisher's exact test, $p = 0.023$, 95%CI 1.03 to 18.40).

Fine-mapping

We performed fine-mapping in the 14 genes with stringent high significance in BRVT or BIPEDAL and cross-wise validation. For BIPEDAL fine-mapping a sliding window of 11 MIPs was selected (see Supplemental Data) resulting in 36 fine-mapped segments with 1 to 6 segments in each of the genes while extended BRVT fine-mapping yielded 19 segments with 1 to 4 segments in 12 of the 14 genes. (As described in the Methods section, BRVT fine-mapping was extended, if feasible, to BIPEDAL fine-mapped segments because it may have been underpowered if applied to single exons only.)

BIPEDAL fine-mapped segments intersected significantly with BRVT fine-mapped segments (35,614 bp, 17 overlapping segments with $p = 0.002$ after 500 permutations, excluding extended segments to avoid any statistical bias), demonstrating the reliability of the two methods.

Intersecting BIPEDAL fine-mapping results with the results of extended BRVT fine-mapping yielded 19 validated segments in 13 genes (88,457 bp of the tiling MIP target regions): 3 in *ATP2C1*, 2 in *BBS7*, 1 in *CADM1*, 1 in *CNTN4*, 1 in *COL6A6*, 1 in *CRBN*, 3 in *CREB5*, 1 in *GLO1*, 1 in *NRG3*, 1 in *NTNG1*, 1 in *STEAP4*, 1 in *SUN1*, and 2 in *VAV3* (details are given in Supplementary Table 4). Fifteen of these segments harbored open chromatin loci or transcription factor binding sites. Of note, the analysis revealed RLS-associated segments relating to brain open chromatin marked by H3K4Me3 patterns in *CNTN4*, *NRG3* and *NTNG1*. *COL6A6* and *CADM1* seemed to be affected by regulatory alterations in RLS cases. Thirteen of 19 segments comprised protein coding sequences of which eight have a functional assignment (Table 3). Among these, the calponin-homology domain of VAV3, the Cation-ATPase-N domain of ATP2C1, and the VOC domain of GLO1 seemed to be affected by moderately or severely detrimental variants. The *ATP2C1* locus overlaps with the *ASTE1* locus, which thus also showed detrimental effects in RLS cases.

Discussion

Common variants contribute to rare disease, and rare variants contribute to common disease[35]. The present paper demonstrates that the latter also applies to RLS, one of the most common neurological disorders. By MIP sequencing of a large case control sample of RLS patients we detected a significant burden of low frequency and rare variants in 14 genes (Table 2). In keeping with the preponderance of RLS GWAS loci in the selection of candidate

genes subjected to the MIP analysis, 9 of the 14 genes resided at one of these 19 known RLS loci. Due to the non-random selection of candidate genes we applied a stringent significance threshold in the gene discovery ($p \leq 0.05/25{,}000$, corresponding to the number of genes in the genome) in order to avoid *p*-value hacking. With a Bonferroni correction for only the number of candidate genes, the set of significant results would have been larger and would have included the leading RLS gene *MEIS1* and its paralogue *MEIS2*.

The detection of genes with rare variants affecting the pathogenesis of RLS leads to the question whether allelic series in these genes may include highly penetrant variants that cause monogenic RLS and are detectable by linkage analysis[4]. A recent study that assessed the segregation pattern of rare protein-altering variants from significant GWAS loci in 7 large French-Canadian families had negative results[36]. We observed 21 of the reported variants, most of them with small effect sizes estimates. When we checked the low-frequency variants detected by the present study for linkage in European RLS families whose index patient we had included in the MIPseq analysis, we also could not detect significant co-segregation (unpublished). While this does not exclude the role of rare variants with nearly complete penetrance in the genetic architecture of RLS, we emphasize that for some multifactorial disorders monogenic subgroups may not exit.

Our study has demonstrated that the search for low frequency and rare variant burden is useful in identifying the putatively causative genes at GWAS loci. This might include surprises such as the identification of *GLO1* at the locus of BTBD9 while the latter previously appeared to be the likely RLS gene[37]. Glyoxalase 1, the gene product of *GLO1*, detoxifies methylglyoxal. Decreased Glo1 activity or increased maternal methylglyoxal levels derange neurogenesis in embryonic mice and cause long-term alterations in cortical neurons postnatally[38], in keeping with the concept of RLS being a disorder of neurodevelopment[11]. Of note, a GWAS locus may contain more than one disease-associated gene, by chance or due to functional or developmental relation between genes within the same chromosomal domain[39]. Indeed, at three RLS GWAS loci we detected two genes each having significant burdens of low frequency variants (Table 2).

Identifying causative genes at GWAS loci is important for guiding further functional research in molecular pathology and pharmacology. We identified *CRBN*, for instance, the gene of cereblon, a substrate receptor of the cullin-4 RING E3 ligase (CRL4). Cereblon is bound by thalidomide which inhibits the binding, ubiquitination and proteosomal degradation of CRL4's

endogenous substrate MEIS2[40, 41]. We therefore assume that thalidomide may have therapeutic potential in RLS cases where its teratogenicity is irrelevant, that is, in men and in women without childbearing potential.

Our MIP analysis aimed for complete assessment of a comparatively large number of genes in a large number of individuals. Accordingly, in order to have as little gaps and dropouts as possible, the chosen MIP design and quality control thresholds were not maximized for sequencing precision as in diagnostic MIP applications, but allowed for potentially false variant callings within the MIP sequencing regions. In order to compensate for the latter, we developed BIPEDAL, a method for analyzing sequencing depth of MIPseq to obtain independent information on the variant burden within the MIP binding regions. As expected, BIPEDAL results showed significant and substantial overlap with the BRVT results. BIPEDAL therefore qualifies as a cost-efficient and time-efficient tool for validation of MIP results if the MIP study is focused on segmental variant burden. Of note, however, BIPEDAL and BRVT may reach their maximal reliability in different scenarios: BIPEDAL is less affected by a high burden of variants (including unrecognized copy number variations) which would lead to many missing variant calls and thereby to an underpowered BRVT so that significant BIPEDAL signals might be of interest even if they are not confirmed by BRVT results. On the other hand, BIPEDAL signals represent the effect of variants on probe binding efficiency which differs between variant positions, for instance, and therefore implies a potential bias so that BIPEDAL cannot reach the performance of BRVT applied to hypothetical error-free genotype data.

As we have demonstrated, BIPEDAL can also be used for fine-mapping of variation-sensitive domains of disease-associated genes. Again, there was a highly significant overlap with BRVT fine-mapping, resulting in the identification of regulatory regions or coding segments with the active centers of putatively causative genes (Table 3) (details are given in Supplementary Table 4). Together, our BIPEDAL results may trigger the reanalysis of existing MIPseq datasets and influence the design of future large-scale probe-based re-sequencing analyses.

In summary, we applied BRVT and BIPEDAL to MIPseq data of a very RLS patient-control sample. Significance thresholds were corrected for the number of genes in the genome, that is, more rigidly than the number of analyzed genes or even of MIP probes would have required. We detected RLS associations of 14 genes, i.e. *AAGAB, ATP2C1*, *CNTN4*,

COL6A6, *CRBN*, *GLO1*, *NTNG1*, *STEAP4*, and *VAV3*, as well as *BBS7*, *CADM1*, *CREB5*, *NRG3*, and *SUN1.* Their products mostly function in calcium transport and neurogenesis. With the exception of *AAGAB,* the association could be fine-mapped to coding and regulatory regions within these genes. The first nine of these genes are located in the vicinity of RLS-GWAS signals, the latter five reside at loci that have not been associated with RLS before.

Acknowledgements

We thank all colleagues and staff at the participating centres for their help with recruitment of RLS patients and the German RLS Patient Organisation for continuously supporting our study. This project was funded by a research grant of the Deutsche Forschungsgemeinschaft (DFG, grant id 310572679) to BS and JW. WHOe is Hertie Senior Research Professor, supported by the Charitable Hertie Foundation. The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The COR study was supported with unrestricted grants to the University of Münster by the German RLS Patient Organisation, the Swiss RLS Patient Association, and a consortium formed by Boeringer Ingelheim Pharma, Mundipharma Research, Neurobiotec, Roche Pharma, UCB (Germany + Switzerland) and Vifor Pharma. Sequencing was done on the platform of the Munich Sequencing Alliance (https://www.munich-sequencing-alliance.de/).

Author Contributions

ET, BS, JW, AVS, AH, BM-M conception and design of the study.

ET, CZ, KO, AH, AAN, PL, JW, EH, BH, WP, CGB, WP, CT, WHO, MH, IF, KB, CG, AP acquisition or analysis of data.

ET, KO drafting of the manuscript and figures.

Potential Conflicts of Interests

The authors declare no conflict of interest..

References

1. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome biology. 2017 Apr 27;18(1):77.
2. International Multiple Sclerosis Genetics Consortium. Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. Cell. 2018 Nov 29;175(6):1679-87.e7.
3. Wainschtein P, Jain DP, Yengo L, et al. Recovery of trait heritability from whole genome sequence data. bioRxiv. 2019:588020.
4. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science (New York, NY). 1996 //;273:1516-7.
5. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3;615(1-2):28-56.
6. Oexle K. A remark on rare variants. J Hum Genet. 2010 Apr;55(4):219-26.
7. Ohayon MM, O'Hara R, Vitiello MV. Epidemiology of restless legs syndrome: a synthesis of the literature. Sleep Med Rev. 2012 Aug;16(4):283-95.
8. Trenkwalder C, Allen R, Hogl B, et al. Comorbidities, treatment, and pathophysiology in restless legs syndrome. The Lancet Neurology. 2018 Nov;17(11):994-1005.
9. Winkelmann J, Schormair B, Lichtner P, et al. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet. 2007 Aug;39(8):1000-6.
10. Stefansson H, Rye DB, Hicks A, et al. A genetic risk factor for periodic limb movements in sleep. N Engl J Med. 2007 Aug 16;357(7):639-47.
11. Schormair B, Zhao C, Bell S, et al. Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. The Lancet Neurology. 2017 2017/11/01/;16(11):898-907.
12. O'Roak BJ, Vives L, Fu W, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (New York, NY). 2012 Dec 21;338(6114):1619-22.
13. Krupp DR, Barnard RA, Duffourd Y, et al. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. Am J Hum Genet. 2017 Sep 7;101(3):369-90.
14. Zhang J, Wang X, de Voer RM, et al. A molecular inversion probe-based next-generation sequencing panel to detect germline mutations in Chinese early-onset colorectal cancer patients. Oncotarget. 2017 Apr 11;8(15):24533-47.
15. Jansen S, Hoischen A, Coe BP, et al. A genotype-first approach identifies an intellectual disability-overweight syndrome caused by PHIP haploinsufficiency. Eur J Hum Genet. 2018 Jan;26(1):54-63.
16. Neveling K, Mensenkamp AR, Derks R, et al. BRCA Testing by Single-Molecule Molecular Inversion Probes. Clin Chem. 2017 Feb;63(2):503-12.

17. Auer PL, Wang G, Leal SM. Testing for rare variant associations in the presence of missing data. Genet Epidemiol. 2013 Sep;37(6):529-38.
18. Boyle EA, O'Roak BJ, Martin BK, et al. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics. 2014 Sep 15;30(18):2670-2.
19. Tilch E. Genetics of Restless Legs Syndrome [monography]. Munich: Technical University of Munich; 2018.
20. Wichmann HE, Gieger C, Illig T. KORA-gen - resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen. 2005 Aug;67 Suppl 1:S26-30.
21. Allen RP, Picchietti D, Hening WA, et al. Restless legs syndrome: diagnostic criteria, special considerations, and epidemiology. A report from the restless legs syndrome diagnosis and epidemiology workshop at the National Institutes of Health. Sleep Medicine. 2003 Mar;4(2):101-19.
22. Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PloS one. 2017;12(10):e0185056.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.
24. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.
25. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93.
26. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.
27. Purcell S. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 //;81:559-75.
28. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
29. R Core Team. R: A language and environment for statistical computing. 2013:http://www.R-project.org/.
30. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome biology. 2016 Jun 06;17(1):122.
31. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6.
32. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012 Sep 6;489(7414):91-100.
33. Cheung I, Shulha HP, Jiang Y, et al. Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. Proc Natl Acad Sci U S A. 2010 May 11;107(19):8824-9.
34. Friedman RC, Farh KK, Burge CB, et al. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009 Jan;19(1):92-105.

35. Oexle K, Winkelmann J. Common Grounds for Family Maladies. Neuron. 2018 May 16;98(4):671-2.
36. Akcimen F, Spiegelman D, Dionne-Laporte A, et al. Screening of novel restless legs syndrome-associated genes in French-Canadian families. Neurology Genetics. 2018 Dec;4(6):e296.
37. Allen RP, Donelson NC, Jones BC, et al. Animal models of RLS phenotypes. Sleep Medicine. 2017 Mar;31:23-8.
38. Yang G, Cancino GI, Zahr SK, et al. A Glo1-Methylglyoxal Pathway that Is Perturbed in Maternal Diabetes Regulates Embryonic and Adult Neural Stem Cell Pools in Murine Offspring. Cell reports. 2016 Oct 18;17(4):1022-36.
39. Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. Cell. 2015 Mar 12;160(6):1049-59.
40. Fischer ES, Bohm K, Lydeard JR, et al. Structure of the DDB1-CRBN E3 ubiquitin ligase in complex with thalidomide. Nature. 2014 Aug 7;512(7512):49-53.
41. Tao J, Yang J, Xu G. The interacting domains in cereblon differentially modulate the immunomodulatory drug-mediated ubiquitination and degradation of its binding partners. Biochem Biophys Res Commun. 2018 Dec 9;507(1-4):443-9.

Figure Legends

Figure 1

Evidence that the target sequencing depth of a substantial proportion of MIPs is reduced by low frequency variants located in the corresponding MIP probe binding regions. Tiling design of the MIPs allowed us to identify low frequency and rare variants in probe binding regions of 16% of all MIPs.

A, B) Two examples of rare variants (SNP rs143456273 and indel rs1308722484) in MIP probe binding regions with effect on the MIP target sequencing depth. Genotypes on the horizontal axis, corrected residual depth (logit of normalized target count minus individual bias, $k_{ij} - d_j$) on the vertical axis. The respective regression (grey dashed line) resulted in the effect size β of -0.71 and -0.75, respectively.

C) β values for all MIPs with low frequency or rare variants in their probe binding regions (solid line). Comparison with sampling 1000 null distributions of random variant-to-MIP assignments (dashed). The substantial shift of the solid curve to the left indicates that the majority of such variants negatively impact the target sequencing depth.

D) Distribution of the inflation parameter λ (solid line) of the 1,000 null distributions (see C) as compared to the observed λ (dashed vertical line). The very large observed λ of 5.2 indicates that most of the examined variants affect the target sequencing depth of the corresponding MIP.

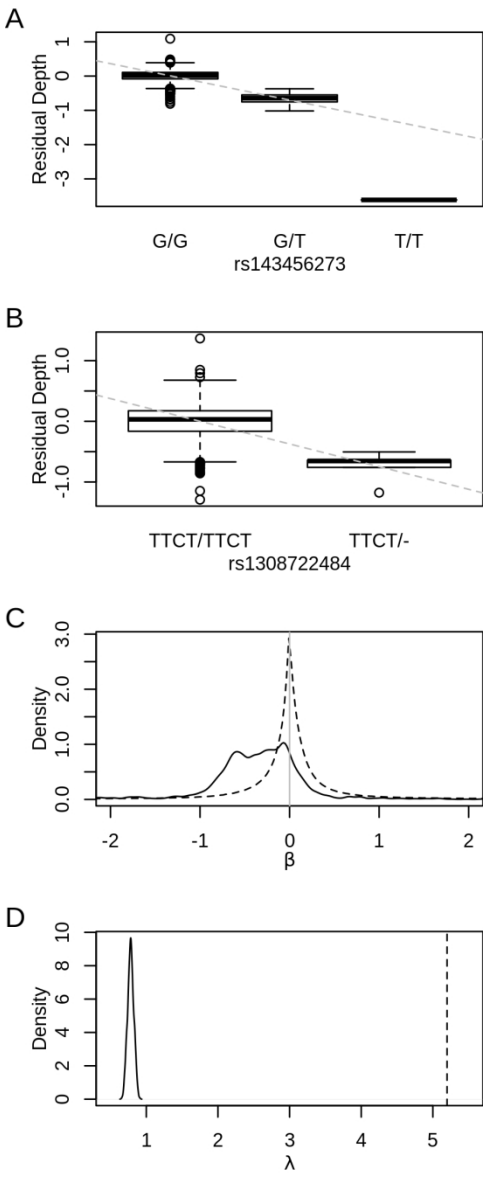

Table 1. Demographics after quality control

	Male/female ratio	**Mean age (SD)**	**Mean age at onset (SD)**
BRVT cases (n = 4,001)	0.52	62.75 (12.60)	42.04 (18.04)
BRVT controls (n = 4,378)	0.94	55.70 (13.18)	NA
BIPEDAL cases (n = 3,975)	0.52	62.67 (12.65)	41.15 (18.03)
BIPEDAL controls (n = 4,261)	0.94	55.62 (13.18)	NA

Table 2. Fourteen genes identified with stringent significance* ($p \leq 0.05/25.000$) in BRVT or BIPEDAL, and cross-wise confirmation by BIPEDAL and BRVT, respectively.

Gene & genomic position (hg19)	**Protein function (extracted from RefSeq Summary, UniProtKB, OMIM)**	p_{BRVT}	$p_{BIPEDAL}$
AAGAB chr15:67,493,012 RLS-GWAS locus	Functions in clathrin-coated vesicle trafficking, e.g. EGFR recycling. Loss of function (LoF) causes autosomal-dominant palmoplantar keratoderma, punctate type I.	8.8E-10*	2.1E-03
ATP2C1 chr3:130,569,368 RLS-GWAS locus	Catalyzes the hydrolysis of ATP coupled with the transport of Ca^{2+}. LoF causes autosomal-dominant Hailey-Hailey skin disease.	7.6E-07*	1.4E-13*
BBS7 chr4:122,745,483	Part of the BBSome, which is required for ciliogenesis. LoF causes autosomal-recessive syndromic intellectual disability (Bardet-Biedl).	4.2E-07*	1.5E-07*
CADM1 chr11:115,044,344	Mediates cell-cell adhesion (Ca^{2+}-independent). May function in synapse assembly, neuronal migration, and axon growth/pathfinding.	3.8E-07*	3.6E-05
CNTN4 chr3:2,140,549 RLS-GWAS locus	Glycosylphosphatidylinositol-anchored axon-associated cell adhesion molecule that functions in neuronal network formation and plasticity.	7.2E-09*	2.4E-09*
COL6A6 chr3:130,279,178 RLS-GWAS locus	Part of the basal lamina of epithelial cells, possibly regulating their cell-fibronectin interactions.	2.1E-03	3.2E-07*
CRBN chr3:3,191,316 RLS-GWAS locus	Substrate recognition component of an ubiquitin-protein ligase (mediates degradation of e.g. MEIS2). May function in memory by regulating neuronal expression of large-conductance Ca^{2+}-activated K^{+}-channels. LoF causes autosomal-recessive nonsyndromic intellectual disability.	7.2E-07*	1.6E-04
CREB5 chr7:28,338,939	Binds CRE (cAMP response element) as a homo-/heterodimer (with c-Jun or CRE-BP1). Functions as a CRE-dependent transactivator.	3.9E-10*	3.5E-07*
GLO1 chr6:38,643,700 RLS-GWAS locus	Synthesis of S-lactoylglutathione. Regulates the TNF-induced transcriptional activity of NF-kappa-B. Required for osteoclastogenesis.	8.6E-07*	9.9E-04
NRG3 chr10:83,635,070	Stimulates activation and phosphorylation of ERBB4. May influence neuroblast population, and act as survival factor in oligodendrocytes.	3.3E-04	7.5E-07*
NTNG1 chr1:107,682,539 RLS-GWAS locus	Functions in patterning and neuronal circuit formation at the laminar, cellular, subcellular and synaptic levels. Promotes neurite outgrowth.	1.1E-06*	9.3E-13*
STEAP4 chr7:87,905,744 RLS-GWAS locus	Transmembrane metalloreductase for Fe^{3+} and Cu^{2+} in the Golgi apparatus. Ubiquitous expression, but not in brain.	6.3E-05	3.0E-11*
SUN1 chr7:855,194	Nuclear envelope protein which is required for radial neuronal migration in the cerebral cortex and glial migration.	8.1E-04	3.0E-08*
VAV3 chr1:108,113,781 RLS-GWAS locus	GTP exchange factors for Rho family GTPases. Functions in actin dynamics, angiogenesis and integrin-mediated cell adhesion.	1.6E-07*	2.7E-10*

Table 3. Of 19 segments determined by intersecting BRVT and BIPEDAL fine-mapping results, ten coded for functional domains in eight genes.

Gene	Functional domain
ATP2C1	Cation transporter/ATPase (N-terminus), Ca2+-binding sites
CRBN	Cereblon domain of unknown activity, binding cellular ligands and thalidomide (CULT) and Lon protease-like (N-terminal)
CREB5	Basic leucine zipper (bZIP) domain
GLO1	Vicinal oxygen chelate (VOC) domain (with substrate and zinc binding sites)
NRG3	Extracellular neuregulin-3 (cleaved from membrane-bound pro-neuregulin-3)
NTNG1	NGL discriminant loops, EGF-like, and laminin-type EGF-like domains
SUN1	SAD1-and-UNC84 (SUN) domain
VAV3	Phorbol-ester/DAG-type zinc finger, calponin homology domain, Dbl homology (DH), pleckstrin homology (PH) and Src homology 3 (SH3) domains