Whole genome sequencing identifies putative associations between genomic polymorphisms and clinical response to the antiepileptic drug *levetiracetam*

Vavoulis DV[1*], Pagnamenta AT[1], Knight SJL[1], Pentony MM[1], Armstrong M[2], Galizia EC[3], Balestrini S[3,4], Sisodiya SM[3,4] & Taylor JC[1*]

1. NIHR Oxford Biomedical Research Centre, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
2. New Medicines, UCB Pharma, Slough, UK
3. Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, UK
4. Chalfont Centre for Epilepsy, London, UK.

*corresponding authors

ABSTRACT

In the context of pharmacogenomics, whole genome sequencing provides a powerful approach for identifying correlations between response variability to specific drugs and genomic polymorphisms in a population, in an unbiased manner. In this study, we employed whole genome sequencing of DNA samples from patients showing extreme response (n=72) and non-response (n=27) to the antiepileptic drug levetiracetam, in order to identify genomic variants that underlie response to the drug. Although no common SNP (MAF>5%) crossed the conventional genome-wide significance threshold of 5×10^{-8}, we found common polymorphisms in genes *SPNS3*, *HDC*, *MDGA2*, *NSG1* and *RASGEF1C*, which collectively predict clinical response to levetiracetam in our cohort with ~91% predictive accuracy (~94% positive predictive value, ~85% negative predictive value). Among these genes, *HDC*, *NSG1*, *MDGA2 and RASGEF1C* are potentially implicated in synaptic neurotransmission, while *SPNS3* is an atypical solute carrier transporter homologous to *SV2A*, the known molecular target of levetiracetam. Furthermore, we performed gene- and pathway-based statistical analysis on sets of rare and low-frequency variants (MAF<5%) and we identified associations between genes or pathways and response to levetiracetam. Our findings include a) the genes *PRKCB* and *DLG2*, which are involved in glutamatergic neurotransmission, a known target of anticonvulsants, including levetiracetam; b) the genes *FILIP1* and *SEMA6D*, which are involved in axon guidance and modelling of neural connections; and c) pathways with a role in synaptic neurotransmission, such as *WNT5A-dependent internalization of FZD4* and *disinhibition of SNARE formation*. Targeted analysis of genes involved in neurotransmitter release and transport further supports the possibility of association between drug response and genes *NSG1* and *DLG2*. In summary, our approach to utilise whole genome sequencing on subjects with extreme response phenotypes is a feasible route to generate plausible hypotheses for investigating the genetic factors underlying drug response variability in cases of pharmaco-resistant epilepsy.

AUTHOR SUMMARY

Levetiracetam (LEV) is a prominent antiepileptic drug prescribed for the treatment of both focal and generalised epilepsy. The molecular mechanism mediating its action is not well understood, but it involves the modulation of synaptic neurotransmition through binding to the synaptic vesicle glycoprotein SV2A. Identifying genomic polymorphisms that predict response to the drug is important, because it can help clinicians prescribe the most appropriate treatment in a patient-specific manner. In this study, we employed whole genome sequencing (WGS) of DNA samples from extreme responders or non-responders to LEV and we identified a small group of common variants, which successfully predict response to the drug in our cohort. These variants are mostly located in genes implicated in synaptic function. Furthermore, we identified significant associations between clinical response to LEV and low-frequency variants in genes and pathways involved in excitatory neurotransmission or in the moulding of neural networks in the brain. Our approach to utilise WGS on subjects with extreme response phenotypes is a feasible route to generate plausible hypotheses on the genomic basis of pharmaco-resistant epilepsy. We expect that the rapidly decreasing cost of WGS will allow conducting similar studies on a larger scale in the near future.

INTRODUCTION

The advent of next-generation sequencing (NGS) has made possible the routine reconstruction of an individual's genetic variation profile across their whole genome[1,2], while the introduction of NGS to clinical practice brings closer the promise of personalised medicine for diagnostic sensitivity and therapeutic precision[3,4]. In the context of pharmacogenomics, whole exome and genome sequencing combined with appropriate bioinformatics and statistical analysis has the potential to identify variants that correlate with clinical response to specific drugs, in a comprehensive, high-resolution and unbiased manner[5–12], i.e. without the need for a prior hypothesis regarding the type (e.g. common or rare), location or distribution of genomic polymorphisms across the whole extent of the genome. We employed whole genome sequencing to better understand response variability to the antiepileptic drug levetiracetam (LEV), a third-generation first-line drug for the treatment of both focal and generalised epilepsies.

Experiments in mice show that *SV2A*, but not its paralogs *SV2B* and *SV2C*, is the molecular target of LEV[13]. *SV2A* is a synaptic glycoprotein with widespread distribution in the brain[14] and a crucial role in synaptic vesicle exocytosis[15]. Mice deficient in SV2 functionality exhibit severe seizures with a concomitant reduction in (inhibitory) GABAergic neurotransmission[16] and an abnormal presynaptic accumulation of calcium leading to increased neurotransmitter release[17]. LEV inhibits presynaptic calcium channels[18] and calcium-dependent vesicle exocytosis[19], and it reverses synaptic deficits due to overexpression of *SV2A*[20]. However, its exact mechanism of action as an antiepileptic drug is not understood.

It is natural to hypothesize that LEV may act by modifying deregulated *SV2A*-dependent neurotransmission and that variability in SV2A functionality may explain differential responsiveness to treatment with LEV. This view is supported by reports showing that partial loss of SV2A functionality is linked to decreased LEV efficacy in several mice seizure models[21], or that levels of SV2A expression in tumour and peri-tumoral tissue predicts clinical response to LEV in patients with glioma[22]. However, neither common nor rare polymorphisms in *SV2A* (including polymorphisms overlapping its binding site with LEV) are associated with clinical response to the drug, based on targeted sequencing approaches[23,24]. Any role of genetic variation (either rare or common) in other genomic loci as potential predictors of LEV efficacy remains to be elucidated.

We analysed whole genome sequencing (WGS) data from 99 people with epilepsy, classified as extreme responders (n=72) or non-responders to LEV (n=27), aiming to explore the genetic differences between the two groups and to identify rare or common polymorphisms that may be predictive of the response/non response phenotype. Using whole genome sequencing (instead of targeted sequencing or genome-wide SNP arrays) facilitates the search for genetic predictors to LEV in a complete, high-resolution and unbiased manner. At the same time, a targeted search for genomic features associated with response to LEV is still possible. Here, we identified common polymorphisms which collectively predict a substantial fraction of clinical response to the drug in our cohort of patients with epilepsy. Furthermore, analysis of groups of low-frequency variants highlights significant associations between response to LEV and genes involved in synaptic neurotransmission, axon guidance and modelling of neural connections.

METHODS

Sample acquisition and whole genome sequencing

The study was approved by the relevant local ethics committee. Patients provided written informed consent, or in the case of people unable to provide consent, assent was obtained from parents or guardians as permitted within the approved protocol.

Ninety-nine unrelated adults with a range of types of epilepsy were recruited from the National Hospital for Neurology and Neurosurgery. Non-responders (n=27; ~27%) were defined as patients who had failed to respond to at least two of the currently established epilepsy treatments and had not responded to maximum tolerated doses of levetiracetam used for at least 12 months. Extreme responders were defined as patients who became seizure-free for at least 12 months after initiation of levetiracetam and who had not previously responded to at least three appropriately chosen and used antiepileptic drugs (AEDs; n=72; ~73%).

Samples from the above subjects were sequenced at the Oxford Genomics Centre using the HiSeq2500 platform, v3 chemistry and the 100bp paired-end read format (Illumina, San Diego, CA). Sequencing was performed across 2.3 lanes per sample at depth 30X (Figure 1).

Bioinformatics analysis

Reads were mapped to hs37d5 using BWA[25] and duplicate reads were removed using the MarkDuplicates option from the Picard toolkit[26] all with default options. Variants were called simultaneously across all 99 samples with Platypus[27] v0.7.9.3 resulting in a multi-sample VCF file. Read alignments were checked visually using the Integrative Genomics Viewer v2.3.5[28].

In total, ~20M variants were called across all samples (Figure 1). We excluded variants in multi-allelic loci or in sex chromosomes, variants with FILTER flag other than PASS, and variants in homopolymers with running length larger than 8 base pairs (HP>8). We excluded genotypes of low quality (PHRED score GQ<20), and with less than 10 reads covering the variant location (DP<10). We also excluded variants not in Hardy-Weinberg Equilibrium[29] (p-value less than 10^{-6}) and with missing genotypes in more than 2 individuals (~2%). Furthermore, we excluded variants in low complexity regions[30], in poor mappability regions[31], in segmental duplications[32] and in the top 1% most variable genes according to Ingenuity IVA[33]. On the remaining ~8.4M variants, we conducted principal component analysis using the *prcomp* function in R[34] and we identified 7 outlier samples, which were excluded from further analysis (Figure 2A,B). We did not find evidence of association between clinical response and sex in the remaining 92 patients (Figure 2C). The filtered data were annotated using the Ensembl Variant Effect Predictor[35] software v90.5 with allele frequency annotations provided by gnomAD r2.0.1[36] and variant IDs provided by dbSNP[37] build 150. Overall, we reviewed ~3.9M common variants (MAF>5%), and ~4M low-frequency (1%<MAF<5%) and rare variants (MAF<1%) (Figure 2D,E).

Statistical analysis

We conducted single-variant tests on common variants, and gene- and pathway-based tests on low-frequency and rare variants (Figure 1). In the case of common variants, we calculated SNP-specific p-values by applying a two-tailed Fisher's exact test on each common variant (Figure 3A). In a pre-specified second stage, we selected a small subset of variants by using all variants with p-value less than the conventional suggestive genome-wide significance threshold of 10^{-5} (n=23 variants; Table S1 and Figure 3A, white dots) as predictors (along with sex) in a penalised logistic regression model[38] (known as the *LASSO*; Figure 3B). An optimal penalisation parameter was estimated using leave-one-out cross-validation. This resulted in the selection of 10 out of 23 variants with maximal predictive power (Figure 3B, red dots). An additional selection step was applied by filtering out all variants (among those selected by the LASSO in the previous step) that had non-protein-coding gene annotation or were annotated as *intergenic*. This resulted in the final selection of 5 variants with protein-coding gene annotations (Table S1). The reason for this final selection step was to avoid overfitting during the downstream analyses described below and because the selection of variants in protein-coding genes (instead of non-protein coding or intergenic variants) facilitates the subsequent investigation of their possible biological relevance. After variant selection, we conducted an analysis of deviance by examining a series of logistic regression models using response to LEV as the dependent variable (Table S2). The BASIC model includes, besides the intercept, a single predictor, sex. The FULL model includes in addition the genotypes of the previously selected variants. A number of intermediate models are simple extensions of the BASIC model through the inclusion of just one of these variants. Finally, we calculated the predictive power of the FULL model using leave-one-out cross-validation and the accuracy (ACC), sensitivity (TPR), specificity (TNR), positive (PPV) and negative (NPV) predictive values, and Matthews correlation coefficient (MCC) as metrics of predictive power. For completeness, we also conducted auxiliary statistical analyses, which included a genome-wide Bayesian analysis and calculation of bespoke genome-wide significance thresholds (see Supplementary Material for more details).

In the case of rare and low-frequency variants, we first calculated a variant-specific p-value by applying a two-tailed Fisher's exact test, as in the case of the common variants. Subsequently, we aggregated all variant-specific p-values in a gene- or pathway-specific statistic using an appropriately corrected Fisher's product method[39] (see Supplementary Material), which takes into account the effective number of independent variants in a group of variants, thus correcting for correlations between variants in the same gene or pathway. The resulting statistic was used to calculate a gene- (Table S3) or pathway-specific (Table S4) p-value for testing the null hypothesis that none of the variants in the gene/pathway are associated with response to LEV, against the alternative hypothesis that at least one variant in the set is associated with response to LEV. P-values were corrected for multiple hypothesis testing across all genes or pathways using Sidak's method.

Finally, we conducted a targeted analysis of common and rare variants in a set of genes implicated in neurotransmitter transport and release and in a set of genes associated with epilepsy (Table S5). For the common variants, we tested each variant individually using a two-tailed Fisher's exact test of independence, as above. We used Sidak's method for multiplicity correction across all genes in each of the two sets. The effective number of independent variants was estimated by first calculating a gene-specific estimate of the number of independent variants

using four alternative methods[39–43], followed by summing these estimates over all genes. All four methods returned consistent results. For the rare variants, we calculated gene-specific p-values followed by multiplicity correction using the Sidak method, as before.

More details on the statistical analysis are given in the Supplementary Material.

RESULTS

Common polymorphisms in genes *SPNS3*, *HDC*, *NSG1*, *MDGA2* and *RASGEF1C* predict clinical response to LEV in our cohort with overall accuracy ~91%

We constructed a statistical model that utilises common genomic variation to predict response to LEV in our cohort. Towards this aim, we first assessed the significance of association between each SNP and response to LEV (Figure 3A). The smallest SNP-specific p-value calculated at this stage was 1.6×10^{-7}, i.e. no p-value crossed the conventional genome-wide significance threshold of 5×10^{-8} (Table S1). This was followed by a principled SNP selection process (see Methods) to identify a minimal set of highly predictive variants (n=5 variants). These are located in the protein-coding genes *SPNS3*, *HDC*, *NSG1*, *MDGA2* and *RASGEF1C*, as indicated by the non-zero coefficients in Figure 1B. Variants with non-zero coefficients in the non-coding genes *RP11-284F21.8*, *RP11-446J8.1* and *RP11-650J17.1*, as well as two intergenic variants in chromosome 15, were not included, in order to keep the model small and avoid overfitting (see Methods for rationale). All these variants are listed in Supplementary Table S1.

At the next stage, we conducted an analysis of deviance on the polymorphisms identified in the previous step (see Methods and Table S2). We found that the inclusion of these SNPs in a logistic regression model reduces the residual deviance from ~107 (BASIC model) to ~28 (FULL model), thus significantly improving the goodness of fit (p-value=1.15×10^{-15} based on a χ^2 test) of the model to the data. The fraction of explained deviance in the data was assessed using a pseudo-R^2 metric, the adjusted D^2, as described in Guisan & Zimmermann[44]. The BASIC and FULL models have an adjusted D^2 equal to 1% and 73%, respectively, which implies that the identified variants in genes *SPNS3*, *HDC*, *NSG1*, *MDGA2* and *RASGEF1C* collectively explain ~72% of the total deviance (Table S2). When considering just a single gene as predictor (as in any of the intermediate models between BASIC and FULL), the improvement in model fit is significant (as indicated by the low p-values). Furthermore, the proportion of explained deviance by SNPs in each gene ranges between 10% (*HDC*) and 21% (*SPNS3*), as inferred by comparing the adjusted D^2 value for each of the intermediate models to the adjusted D^2 value of the BASIC model.

Subsequently, we assessed the predictive power of the FULL model using leave-one-out cross-validation. In brief, this involves fitting the FULL model in all but one subjects and predicting the response phenotype of the held-out subject using the fitted model. This process of model fitting and prediction is repeated until all 92 subjects have been used for prediction. We found that the FULL model correctly predicts clinical response to LEV in 62 responders and 22 non-responders, which corresponds to ~94% sensitivity (TPR) and positive predictive value (PPV), ~85% specificity (TNR) and negative predictive value (NPV), and ~91% overall predictive accuracy

(ACC). The Matthews correlation coefficient (MCC), a balanced performance metric for binary classifiers even when the two classes are of very different size, was equal to ~79%.

Local genomic structure near the identified variants and possible biological relevance

For gene *NSG1* on chromosome 4, three highly correlated SNPs (rs7695197, rs3981 and rs12641832) are located ~5kb upstream of the gene, less than 3kb upstream or downstream of transcription factor binding sites (TFBS) and DNaseI hypersensitivity sites (DHS), and less than 5kb upstream of a small cluster of conserved elements (CE; Figure 4A). The odds ratio for a recessive model (with respect to the ALT allele) is ~23 times in favour of the non-responders, while the corresponding odds ratio for a dominant model is ~2.7 (see Table S1 for the number of homozygous/heterozygous cases in each group). In other words, non-responders to LEV are ~23 times more likely to be homozygous for the alternative allele than responders. *NSG1* (Neuronal Vesicle Trafficking Associated 1) is abundantly expressed in the brain[45,46] and it plays a role in synaptic neurotransmission and plasticity due to its involvement in recycling and trafficking of receptors, such as the glutamate receptor AMPA, the amyloid precursor protein (APP), and the L1 cell adhesion molecule (L1CAM)[47].

The intronic variant rs34570575 in gene *RASGEF1C* on chromosome 5 overlaps a DHS and it is located ~5kb upstream of a TFBS and a cluster of CE (Figure 4B). The odds ratio for a dominant model of inheritance (with respect to the ALT allele) is slightly higher than that of a recessive model (~9.5 and ~8, respectively; Table S1). *RASGEF1C* (RAS guanyl-nucleotide exchange factor domain family member 1C) is abundantly expressed in the brain[45,46]. It belongs to a family of proteins containing the RASGEF domain, which regulates the GTPase activity of RAS-like proteins. These comprise a superfamily of membrane-associated signalling molecules involved in a variety of essential cellular processes, including vesicle trafficking and synaptic function[48–50].

In gene *MDGA2* on chromosome 14, rs1952220 is an intronic variant, less than ~4kb from CE, TFBS and DHS (Figure 4C). The odds ratios for recessive and dominant models (with respect to the ALT allele) are 0.11 and 0.61 in favour of the non-responders, respectively, suggesting a recessive model where non-responders to LEV are ~9 times less likely to be homozygous for the alternative allele than responders (Table S1). The *MDGA2* (MAM Domain Containing Glycosylphosphatidylinositol Anchor 2) mRNA is expressed in the cerebral cortex[45,46]. MDGAs are Ig superfamily cell adhesion molecules that contribute to the radial migration of cortical neurons during early neural development. They play an important, neuroglin-2-dependent role in controlling the function of inhibitory synapses, and they have been associated with autism spectrum disorders and schizophrenia[51,52].

In gene *HDC* on chromosome 15, rs7182203 is an intronic variant that overlaps a TFBS and a DHS, and it is within 5kb of upstream or downstream CE (Figure 4D). From Table S1, the odds ratios for recessive and dominant models (with respect to the ALT allele) are 1.1 and 0.12 in favour of the non-responders, respectively. This implies that patients that respond to LEV are ~8 times more likely to be homozygous or heterozygous for the alternative allele in comparison to non-responders. *HDC* (histidine decarboxylase) is expressed in the brain[45,46], and it catalyses the synthesis of histamine, which is implicated, among others, in neurotransmission and smooth muscle tone. Elevated levels of histamine in the brain appear to suppress seizures and confer neuroprotection, thus antiepileptic agents that boost the levels of histamine in the brain may act by increasing *HDC* activity[53]. Furthermore, *HDC*

has been linked to the pathogenesis of Tourette's syndrome[54]. Interestingly, LEV has been used for the treatment of Tourette's syndrome, although its efficacy has not been established[55–57].

Finally, in gene *SPNS3* on chromosome 17, the intronic variants rs2047231, rs2047232 and rs2047233 overlap a DHS and a cluster of CE, and they are located within 5kb of upstream or downstream TFBS (Figure 4E). From Table S1, the odds ratio for a recessive model (with respect to the ALT allele) ranges among these three SNPs between 0.07 and 0.09 in favour of non-responders. This implies that patients responding to LEV are between ~11 and ~14 times more likely to be homozygous for the alternative allele than non-responders. *SPNS3* (a putative sphingolipid transporter 3) is expressed in the cerebral cortex[45,46]. Both *SPNS3* and *SV2A,* the known target of LEV, are atypical solute carrier (SLC) transporters. They belong to the Major Facilitator Superfamily (MFS) of membrane transporters, and they share a common structure consisting of 12 transmembrane segments, which is necessary for optimal transporter activity[58,59].

Tests on sets of low frequency variants (MAF<5%)

Next, we studied variants with MAF<5%, i.e. low-frequency and rare variants. Among the approximately 4M variants with MAF<5%, we focused on the top 5% genotypically most variable variants across all 92 samples in our cohort. These included ~182K variants with MAF between 0.003% and 5%. A common strategy for increasing statistical power when studying low-frequency and rare variants is to analyse sets of variants, instead of individual variants. Therefore, we examined gene- and pathway-based sets of variants (see Methods).

Gene-based tests indicate that low-frequency variants in genes PRKCB, DLG2, FILIP1, SEMA6D and LINC01090 are associated with response to LEV

We conducted 19,824 gene-based tests, which is the number of genes harbouring at least one of the ~182K low-frequency and rare variants in our data. We found that four protein-coding genes (*PRKCB*, *DLG2*, *FILIP1* and *SEMA6D*) and a long intergenic non-protein-coding RNA (*LINC1090*) had a Family-Wise Error Rate (FWER) less than 10%, and they were kept for further study (Table S3 and Figure 5).

The top hit, *PRKCB*, encodes a protein kinase C, a family of serine- and threonine-specific protein kinases, which can be activated by calcium and second messenger diacylglycerol[47]. There are 78 variants in *PRKCB* with MAF between 2.2% and 4.9%. Forty-five of them have p-values less than 0.05 and they aggregate towards the 5' end of the gene (Figure 5A). Associated Reactome pathways are *glutamate binding*, *activation of AMPA receptors* and *synaptic plasticity*[60]. *PRKCB* is implicated in the trafficking of GluR2-containing AMPA receptors[60]. It is known that fast synaptic excitation relevant to epilepsy is mediated mainly by AMPA receptors, thus rendering the latter potential targets of antiepileptic treatment[61]. There is evidence suggesting that LEV interacts with AMPA receptors[62] and that its antiepileptic action is mediated by inhibiting glutamatergic neurotransmission through presynaptic calcium channels[63], but the precise molecular mechanism that mediates its action remains unclear.

A second hit of interest, *DLG2*, encodes a membrane-associated guanylate kinase, which is implicated in the clustering of receptors (including NMDARs), ion channels, and associated signalling proteins at postsynaptic sites of excitatory synapses[47]. We found 208 variants in this gene with MAF between 0.96% and 4.97%, 53 of which

have p-values less than 0.05 (Figure 5B). A related Reactome pathway is *protein-protein interactions at synapses*[60]. There is evidence supporting the role of NMDARs in epilepsy, and as a potential therapeutic target of antiepileptic drugs, including LEV[64]. It is possible that LEV blocks epileptiform bursting induced by NMDA *in vitro* without affecting normal synaptic transmission[65] and that it inhibits NMDA-dependent excitatory postsynaptic currents[63], although its precise molecular mechanism of action remains unclear.

Among the remaining three hits (Figure 5C-E), *FILIP1* includes 12 variants (11 with p-values less than 0.01) with MAF between 1.8% and 5%, *SEMA6D* has 41 variants (17 with p-values less than 0.05) with MAF between 1.9% and 5% and *LINC01090* harbours 35 variants (18 with p-values less than 0.01) with MAF between 1.7% and 5%. *FILIP1* encodes a protein that stimulates filamin A degradation, which may regulate cortical neuron migration, dendritic spine morphology, and normal excitatory signalling[47]. *SEMA6D* encodes a transmembrane semaphorin, a class of proteins involved in axon guidance, and maintenance and remodelling of neural connections[47]. Finally, *LINC01090* is transcribed into a long intergenic non-protein-coding RNA[47], which is associated with post-traumatic stress disorder[66].

Pathway-based tests indicate that associations between response to LEV and Reactome pathways are driven mainly by low-frequency variants in gene PRKCB

We conducted tests using gene sets, instead of single genes, as the organisational unit for grouping individual variants together. We have used all pathways from *Reactome*, a curated, peer-reviewed database of interacting signalling and metabolic molecules, which are organised into groups of higher order structures (pathways) with well-defined biological relevance[60]. In total, we considered 2,028 pathways, of which 1,979 harboured at least one of the ~182K low-frequency highly-variable variants in our data. Among these, we identified six pathways with FWER<5% and one pathway with FWER<10% (Table S4).

The top hit is *activation of NF-kappaB (nuclear factor kappaB) in B cells*. NF-kappaB is a ubiquitous transcription factor, which is instrumental in gene regulation relevant to cell death and survival and to the immune system's response to inflammation. The next hit is *WNT5A-dependent internalization of FZD4*. WNT5A regulates multiple intracellular signalling cascades via internalisation of its receptors. These include FZD4, a member of the frizzled gene family, which encode seven-transmembrane domain proteins[47]. Importantly, the WNT5A-dependent uptake of FZD4 occurs in a clathrin-dependent manner[67]. Clathrins are adaptor proteins, which are essential in the formations of synaptic vesicles, and which are known to interact with SV2A, the molecular target of LEV[68].

Another interesting pathway is *disinhibition of SNARE formation*. SNARE is a family of proteins, which are important components of the mechanism responsible for membrane fusion, thus playing an important role in docking of synaptic vesicles with the presynaptic membrane, and neurotransmitter release. It is known that SV2A, the target of LEV, regulates the formation of SNARE complexes: kindling epileptogenesis triggers the long-term accumulation of both SNARE and SV2A in the ipsilateral hippocampus, a molecular process which is reversed by LEV[69].

Next, we asked which genes underlie these findings. In total, in these pathways, there are 108 genes harbouring low-frequency mutations (Table S4). In Figure 6, we illustrate these genes, as well as their pathway membership. *PRKCB* is mutated in all but the least significant pathway with FWER<10%, followed by its paralog, *PRCKA*, which is mutated in three pathways (central panel). The remaining genes are mutated in only 1 or 2 pathways. Furthermore, *PRKCB* and *PRKCA* harbour the largest number of low-frequency mutations, along with *RUNX1* (top panel). However, in *PRKCB*, more than half of these mutations have p-values less than 0.05 (see also Table S3) leading to a low gene- and pathway-based p-value (Tables S3 and S4), while only a very small proportion of mutations in *PRKCA* and *RUNX1* have p-values less than 0.05. Among the other highly mutated genes, *LPR6* and *IKBKB* also harbour a large proportion of mutations with low p-values, but they participate in only 1 and 2 pathways, respectively. We conclude that the significant associations in Table S4 are driven mainly by *PRKCB* in all but the least significant pathway with FWER<10%. Associations in this last pathway (*disassembly of the destruction complex and recruitment of AXIN to the membrane*) are driven mainly by *LRP6*, a transmembrane low density lipoprotein (LDL) receptor[47]. Neuronal *LRP6*-mediated Wnt signalling is critical for synaptic function and cognition[70,71].

Targeted analysis of genes implicated in neurotransmitter transport and release highlights the previously identified genes *NSG1* and *DLG2*, but not *SV2A*, *SV2B* or *SV2C*

Whole genome sequencing permits focused analysis of identified sets of variants, in addition to unbiased analysis over the whole genome. Furthermore, by testing only a small subset of variants, we can ameliorate the effect of multiple hypothesis testing, thus increasing the power of statistical analysis. We conducted targeted analysis of common and rare variants on a set of 294 genes implicated in neurotransmitter transport and release (Table S5). These genes (SYNAPTIC) were identified based on their Gene Ontology[72] terms and they included *SV2A* and its paralogs, *SV2B* and *SV2C*. In addition, we tested all 402 high-confidence (i.e. "green") genes (EPILEPSY) in the genetic epilepsy syndromes panel v1.35 provided by Genomics England[73] (Table S5).

In the case of SYNAPTIC genes, which harboured ~51K common variants, we found evidence of association with response to LEV in the previously identified gene *NSG1* at a FWER<5% or <10%, depending on the methodology used for calculating the effective number of independent tests (Table S5). Furthermore, when examining the ~2.6K rare variants found in SYNAPTIC genes, *DLG2* was found associated with response to LEV at a FWER<0.1% (Table S5). This is not surprising, since *DLG2* was also found associated with LEV response in the previously conducted gene-based tests. We did not find any evidence of association between LEV response and *SV2A* or its paralogs, *SV2B* and *SV2C*, using either SNP- or gene-based tests. Furthermore, we did not find any evidence of association between ~66K common variants in EPILEPSY genes and response to LEV or between ~3.7K rare variants in the same genes and response to LEV at a FWER<10% (Table S5).

DISCUSSION

Although the anticonvulsant properties of the prominent antiepileptic drug LEV have been linked to the activity levels of its molecular target, the synaptic glycoprotein SV2A[21,22], targeted sequencing did not reveal any associations between common[24] or rare[23] variation in this gene and LEV efficacy. This leaves open the question as to whether genetic variation is a component of response variability and, if so, the identity of the genomic variants

underlying clinical response to LEV. In the present study, we followed a whole genome sequencing approach in an unbiased search of genomic polymorphisms that underlie clinical response to this drug.

According to one possible hypothesis for explaining variability in drug response, one or more common polymorphisms occur with different frequencies between responders and non-responders. Our analysis indicates that common polymorphisms in genes *NSG1*, *HDC, MDGA2*, *RASGEF1C* and *SPNS3* collectively predict clinical response to LEV in our cohort with overall accuracy ~91%. These genes are attractive candidates, since the first four are potentially implicated in synaptic neurotransmission, while the fourth is a transmembrane transporter protein homologous to *SV2A*.

A second hypothesis asserts that multiple rare variants act synergistically to influence a patient's response to the drug. Our analysis showed that groups of low-frequency variants in genes *PRKCB*, *DLG2*, *FILIP1* and *SEMA6D*, and in pathways involving *PRKCB* (and *LRP6*) demonstrate significant associations with the response/non-response phenotype.

From a neurophysiological perspective, there are three major, not mutually exclusive hypotheses for explaining pharmaco-resistant epilepsy[74]. First, the drug target hypothesis postulates that alterations in the activity of the molecular target of the drug (e.g. due to genomic polymorphisms coding for the drug-target binding site) result in reduced drug efficacy. Our analysis did not provide any evidence that *SV2A* or its paralogs (*SV2B* and *SV2C*) are associated with response to LEV, in agreement with previous studies[23,24]

Second, the drug transporter hypothesis states that reduced efficacy of antiepileptic drugs are due to low concentration of the drug at its target site due to over-active efflux drug transporters. A common intronic polymorphism in *SPNS3*, a gene homologous to *SV2A*, may be of interest in relation to this hypothesis. Both *SPNS3* and *SV2A* (and its paralogs) are structurally similar to the solute carrier family 22 (SLC22), a large family of transmembrane drug transporters. It should, however, be emphasised that homology (as established through structural similarity) is not definitive proof of biological relevance.

Finally, the intrinsic severity hypothesis postulates that severe epilepsy (manifested, for example, as high-frequency seizures) is linked to reduced response to antiepileptic drugs. Neurophysiological processes that are proposed to underlie the severity of epilepsy include neuroinflammation, aberrations in synaptic neurotransmission, and restructuring of neural networks[75]. Our analysis has identified common and low-frequency polymorphisms in genes and pathways, which are putatively related to these processes; for example, genes *HDC*, *NSG1*, *MDGA2*, *RASGEF1C*, *PRKCB* and *DLG2* (synaptic neurotransmission) and genes *FILIP1* and *SEMA6D* (restructuring of neural networks).

Whilst highlighting the approaches now available through the advent of NGS technologies, the findings in the present study need independent replication and potentially functional validation to confirm their role in determining response to LEV. Furthermore, we expect that the rapidly decreasing cost of WGS will allow conducting

similar studies with a larger sample size in the near future. Nevertheless, our approach of using extremes of response is a pragmatic way to derive hypotheses for experimental testing. It is interesting to postulate what the remaining factors are that determine response to LEV. Drug response is likely to be a complex interaction of many factors, including interacting genetic factors, which should be explored through polygenic risk score analysis and integrative analysis of multiple data modalities utilizing machine learning approaches.

In summary, we have identified common and low-frequency variants in genes and pathways, which may influence clinical response to LEV in a cohort of 99 patients with epilepsy. We conclude that whole genome sequencing can be a useful approach for investigating the genomic correlates of pharmaco-resistant epilepsy.

FUNDING

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford. This work was supported by Epilepsy Society, UK. UCB Pharma (Brussels, Belgium) funded the sequencing and bioinformatics work and provided input into the analytical approaches used. Part of this work was undertaken at University College London Hospitals, which received a proportion of funding from the NIHR Biomedical Research Centres funding scheme. S Balestrini was supported by the Muir Maxwell Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This publication presents independent research commissioned by the Health Innovation Challenge Fund (R6-388 / WT 100127), a parallel funding partnership between the Wellcome Trust and the Department of Health. The views expressed in this publication are those of the authors and not necessarily those of the Wellcome Trust or the Department of Health.

ACKNOWLEDGMENTS

The authors would like to thank Dr Chris Spencer of Genomics PLC Oxford for critically reviewing an earlier version of the paper.

REFERENCES

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351 (2016).
2. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31–46 (2010).
3. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47**, 717–726 (2015).
4. Ashley, E. A. Towards precision medicine. *Nat Rev Genet* **17**, 507–522 (2016).
5. Choi, J., Tantisira, K. G. & Duan, Q. L. Whole genome sequencing identifies high-impact variants in well-known pharmacogenomic genes. *Pharmacogenomics J* (2018).
6. Chua, E. W. *et al.* Exome sequencing and array-based comparative genomic hybridisation analysis of preferential 6-methylmercaptopurine producers. *Pharmacogenomics J* **15**, 414–421 (2015).
7. Katsila, T. & Patrinos, G. P. Whole genome sequencing in pharmacogenomics. *Front Pharmacol* **6**, 61 (2015).
8. Mak, A. C. Y. *et al.* Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma. *Am J Respir Crit Care Med* **197**, 1552–1564 (2018).
9. Mizzi, C. *et al.* Personalized pharmacogenomics profiling using whole-genome sequencing. *Pharmacogenomics* **15**, 1223–1234 (2014).
10. Price, M. J. *et al.* FIRST PHARMACOGENOMIC ANALYSIS USING WHOLE EXOME SEQUENCING TO IDENTIFY NOVEL GENETIC DETERMINANTS OF CLOPIDOGREL RESPONSE VARIABILITY: RESULTS OF THE GENOTYPE INFORMATION AND FUNCTIONAL TESTING (GIFT) EXOME STUDY. *J Am Coll Cardiol* **59**, E9 (2018).
11. Weeke, P. *et al.* Exome sequencing implicates an increased burden of rare potassium channel variants in the risk of drug-induced long QT interval syndrome. *J Am Coll Cardiol* **63**, 1430–1437 (2014).
12. Yang, G. *et al.* SIRT1/HERC4 Locus Associated With Bisphosphonate-Induced Osteonecrosis of the Jaw: An Exome-Wide Association Analysis. *J Bone Min. Res* **33**, 91–98 (2018).
13. Lynch, B. A. *et al.* The synaptic vesicle protein SV2A is the binding site for the antiepileptic drug levetiracetam. *Proc Natl Acad Sci U A* **101**, 9861–9866 (2004).
14. Bajjalieh, S. M., Frantz, G. D., Weimann, J. M., McConnell, S. K. & Scheller, R. H. Differential expression of synaptic vesicle protein 2 (SV2) isoforms. *J Neurosci* **14**, 5223–5235 (1994).
15. Chang, W.-P. & Südhof, T. C. SV2 renders primed synaptic vesicles competent for Ca2+ -induced exocytosis. *J Neurosci* **29**, 883–897 (2009).
16. Crowder, K. M. *et al.* Abnormal neurotransmission in mice lacking synaptic vesicle protein 2A (SV2A). *Proc Natl Acad Sci U A* **96**, 15268–15273 (1999).
17. Janz, R., Goda, Y., Geppert, M., Missler, M. & Südhof, T. C. SV2A and SV2B function as redundant Ca2+ regulators in neurotransmitter release. *Neuron* **24**, 1003–1016 (1999).
18. Vogl, C., Mochida, S., Wolff, C., Whalley, B. J. & Stephens, G. J. The synaptic vesicle glycoprotein 2A ligand levetiracetam inhibits presynaptic Ca2+ channels through an intracellular pathway. *Mol Pharmacol* **82**, 199–208 (2012).
19. Harada, S. *et al.* Inhibition of Ca(2+)-regulated exocytosis by levetiracetam, a ligand for SV2A, in antral mucous cells of guinea pigs. *Eur J Pharmacol* **721**, 185–192 (2013).
20. Nowack, A. *et al.* Levetiracetam reverses synaptic deficits produced by overexpression of SV2A. *PLoS One* **6**, e29560 (2011).

21. Kaminski, R. M. *et al.* Proepileptic phenotype of SV2A-deficient mice is associated with reduced anticonvulsant efficacy of levetiracetam. *Epilepsia* **50**, 1729–1740 (2009).

22. de Groot, M., Aronica, E., Heimans, J. J. & Reijneveld, J. C. Synaptic vesicle protein 2A predicts response to levetiracetam in patients with glioma. *Neurology* **77**, 532–539 (2011).

23. Dibbens, L. M. *et al.* Rare protein sequence variation in SV2A gene does not affect response to levetiracetam. *Epilepsy Res* **101**, 277–279 (2012).

24. Lynch, J. M. *et al.* No major role of common SV2A variation for predisposition or levetiracetam response in epilepsy. *Epilepsy Res* **83**, 44–51 (2009).

25. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **26**, 589–595 (2010).

26. Picard Tools - By Broad Institute. Available at: https://broadinstitute.github.io/picard/. (Accessed: 5th April 2019)

27. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

28. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

29. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887–893 (2005).

30. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

31. ARGOS. Available at: http://www.cibiv.at/~niko/argos/#tracks. (Accessed: 31st May 2019)

32. UCSC Genome Browser Home. Available at: https://genome.ucsc.edu/index.html. (Accessed: 31st May 2019)

33. Ingenuity Variant Analysis. *QIAGEN Bioinformatics*

34. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).

35. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

36. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210

37. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

38. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).

39. Galwey, N. W. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genet Epidemiol* **33**, 559–568 (2009).

40. Cheverud, J. M. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87**, 52–58 (2001).

41. Nyholt, D. R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765–769 (2004).

42. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).

43. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* **32**, 361–369 (2008).

44. Guisan, A. & Zimmermann, N. E. Predictive habitat distribution models in ecology. *Ecol Modell* **135**, 147–186 (2000).

45. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

46. The Human Protein Atlas. Available at: https://www.proteinatlas.org/. (Accessed: 3rd April 2019)

47. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1-1.30.33 (2016).

48. Brambilla, R. *et al.* A role for the Ras signalling pathway in synaptic transmission and long-term memory. *Nature* **390**, 281–286 (1997).

49. Ye, B. *et al.* GRASP-1: A Neuronal RasGEF Associated with the AMPA Receptor/GRIP Complex. *Neuron* **26**, 603–617 (2000).

50. Stornetta, R. L. & Zhu, J. J. Ras and Rap Signaling in Synaptic Plasticity and Mental Disorders. *Neurosci. Rev. J. Bringing Neurobiol. Neurol. Psychiatry* **17**, 54–78 (2011).

51. Pettem, K. L., Yokomaku, D., Takahashi, H., Ge, Y. & Craig, A. M. Interaction between autism-linked MDGAs and neuroligins suppresses inhibitory synapse development. *J Cell Biol* **200**, 321–336 (2013).

52. Lee, K. *et al.* MDGAs interact selectively with neuroligin-2 but not other neuroligins to regulate inhibitory synapse development. *Proc. Natl. Acad. Sci.* **110**, 336–341 (2013).

53. Bhowmik, M., Khanam, R. & Vohora, D. Histamine H3 receptor antagonists in relation to epilepsy and neurodegeneration: a systemic consideration of recent progress and perspectives. *Br J Pharmacol* **167**, 1398–1414 (2012).

54. Baldan, L. C. *et al.* Histidine decarboxylase deficiency causes tourette syndrome: parallel findings in humans and mice. *Neuron* **81**, 77–90 (2014).

55. Awaad, Y., Michon, A. M. & Minarik, S. Use of levetiracetam to treat tics in children and adolescents with Tourette syndrome. *Mov Disord* **20**, 714–718 (2005).

56. Hedderick, E. F., Morris, C. M. & Singer, H. S. Double-blind, crossover study of clonidine and levetiracetam in Tourette syndrome. *Pediatr Neurol* **40**, 420–425 (2009).

57. Martínez-Granero, M. A., García-Pérez, A. & Montañes, F. Levetiracetam as an alternative therapy for Tourette syndrome. *Neuropsychiatr Treat* **6**, 309–316 (2010).

58. Jacobsson, J. A., Haitina, T., Lindblom, J. & Fredriksson, R. Identification of six putative human transporters with structural similarity to the drug transporter SLC22 family. *Genomics* **90**, 595–609 (2007).

59. Perland, E., Bagchi, S., Klaesson, A. & Fredriksson, R. Characteristics of 29 novel atypical solute carriers of major facilitator superfamily type: evolutionary conservation, predicted structure and neuronal co-expression. *Open Biol* **7**, (2017).

60. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649–D655 (2018).

61. Rogawski, M. A. AMPA receptors as a molecular target in epilepsy therapy. *Acta Neurol Scand Suppl* 9–18 (2013).

62. Carunchio, I., Pieri, M., Ciotti, M. T., Albo, F. & Zona, C. Modulation of AMPA receptors in cultured cortical neurons induced by the antiepileptic drug levetiracetam. *Epilepsia* **48**, 654–662 (2007).

63. Lee, C.-Y., Chen, C.-C. & Liou, H.-H. Levetiracetam inhibits glutamate transmission through presynaptic P/Q-type calcium channels on the granule cells of the dentate gyrus. *Br J Pharmacol* **158**, 1753–1762 (2009).

64. Ghasemi, M. & Schachter, S. C. The NMDA receptor complex as a therapeutic target in epilepsy: a review. *Epilepsy Behav* **22**, 617–640 (2011).

65. Birnstiel, S., Wülfert, E. & Beck, S. G. Levetiracetam (ucb LO59) affects in vitro models of epilepsy in CA3 pyramidal neurons without altering normal synaptic transmission. *Naunyn Schmiedebergs Arch Pharmacol* **356**, 611–618 (1997).

66. Rappaport, N. *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* **45**, D877–D887 (2017).

67. Yu, A. *et al.* Association of Dishevelled with the clathrin AP-2 adaptor is required for Frizzled endocytosis and planar cell polarity signaling. *Dev Cell* **12**, 129–141 (2007).

68. Yao, J., Nowack, A., Kensel-Hammes, P., Gardner, R. G. & Bajjalieh, S. M. Cotrafficking of SV2 and synaptotagmin at the synapse. *J Neurosci* **30**, 5569–5578 (2010).

69. Matveeva, E. A., Vanaman, T. C., Whiteheart, S. W. & Slevin, J. T. Levetiracetam prevents kindling-induced asymmetric accumulation of hippocampal 7S SNARE complexes. *Epilepsia* **49**, 1749–1758 (2008).

70. Liu, C.-C. *et al.* Deficiency in LRP6-mediated Wnt Signaling Contributes to Synaptic Abnormalities and Amyloid Pathology in Alzheimer's Disease. *Neuron* **84**, 63–77 (2014).

71. Zhou, C.-J., Borello, U., Rubenstein, J. L. R. & Pleasure, S. J. Neuronal production and precursor proliferation defects in the neocortex of mice with loss of function in the canonical Wnt signaling pathway. *Neuroscience* **142**, 1119–1131 (2006).

72. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

73. Genetic epilepsy syndromes (Version 1.35). Available at: https://panelapp.genomicsengland.co.uk/panels/402/. (Accessed: 8th April 2019)

74. Schmidt, D. & Löscher, W. New developments in antiepileptic drug resistance: an integrative view. *Epilepsy Curr* **9**, 47–52 (2009).

75. Mirza, N., Vasieva, O., Marson, A. G. & Pirmohamed, M. Exploring the genomic basis of pharmacoresistance in epilepsy: an integrative analysis of large-scale gene expression profiling studies on brain tissue from epilepsy surgery. *Hum Mol Genet* **20**, 4381–4394 (2011).

FIGURES

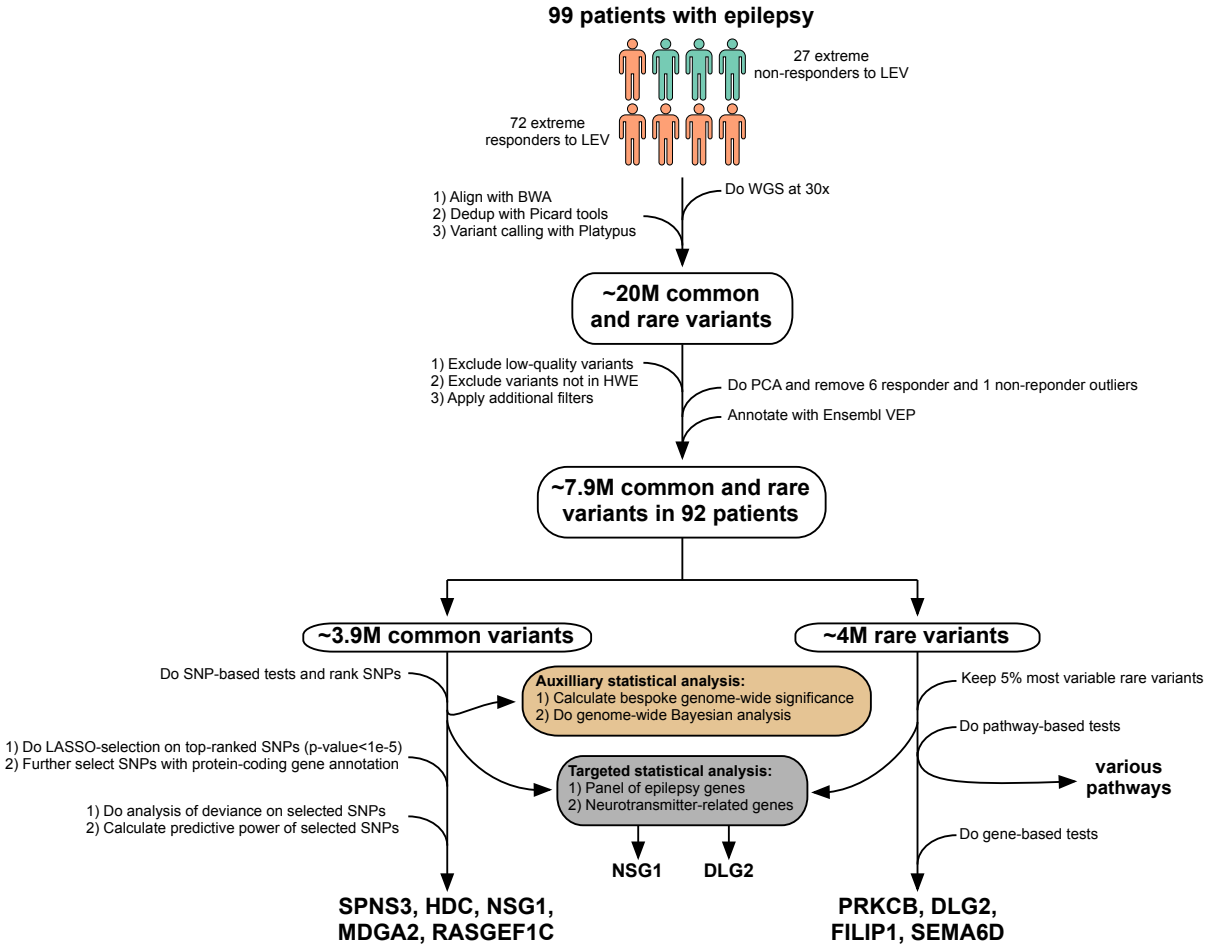

Figure 1: Overview of the study. We recruited 99 patients with epilepsy (72 extreme responders and 27 extreme non-responders to LEV). After performing WGS, alignment and variant calling, we identified ~20M unfiltered variants. After filtering across variants and samples, we ended up with ~3.9M common (MAF>5%) variants and ~4M low-frequency and rare (MAF<5%) variants across 92 patients. Subsequently, we calculated p-values for each common variant using a two-tailed Fisher's exact test. In the next step, we performed penalised logistic regression (LASSO) on all common variants with p-value less than the suggestive genome-wide significance threshold of 10^{-5} (n=23 variants; Supplementary Table S1). This was followed by further selecting variants with protein-coding gene annotation. In the last step, we performed analysis of deviance on the finally selected variants (n=5 variants) and we calculated their collective predictive accuracy using a cross-validation approach. For completeness, we also conducted additional auxiliary statistical analyses on the common variants (see Supplementary Material). In the case of low-frequency and rare variants, we focused on the top 5% most variable variants in our cohort and, by performing gene- and pathway-based tests on these, we identified associations between several genes or pathways and clinical response to LEV. Finally, for both common and low-frequency/rare variants, we conducted targeted analysis on a panel of epilepsy genes and on genes related to neurotransmitter transport and release.

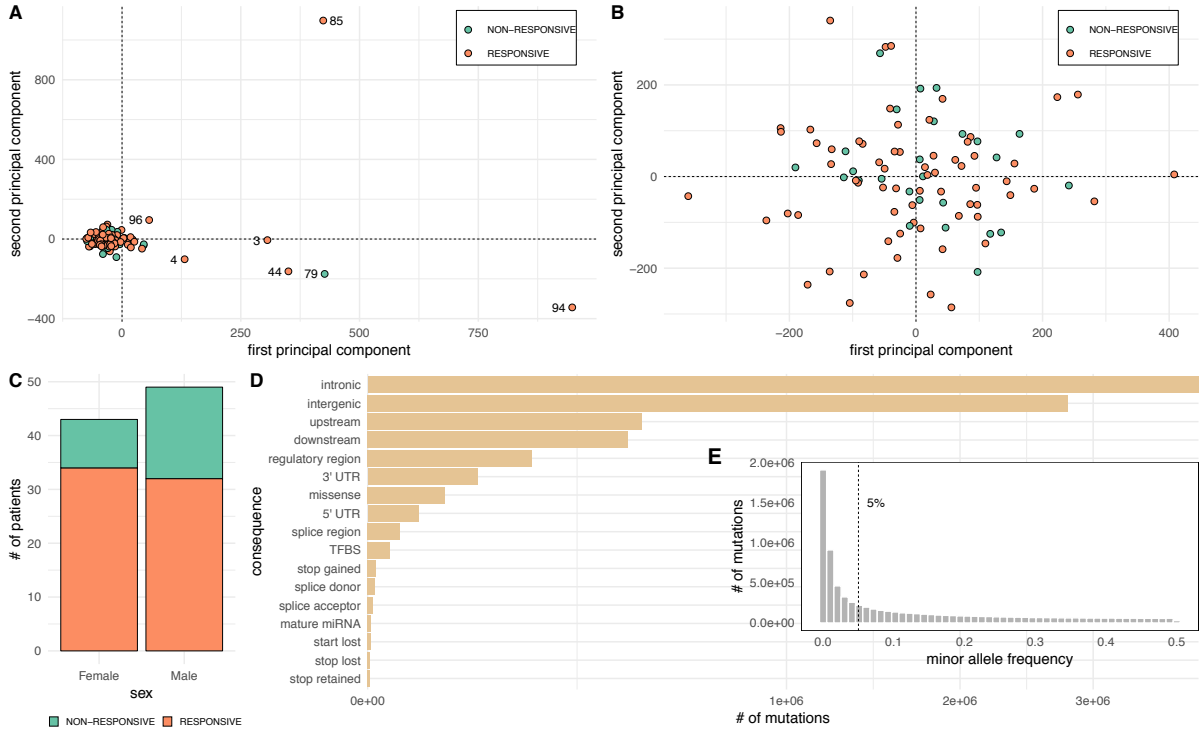

Figure 2: Overview of the WGS data from 99 extreme responders and non-responders to LEV. A) Principal component analysis (PCA) of the matrix of genotypes across all samples and variants. The first two principal components are illustrated. Seven samples appear as outliers. B) Repeating the PCA after removing the seven outliers identified in (A) indicates lack of any stratification (e.g. due to population structure) in the data. C) Number of male and female subjects among responders and non-responders to LEV. There are almost twice as many non-responders among 49 males (n=17), as among 43 females (n=9) in the data. A two-tailed Fisher's exact test of independence indicates that this difference is not statistically significant (odds ratio: 1.99; 95% CI: 0.72-5.86; p-value: 0.17). D) Consequences of all variants identified by WGS. Most variants are intronic, intergenic, or located immediately upstream or downstream of protein-coding genes. E) Minor allele frequencies (MAF) of all variants identified by WGS. A cut-off of 5% was chosen to discriminate between common and low-frequency or rare variants.

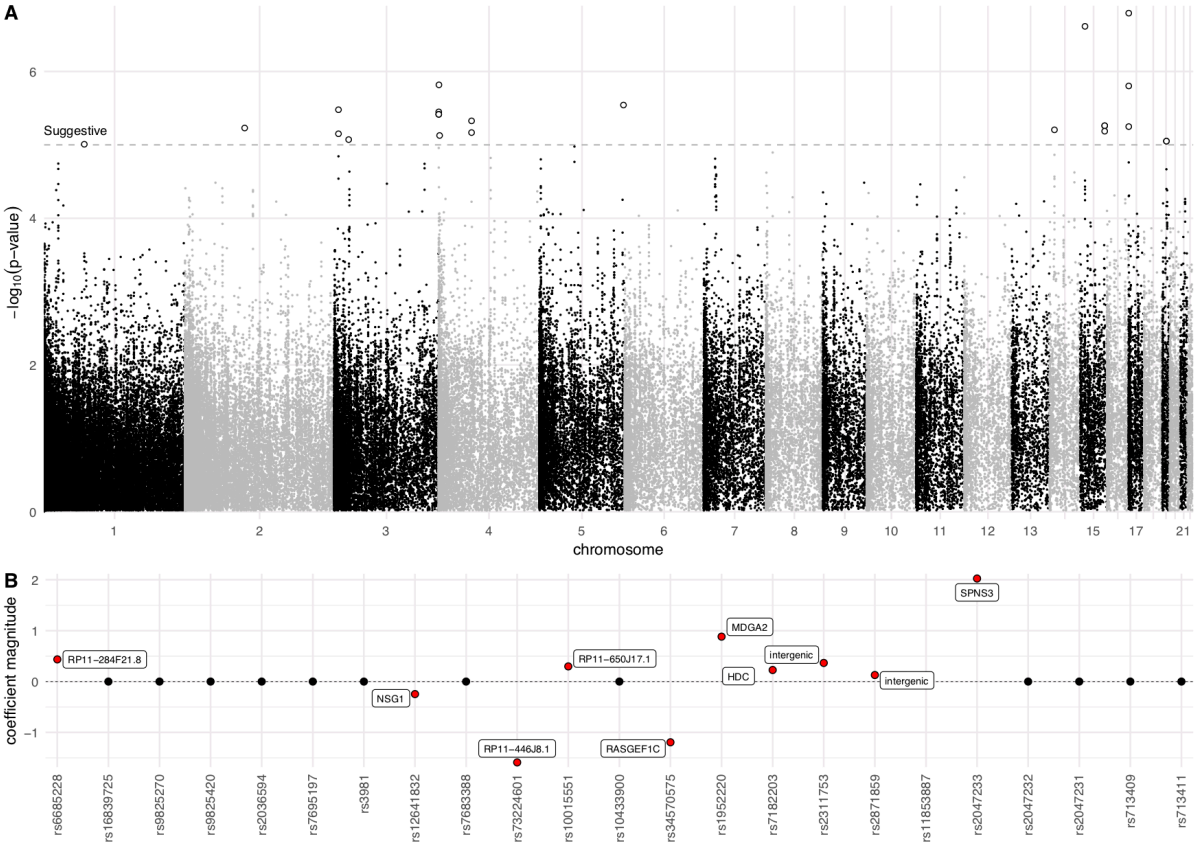

Figure 3: Genome-wide selection of a minimal set of common (MAF>5%) variants with maximal predictive power. A) Manhattan plot summarising SNP-based tests using a two-tailed Fisher's exact test of independence. All variants with p-values below a suggestive significance threshold of 10^{-5} are indicated with white circles (n=23). B) Summary of variable selection using penalised logistic regression (LASSO). All SNPs crossing the suggestive genome-wide significance threshold in (A) were used as predictors. Variants selected through this process have non-zero regression coefficients (red dots). Among these, the variants with protein-coding gene annotation (i.e. *SPNS3*, *HDC*, *MDGA2*, *NSG1* and *RASGEF1C*) were selected for further analysis.

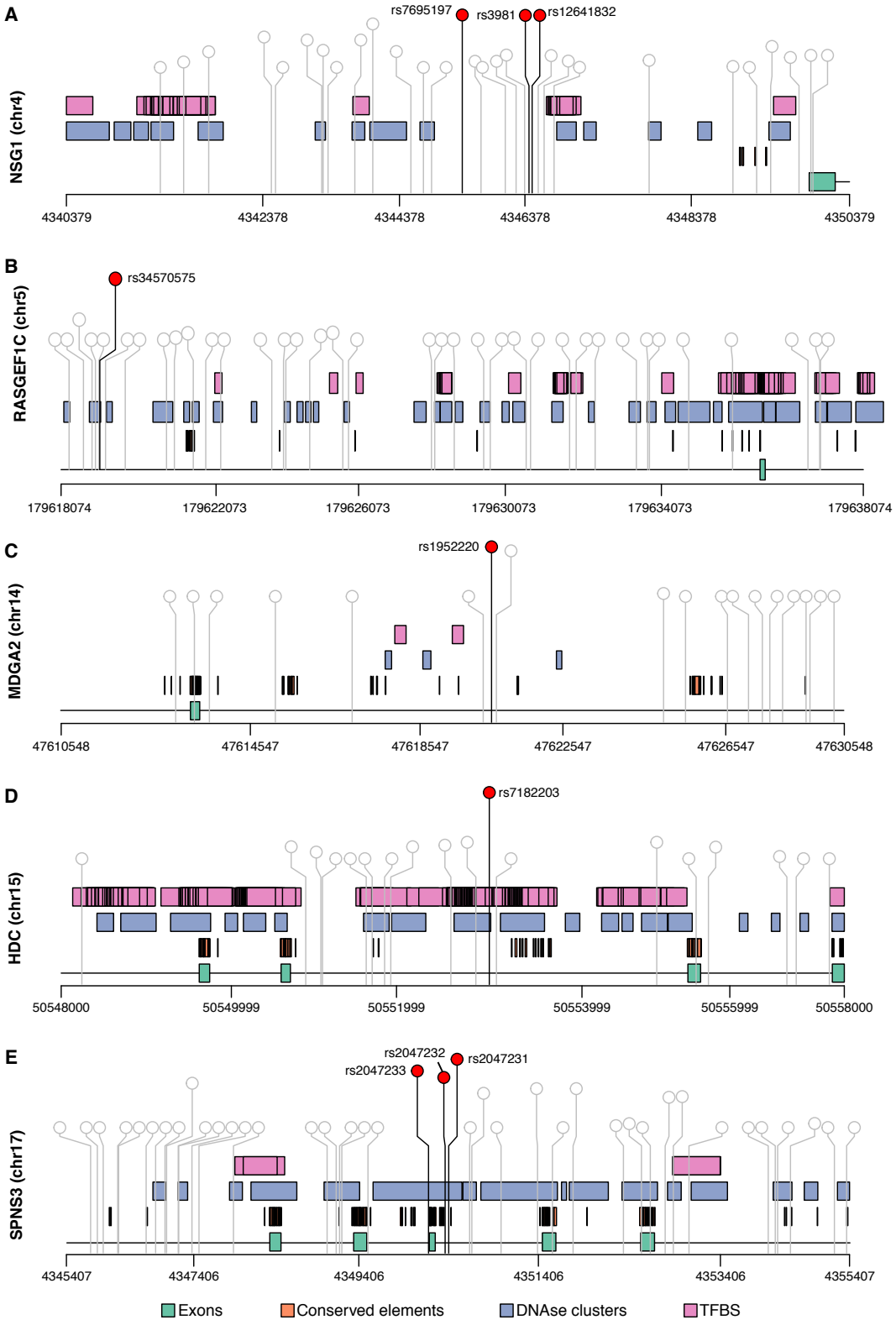

Figure 4: Local genomic structure near the most significant SNPs in genes *NSG1*, *RASGEF1C*, *MDGA2*, *HDC* and *SPNS3*. Common variants in these genes are strong predictors of clinical response to LEV in our cohort. SNPs crossing the suggestive genome-wide significance threshold of 10^{-5} are indicated in red.

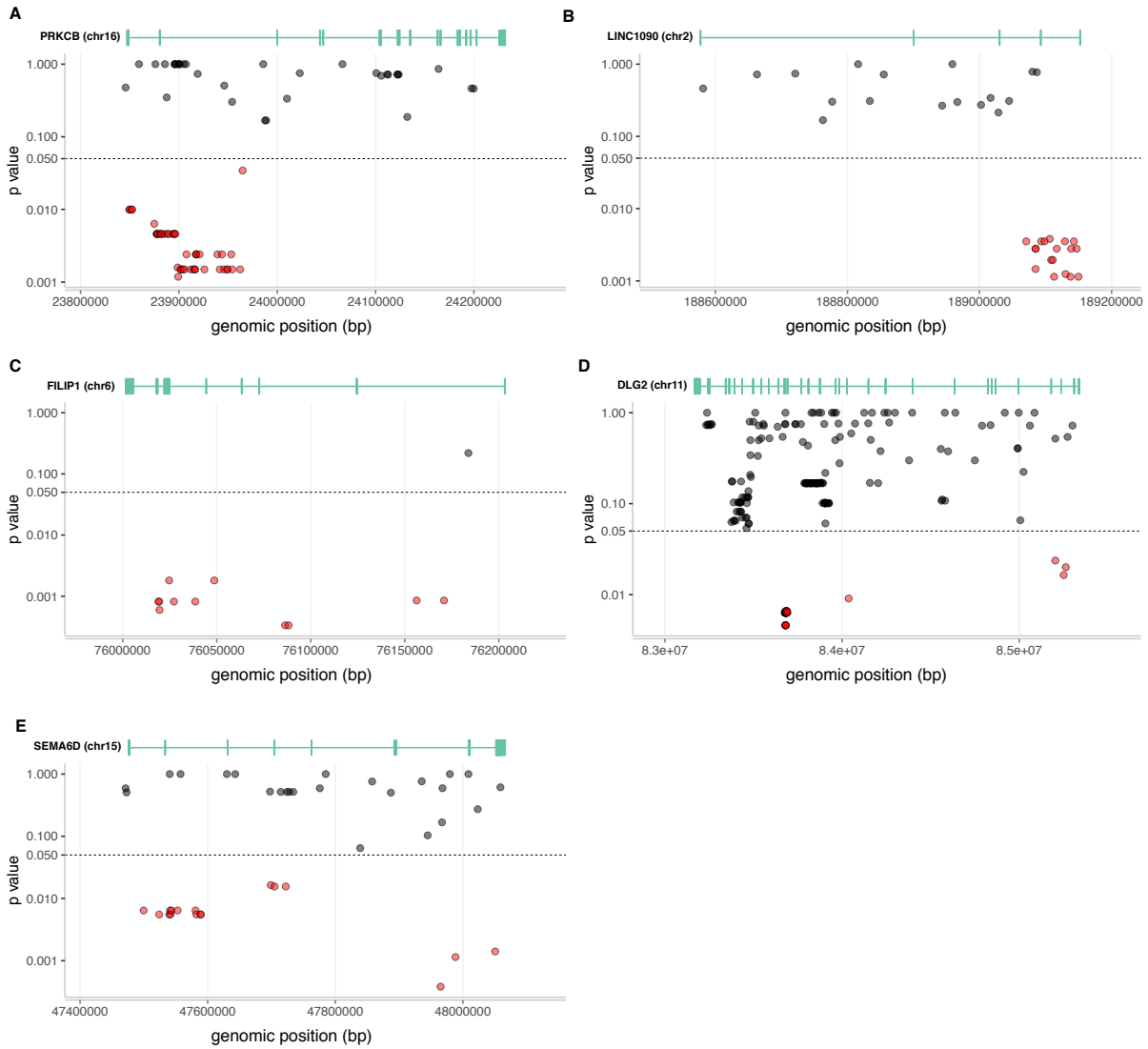

Figure 5: Distribution of rare variants along genes *PRKCB*, *LINC01090*, *FILIP1*, *DLG2* and *SEMA6D*. Based on gene-based tests, these genes are significantly associated with response to LEV at a FWER<10%. Rare variants with p-values below 5% (calculated using a two-tailed Fisher's exact test of independence) are indicated in red.

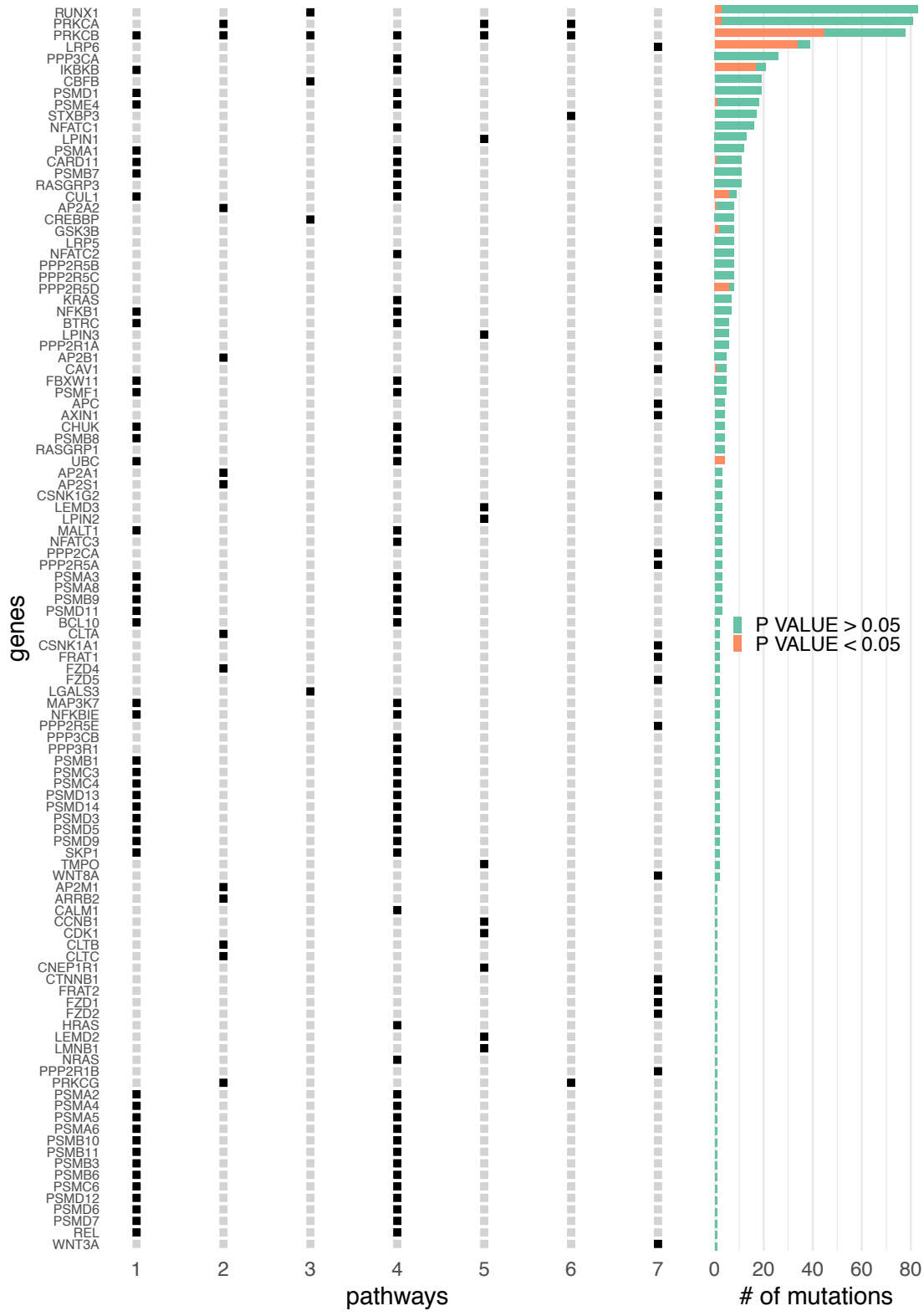

Figure 6: Summary of the results from Reactome pathway-based tests. In the left panel, pathways are as follows: *1-Activation of NF-kappaB in B cells; 2-WNT5A-dependent internalization of FZD4; 3-RUNX1 regulates transcription of genes involved in differentiation of myeloid cells; 4-Downstream signalling events of B Cell Receptor (BCR); 5-Depolymerisation of the Nuclear Lamina; 6-Disinhibition of SNARE formation; 7-Disassembly of the*

destruction complex and recruitment of AXIN to the membrane. Among all genes harbouring the highest number of rare variants (*RUNX1*, *PRKCA*, *PRKCB*, *LRP6*), *PRKCB* and *LRP6* have the highest proportion of rare variants with p-values less than 5% (right panel). *PRKCB* is participating in all but one pathway, with the remaining genes participating in only 1, 2 or 3 pathways (left panel).

SUPPORTING INFORMATION LEGENDS

Figure S1: Auxiliary statistical analysis of common (MAF>5%) variants across the whole genome. A) Manhattan plot summarising SNP-based tests using a two-tailed Fisher's exact test of independence. Genome-wide significance thresholds (calculated using four different methodologies) are indicated, including a suggestive significance threshold of 10^{-5}. Two variants (in red) cross the three least conservative thresholds and they could be considered statistically significant with respect to these thresholds. B) Manhattan plot summarising the Bayesian analysis of single SNPs. We assumed a prior retrospective probability of association (rPPA) equal to $\pi=10^{-4}$. Variants with rPPA values above 50% are more likely than not to be associated with response to the drug.

Table S1: Summary of all common variants with p-values less than a suggestive genome-wide significance threshold of 10^{-5} (n=23). Among these, those selected by the LASSO are indicated in green and orange (n=10; also see Figure 3B). The variants indicated in green (n=5) have protein-coding genes annotations and they were selected for further analysis. Those indicated in orange (n=5) have non-protein-coding gene annotations or are annotated as *intergenic* and they were not selected for further analysis. In order to avoid division by zero in the calculation of odds ratios, the matrix of counts for each variant was pre-processed using Lidstone smoothing with pseudo-count parameter equal to 1 (i.e. add-one smoothing).

Table S2: Analysis of deviance using seven different logistic regression models of increasing complexity. The BASIC model includes only sex and the intercept as predictors. The FULL model includes in addition five SNPs from genes *SPNS3*, *HDC*, *NSG1, RASGEF1C* and *MDGA2*, which were previously selected using penalised logistic regression. Intermediate models include only sex, the intercept and the SNP harboured by the indicated gene. DF: degrees of freedom

Table S3: Summary of gene-based tests. Only genes with FWER<10% are shown. The low-frequency variants harboured by these genes are listed in the second spreadsheet.

Table S4: Summary of Reactome pathway-based tests. Only pathways with FWER<10% are shown. The rare variants harboured by genes in these pathways are listed in the second spreadsheet.

Table S5: Summary of results from the targeted analysis. We list the results from SNP- and gene-based tests on SYNAPTIC and EPILEPSY genes. A complete list of genes in each of these two groups is also provided in the last spreadsheet.

Table S6: Summary of results from the auxiliary statistical analysis on common variants. Only variants with a p-value below a suggestive genome-wide significance threshold of 10^{-5} are presented. P-values were corrected for

multiplicity using Sidak's method and four different estimates of the effective number of independent tests, as indicated. The results of Bayesian analysis for different choices of the priors π (10^{-4}, 10^{-5} and 10^{-6}) and (a, b, c) (flat or empirical) are also given. rPPA: retrospective probability of association. FWER: family-wise error rate.