Supplementary Methods

Genome-wide association study identifies candidate loci associated with chronic pain and postherpetic neuralgia

Daisuke Nishizawa, Masako Iseki, Hideko Arita, Kazuo Hanaoka, Choku Yajima, Jitsu Kato, Setsuro Ogawa, Ayako Hiranuma, Shinya Kasai, Junko Hasegawa, Masakazu Hayashida, Kazutaka Ikeda

Supplementary Methods

1. Whole-genome genotyping, quality control, and gene-based and gene-set analyses

1.1. Whole-genome genotyping and quality control

A total of 194 DNA samples from the patients were used for genotyping. Total genomic DNA was extracted from whole-blood samples using standard procedures. The extracted DNA was dissolved in TE buffer (10 mM Tris-HCl and 1 mM ethylenediaminetetraacetic acid [EDTA], pH 8.0). The DNA concentration was adjusted to 50 or 100 ng/μl for whole-genome genotyping using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Briefly, whole-genome genotyping was performed using the Infinium assay II with an iScan system (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Two kinds of BeadChips were used for genotyping 153 and 41 patient samples, respectively: HumanOmni1-Quad v1.0 (total markers: 1,134,514) and HumanOmniExpress-12 v1.1 (total markers: 719,665). Approximately 697,000 markers were commonly included in both of the BeadChips. For genotyping 282 control samples, the HumanOmniExpressExome-8 v1.2 BeadChip (total markers: 964,193) was used. The BeadChips included a number of probes that were specific to copy number variation markers, but most of the BeadChips were for SNP markers on the human autosome or sex chromosome.

The data for the whole-genome-genotyped samples were analyzed using GenomeStudio with the Genotyping module v3.3.7 (Illumina) to evaluate the quality of the results. The genotype data from the two kinds of BeadChips that were used for genotyping patient samples were merged to analyze all of the samples simultaneously (i.e., only markers that were common to the two kinds of BeadChips were included in the analysis, and the

others were automatically excluded). In the data-cleaning process, the samples with a genotype call rate of less than 0.95 were excluded from further analyses. As a result, three patient samples were excluded from further analyses, whereas no control samples were excluded based on this criterion. Markers with a genotype call frequency of less than 0.95 or "Cluster sep" (i.e., an index of genotype cluster separation) of less than 0.1 were excluded from the subsequent association study. A total of 684,499 and 954,793 SNP markers survived this filtration process in the patient and control samples, respectively. For the study of the effects of drugs in patients, markers were filtered based on minor allele frequencies (MAFs) and Hardy-Weinberg equilibrium (HWE) test results. Markers with MAF values less than 0.1 were excluded in the present study to examine relatively common SNPs, whereas markers with p values (df = 1) less than approximately 7×10^{-8} (0.05 / 684,499) were considered significantly deviated in the HWE tests and thus were excluded. Finally, 447,634 SNPs survived the entire filtration process and were used in the study. For the case-control study to compare genotypes between the patient and control subjects, more stringent criteria were used for filtration to remove spurious results. Initially, the genotype data that were available in the patient and control samples were merged to analyze both samples simultaneously (i.e., only markers that were common to both samples were included in the analysis, and the others were automatically excluded). As a result, a total of 681,151 SNP markers were selected. Fisher's exact tests were conducted in the genotypic model to compare genotype distributions between a patient sample group, for which the HumanOmni1-Quad v1.0 BeadChip was used for genotyping, and the other patient sample group, for which the HumanOmniExpress-12 v1.1 BeadChip was used for genotyping. Single-nucleotide polymorphisms with p values (df = 2) less than approximately 7×10^{-8} were excluded from the study to avoid the risk of detecting erroneous associations that can be caused by differences in possibly inappropriate probes between the two kinds of BeadChips. Again, markers with MAF values less than 0.1

were excluded, and markers with p values (df = 1) less than 7×10^{-8} in the entire patient and control samples in the HWE tests were excluded. Additionally, markers with p values (df = 1) equal to or less than 0.001 in either the patient or control samples in the HWE tests were also excluded. Finally, 445,723 SNPs survived the entire filtration process and were used in the study. Furthermore, each genotype cluster of the top 20 candidate SNPs depicted by GenomeStudio software was checked for separation state in both the study of the effects of drugs in patients and case-control study. The TaqMan allelic discrimination assay (Life Technologies, Carlsbad, CA, USA) was performed to confirm the genotype data if the cluster separation was apparently insufficient and the data were suspected to be dubious.

A log quantile-quantile (QQ) p -value plot as a result of the GWAS for the entire sample was subsequently drawn to check the pattern of the generated p-value distribution in the association studies between the SNPs and disease status, in which the observed p values against the values that were expected from the null hypothesis of a uniform distribution, calculated as $-\log_{10}$ (p value), were plotted for each model. All of the plots were mostly concordant with the expected line (y = x), especially over the range of $0 < -\log_{10}$ (p value) < 4 for each model, indicating no apparent population stratification of the samples that were used in the study (Supplementary Fig. S1, S2). In addition, results of GWAS were displayed as "Manhattan" plots, in which $-\log_{10}$ (p value) for each SNP on each chromosome was plotted simultaneously.

1.2. Gene-based and gene-set analyses

To further understand the genetic backgrounds and molecular mechanisms that underlie complex traits, such as chronic pain and PHN, gene-based and gene-set approaches were adopted with Multi-marker Analysis of GenoMic Annotation (MAGMA) v1.06 [1], which is also available on the Functional Mapping and Annotation of Genome-Wide

Association Studies (FUMA GWAS) v1.3.3 platform [2]. In the gene-based analysis, genetic marker data were aggregated to the level of whole genes to test the joint association between all markers in the gene and the phenotype [1]. Similarly, in the gene-set analysis, individual genes were aggregated to groups of genes that shared certain biological, functional, or other characteristics. Therefore, gene-set analysis can provide insights into the involvement of specific biological pathways or cellular functions in the genetic etiology of a phenotype [1]. In these analyses, associations were explored for genes on autosome 1-22 and the X chromosome, and the window of the genes to assign SNPs was set to 20 kb, thereby assigning SNPs within the 20 kb window of the gene (both sides) to that gene. For the reference panel, the 1000 Genome Phase3 EAS population was selected (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/; accessed July 10, 2019). In the gene-set analysis, gene sets were defined using the Molecular Signatures Database (MSigDB) v6.1 [3], and a total of 10654 gene sets (curated gene sets: 4737, GO terms: 5917) from MsigDB were tested. In both analyses, Bonferroni correction for multiple testing was performed for all of the tested genes and gene sets, and adjusted values of $p < 0.05$ in the results were considered significant.

2. Additional *in silico* analysis

2.1. Power analysis

Statistical power analyses were preliminarily performed using G*Power 3.0.5 software [4]. Power analyses for Fisher's exact tests, with degrees of freedom set at 2, indicated that the expected power (1 minus type II error probability) was 80.0% for the type I error probability that was set at 1.858×10^{-7} when the risk allele frequencies for patients and controls were 0.3259 and 0.1000 and the sample sizes for patients and controls were 191 and 282, respectively, in the present study. However, for the same type I error probability and

sample sizes of 191 and 282, the expected power decreased to 50.0% when the risk allele frequencies for patients and controls were 0.2897 and 0.1000, respectively. Conversely, the estimated risk allele frequencies for patients and controls were 0.3452 and 0.1000 for the same type I error probability and sample sizes of 191 and 282, respectively, to achieve 90% power. Therefore, a single analysis in the present study was expected to detect true associations with the phenotype with 80% statistical power for effect sizes from large to moderately medium but not small, although the exact effect size has been poorly understood in cases of single-nucleotide polymorphisms (SNPs) that significantly contribute to chronic pain and postherpetic neuralgia (PHN).

2.2. Linkage disequilibrium analysis

To identify relationships between SNPs in the *ABCC4* gene and flanking region, which includes the most potent SNPs that were selected after the GWAS, linkage disequilibrium (LD) analysis was performed for SNPs with a minor allele frequency ≥ 0.05 in the genomic position around the rs4773840 SNP, which spans from ~30 kbp upstream of the rs4773840 SNP to ~40 kbp downstream of the 3' end of the gene, using Haploview 4.2 software [5] based on genotype data for the 191 patient subjects that passed the quality control criteria (Supplementary Fig. S3). For the estimation of LD strength between SNPs, the commonly used r^2 values were pairwise-calculated using the genotype dataset of each SNP. Linkage disequilibrium blocks were defined among the SNPs that showed "strong LD," based on the default algorithm of Gabriel et al. [6], in which the upper and lower 95% confidence limits on *D'* for a strong LD were set to 0.98 and 0.7, respectively.

2.3. Reference to databases

To scrutinize the candidate locus, SNPs, and genes that may be associated with

human chronic pain or PHN, several databases and bioinformatic tools were referenced, including the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov; accessed July 9, 2019), Genotype-Tissue Expression (GTEx) portal (https://www.gtexportal.org/home/; accessed July 10, 2019) [7], Exome Aggregation Consortium (ExAC) Browser (Beta; http://exac.broadinstitute.org/; accessed July 10, 2019) [8], SNPinfo Web Server (https://snpinfo.niehs.nih.gov/; accessed July 10, 2019) [9], and SNPnexus (https://snp-nexus.org/index.html; accessed July 10, 2019) [10-13]. The GTEx portal provides open access to such data as gene expression, quantitative trait loci, and histology images, based on the GTEx project, which is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation [7]. The ExAC Browser provides datasets that span 60,706 unrelated individuals sequenced as a part of various disease-specific and population genetic studies [8]. The SNPinfo Web Server is a set of web-based SNP selection tools (freely available at http://www.niehs.nih.gov/snpinfo; accessed July 10, 2019) where investigators can specify genes or linkage regions and select SNPs based on GWAS results, LD, and predicted functional characteristics of both coding and non-coding SNPs [9]. SNPnexus was designed to simplify and assist in the selection of functionally relevant SNPs for large-scale genotyping studies of multifactorial disorders and provides the scientific community with a user-friendly web interface to extract the broadest annotations for query variants, all from a single location [10-13].

References

1. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11(4):e1004219.
2. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and

annotation of genetic associations with FUMA. Nat Commun 2017;8(1):1826.

3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545-50.

4. Faul F, Erdfelder E, Lang AG, Buchner A: G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191.

5. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21(2):263-265.

6. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M et al: The structure of haplotype blocks in the human genome. Science 2002;296(5576):2225-2229.

7. GTEx Consortium: The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45(6):580-585.

8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al: Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536(7616):285-291.

9. Xu Z, Taylor JA: SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res 2009;37(Web Server issue):W600-W605.

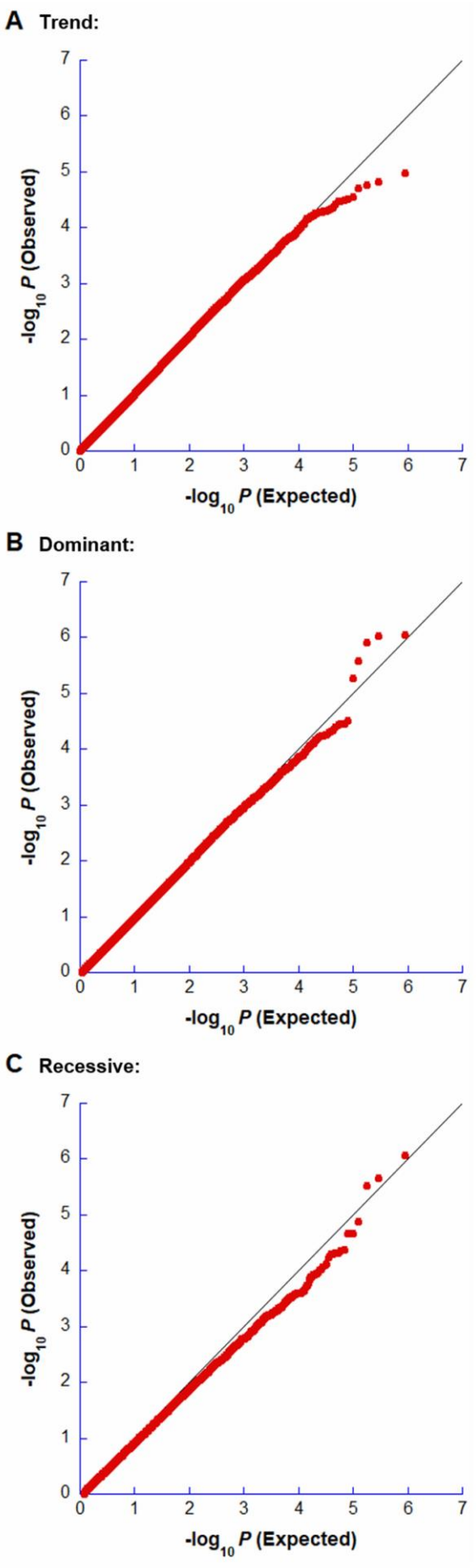
10. Chelala C, Khan A, Lemoine NR: SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 2009;25(5):655-661.

11. Dayem Ullah AZ, Lemoine NR, Chelala C: SNPnexus: a web server for functional

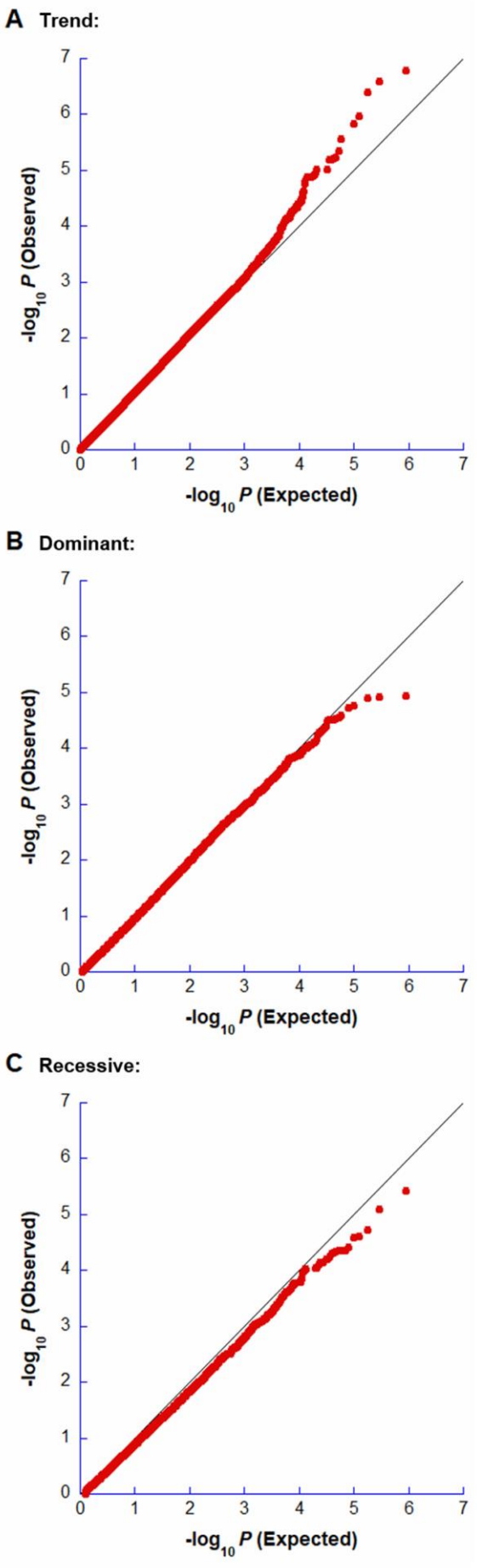
annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res 2012;40(Web Server issue):W65-W70.

12. Dayem Ullah AZ, Lemoine NR, Chelala C: A practical guide for the functional annotation of genetic variations using SNPnexus. Brief Bioinform 2013;14(4):437-447.

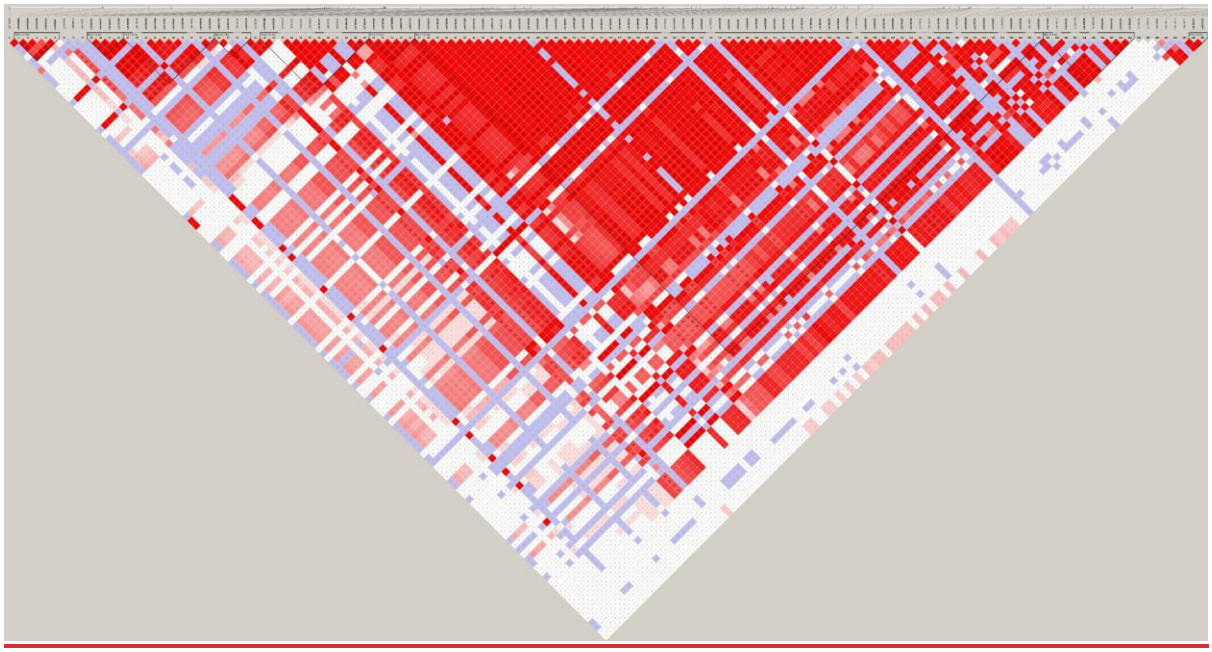
13. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C: SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Res 2018;46(W1):W109-W113.


A Trend:
-log10 P (Observed)
-log10 P (Expected)
B Dominant:
-log10 P (Observed)
-log10 P (Expected)
C Recessive:
-log10 P (Observed)
-log10 P (Expected)

Supplementary Figure S1. Log quantile-quantile (QQ) *p*-value plot as a result of the GWAS for all patient samples. (A) Plot of the results from the trend model. (B) Plot of the results from the dominant model. (C) Plot of the results from the recessive model.


A Trend:
-log10 P (Observed)
-log10 P (Expected)
B Dominant:
-log10 P (Observed)
-log10 P (Expected)
C Recessive:
-log10 P (Observed)
-log10 P (Expected)

Supplementary Figure S2. Log quantile-quantile (QQ) *p*-value plot as a result of the GWAS for the PHN patient samples. (A) Plot of the results from the trend model. (B) Plot of the results from the dominant model. (C) Plot of the results from the recessive model.



Supplementary Figure S3. State of linkage disequilibrium (LD) between SNPs in the genomic region of the *ABCC4* gene. The genotype data for the 191 patient subjects that passed the quality control criteria were used in the LD analysis, and SNPs with a minor allele frequency ≥ 0.05 were selected in the genomic position around the rs4773840 SNP, which spans from ~30 kbp upstream of the rs4773840 SNP to ~40 kbp downstream of the 3' end of the gene. Numbers in squares in which two SNPs face each other represent the percentage of the r^2 values that were calculated from the genotype data of the SNPs. Blank squares represent $r^2 = 1$.