

***In trans* variant calling reveals enrichment for compound heterozygous variants in genes involved in neuronal development and growth.**

Authors: Authors: Allison J. Cox^{1,2}, Fillan Grady³, Gabriel Velez^{3,4}, Vinit B. Mahajan^{4,5}, Polly J. Ferguson¹, Andrew Kitchen⁶, Benjamin W. Darbro¹, and Alexander G. Bassuk^{1,2*}

¹*Department of Pediatrics, The University of Iowa, Iowa City, IA*

²*Interdisciplinary Graduate Program in Genetics, The University of Iowa, Iowa City, IA*

³*Medical Scientist Training Program, University of Iowa, Iowa City, IA*

⁴*Omics Laboratory, Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA*

⁵*Palo Alto Veterans Administration, Palo Alto, CA*

⁶*Department of Anthropology, The University of Iowa, Iowa City, IA*

*Correspondence to: Alexander Bassuk, MD, PhD, alexander-bassuk@uiowa.edu

***In trans* variant calling reveals enrichment for compound heterozygous variants in genes involved in neuronal development and growth.**

Abstract

Compound heterozygotes occur when different mutations at the same locus on both maternal and paternal chromosomes produce a recessive trait. Here we present the tool VarCount for the quantification of mutations at the individual level. We used VarCount to characterize compound heterozygous coding variants in patients with epileptic encephalopathy and in the 1000 genomes participants. The Epi4k data contains variants identified by whole exome sequencing in patients with either Lennox-Gastaut Syndrome (LGS) or Infantile Spasms (IS), as well as their parents. We queried the Epi4k dataset (264 trios) and the phased 1000 genomes data (2504 participants) for recessive variants. To assess enrichment, transcript counts were compared between the Epi4k and 1000 genomes participants using minor allele frequency (MAF) cutoffs of 0.5% and 1.0%, and including all ancestries or only probands of European ancestry. In the Epi4k participants, we found enrichment for rare, compound heterozygous mutations in six genes, including three involved in neuronal growth and development – *PRTG* ($p=0.00086$, 1% MAF, combined ancestries), *TNC* ($p=0.0221\%$ MAF, combined ancestries), and *MACFI* ($p=0.0245$, 0.5% MAF, EU ancestry). Due the total number of transcripts considered in these analyses, the enrichment detected was not significant after correction for multiple testing and higher powered or prospective studies are necessary to validate the candidacy of these genes. However, *PRTG*, *TNC*, and *MACFI* are potential novel recessive epilepsy genes and our results highlight that compound heterozygous mutations should be considered in sporadic epilepsy.

Keywords: epilepsy; genetics; compound heterozygous, bioinformatics

Introduction

Using the premise that effective variants are in linkage disequilibrium (LD) with common polymorphisms and haplotypes, linkage and association studies have identified genes involved in the development of traits and pathologies. Upon their identification, the regions flanking associated markers are sequenced to find the linked, penetrant mutation. However, rare variants are often not detectable using LD-based methods. This problem has been alleviated by recent advances in next-generation sequencing (NGS), and the detection of highly penetrant rare variants associated with disease has reduced the heritability gap for such diseases as autism, Crohn's disease, and osteoporosis (Bomba, Walter, & Soranzo, 2017; Kosmicki, Churchhouse, Rivas, & Neale, 2016). Despite these advances, for most traits and complex disorders the underlying genes and mutations remain elusive.

Recessive traits are caused by mutations in both copies of a gene. The mutations may be homozygous, i.e. identical, or compound heterozygous. Compound heterozygous (CH) mutations are two different mutations in a gene on opposite alleles of a chromosome and it is speculated that compound heterozygous mutations account for many recessive diseases (Li et al., 2010; Sanjak, Long, & Thornton, 2017). Lack of detection of CH may explain a significant portion of missing heritability for all phenotypes (Li et al., 2010; Sanjak et al., 2017; Zhong, Karssen, Kayser, & Liu, 2016). Association studies using polymorphisms are LD-based and recent association studies using rare variants compare total variant burden between cases and controls to account for the contributions of multiple alleles at a locus to phenotype. Importantly, because LD-based studies require recessive mutations to be on the same genetic background and total variant burden analyses are not allele-specific, neither discerns between dominant and recessive models of inheritance.

Burden tests may account for compound heterozygosity if the variants are allocated to one of the two alleles for a gene, i.e. phased. Relatively common variants may be phased assuming linkage to surrounding haplotypes; in families, rare variants are phased using parental genotypes. Once mutations are phased, it may be determined if an individual's mutations are on different chromosomes, and burden tests that aggregate using an indicator function (i.e. presence of qualifying variants) may assess enrichment for recessive variants.

Here we provide a publicly available tool, VarCount, that is user-friendly and effective for researchers seeking to quantify the presence or absence of a mutation or mutations in a gene at the individual level. VarCount is useful for the quantification of heterozygous, homozygous, or CH mutations per sample. We used VarCount to query the Epi4k (Epi et al., 2013) dataset for rare homozygous and CH mutations and found enrichment for rare, compound heterozygous mutations in six genes, including three involved in neuronal development or growth (*PRTG*, *TNC*, and *MACF1*). The variants in the 1000 genomes database are now phased (Genomes Project et al., 2010; Genomes Project et al., 2012; Genomes Project et al., 2015), and so genes may be queried for *in trans* combinations of variants. The Epi4k enrichment was identified in comparison to the 1000 genomes participants combining all ancestries and considering only individuals of European ancestry.

Materials and Methods

Processing of Epi4k vcf files:

The Epi4k data (Epi et al., 2013) were accessed by permission via the Database of Genotypes and Phenotypes (dbGaP Study Accession, phs000653.v2.p1). Individual vcf files were combined using the CombineVariants function in GATK (McKenna et al.,

2010). The vcf files were then annotated with minor allele frequencies (MAFs) from EVS (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: <http://evs.gs.washington.edu/EVS/>)), 1000 genomes and ExAC (Monkol Lek et al., 2015), and with information regarding the effect of each variant using SnpSift/SNPeff (Cingolani et al., 2012). The databases used for annotation were dbNSFP2.9 (for MAF and CADD score) and GRCh37.75 for protein effect prediction. SnpSift was used to remove any variants not inducing a protein-changing event (not “HIGH” or “MODERATE” impact) based on SNPeff annotation – this includes missense, nonsense, splice-site and insertion/deletion variants. Variants with quality flags and multiallelic variants, i.e. those with more than two known nucleotide values, were also removed. Variants remaining after filtering were cross-referenced with the 1000 genomes variants from the same MAF threshold to ensure that any variants removed from one dataset were removed from the other. The annotated vcf was used as input for VarCount. Ancestry for each exome was determined using LASER (Wang et al.) and this information was input to VarCount via the SampleInfo.txt file. Ancestry and phenotype information for each proband are described in Supplementary Table 1. In addition to the annotated vcf file, the parameters.txt and subjectinfo.txt (containing sex and ancestry information) were used as input. Within the parameters file, the following qualifications were selected: (1) counting at the transcript (rather than gene) level, (2) protein-changing effects, (3) MAF threshold of either 0.005 or 0.01, (4) all within-dataset and annotated (1000 genomes, ExAC and EVS) MAFs, and (5) either compound heterozygous or homozygous mutations. Analyses were run separately for the two MAFs and using all Epi4k probands (264) and only those of European ancestry (207). Because the variants were not phased, VarCount was used to query the vcf file for individuals with two or more mutations in each transcript. The output, a list of

counts for each transcript was then used to query the parental vcf files for genotype information to determine which sets of variants composed *in trans* combinations of mutations. Final counts were determined using parental genotype information. Custom python scripts were used to query for parental genotypes and to count true compound heterozygotes or homozygotes. *De novo* mutations were excluded in the determination of true *in trans* mutations.

Processing of 1000 genomes vcf files:

Vcf files for the 2504 participants in the 1000 genomes sequencing project (Genomes Project et al., 2015) were downloaded by chromosome from the 1000 genomes ftp site. To reduce input file size, the genomic regions for the hg19 mRNA transcripts were downloaded via UCSC's Table Browser and used to remove non-coding regions from the vcf files. Including all exons from UCSC allowed for a more conservative analysis, given that the Epi4k data were sequenced using various exome captures, which are not inclusive of all possible exons. The variants were annotated and filtered via the same steps as the Epi4k vcf file. Multi-allelic variants were also removed prior to analysis by VarCount. A diagram showing the steps involved in processing and analyzing the variant files is shown in (Figure 1).

VCF files were queried for homozygous and compound heterozygous mutations using VarCount. Because the variants in the 1000 genomes vcf files are phased, determining true compound heterozygotes is automatic using VarCount. In addition to the annotated vcf file, the parameters.txt and subjectinfo.txt (containing sex and ancestry information) were used as input. Within the parameters file, the same qualifications used in the Epi4k analysis were selected: Analyses were run for each of the two MAFs (0.5% and 1.0%) and for all 1000 genomes participants and using only

those of European (EUR) ancestry. The final output from analyses was for each MAF cutoff and for each population, counts for every transcript in which at least one individual harbored recessive variants.

Epi4k statistical analysis:

Using R statistical software, a Fisher's exact test was used to detect transcripts with significant differences in the proportion of individuals with homozygous or compound heterozygous variants between the Epi4k dataset and the 1000 genomes dataset. Odds ratios and p-values were calculated using the number of individuals with and without qualifying mutations in each superpopulation. Analyses were performed using all ancestries, and for only individuals of European ancestry. Both Bonferroni and Benjamini-Hochberg adjustments were used to determine significance thresholds after correction for multiple testing. The number of tests was based on the number of transcripts with at least one individual in either the Epi4k or 1000 genomes dataset with *in trans* coding variants with minor allele frequencies below the set threshold.

Structural modeling of PRTG:

The three-dimensional structure of Protogenin (PRTG) was modeled off the crystal structure of the human receptor protein tyrosine phosphatase sigma (PDB: 4PBX; 25.1% sequence identity) using MODELLER 9.14 (Webb & Sali, 2016). The resultant model superimposed with the template had an RMSD of 4.94 Å over 442 C α atoms. Charges and hydrogen atoms were added to the wild-type and mutant FGR models using PDB2PQR (Dolinsky, Nielsen, McCammon, & Baker, 2004). Electrostatic potentials were calculated using APBS (Konecny, Baker, & McCammon, 2012) as described previously (Cox et al., 2017; Moshfegh et al., 2016; Toral et al., 2017).

Protein and solvent dielectric constants were set to 2.0 and 78.0, respectively. All structural figures were generated by PyMOL(W).

Results

Varcount: Mutation Quantification at the Individual level:

Varcount is a free, open source tool useful for the quantification of heterozygous, homozygous, or compound heterozygous mutations per sample. Input variants may be phased or unphased. All python scripts and supporting files may be downloaded from Github at <https://github.com/GeneSleuth/VarCount>. Supporting files include the “parameters.txt” file where the user may select variant filters for mutation effect, minor allele frequency, and inheritance pattern (homozygous, compound heterozygous, one mutation or two mutations), and Sample filters based on information entered into the “SampleInfo file”. Input vcf files must be annotated with SNPSift/SNPEff using the dbNSFP and GRCh37/38 databases. A readme file with instructions is also provided. A flow diagram with the steps involved in processing of data is depicted in Figure 1.

[Figure 1 near here]

Compound heterozygous mutations in Epi4k probands reveal novel epilepsy genes:

We used VarCount to query the Epi4k dataset for rare homozygous and CH mutations. The Epi4k data are whole exome data from 264 trios with a child affected by epileptic encephalopathy, either infantile spasms (IS) or Lennox-Gastaut syndrome (LGS) (6). Counts were performed using individuals of all ancestries or just those of European ancestry (207/264). Individuals from the 1000 genomes study were used as controls. The individual counts and p-values for the analyses are listed in (Supplementary Tables 2-5). Including only rare variants (MAFs below 0.5% and 1.0%) determined enrichment

for compound heterozygous mutations in six genes. For combined ancestries, the six genes are in order of significance: *OSBP2*, *PRTG*, *ABCC11*, *MACF1*, *STAB1*, and *TNC*. *PRTG* and *TNC* were also highly ranked in the 1% MAF analysis, with one additional count for each transcript. Variants for all six genes are listed in Table 1. In our analysis of just individuals of European ancestry, *MACF1* was the most significantly enriched gene using a 0.5% MAF. The p-values indicated in Table 1 are for individual tests; there were no p-values significant after correction for multiple testing.

[Table 1 near here]

The variants for the three individuals with CH *PRTG* mutations are depicted in Figure 2A. Because of the concentration of mutations at position E104, we performed structural modeling to predict the pathogenicity of the *PRTG* mutations. The p.Glu104Gly and p.Glu104Asp mutations localize to the immunoglobulin (Ig)-like domain 1 (Fig. 2A). Ig-like domains are responsible for mediating protein-protein and protein-peptide interactions. The p.Glu104Gly disrupts a negative charge in the Ig-like 1 domain. This loss of charge may disrupt interactions with putative *PRTG*-binding partners (Fig. 2B).

[Figure 2 near here]

The *de novo* variants identified by Epi4K Consortium and the Epilepsy Phenome/Genome Project (Epi et al., 2013) in the nine probands with either *PRTG*, *TNC*, or *MACF1* recessive variants are described in Table 2. For the three patients with compound heterozygous *PRTG* variants, one patient harbors a *de novo* missense variant in *HSF2*, the second has a nonsense variant in *CELSRI*, and the third patient has two *de novo* variants – a missense in *Fam102A* and a 3'UTR variant in *USP42*. *De novo* mutations were only reported in one of the probands with *in trans* *TNC* variants – a

missense variant in *DIP2C* and a splice donor change in *IFT172*. All three patients with compound heterozygous variants in *MACF1* were reported to have *de novo* mutations. The first patient has a 5' and 3'UTR *de novo* variant in *FAM19A2* and *GLRA2*, respectively and the second patient also has a 3'UTR *de novo* change in the gene *LRRC8D* and a missense change in *SNX30*. One *de novo* variant was identified in the third proband in the gene *FAM227A*. Polyphen2 categories and CADD scores for each *de novo* variant as well as missense and loss-of-function constraint metric values for each gene (from ExAC) are also listed in Table 2. The z-score is a ratio of expected to identified missense variants in a particular gene, and pLI is a gene's probability of being loss-of-function intolerant. These constraint metrics are calculated using genomic data from controls without severe genetic diseases in the ExAC database (Monkol Lek et al., 2015).

[Table 2 near here]

Discussion:

Epileptic encephalopathies are a group of severe, early-onset seizure disorders with consistent EEG abnormalities that over time interfere with development and cause cognitive decline (Covanis, 2012). The Epi4k dataset contains exome sequence from 264 trios that include a proband with epileptic encephalopathy, either Lennox-Gastaut Syndrome (LGS) or Infantile Spasms (IS). LGS is characterized by frequent, mixed epileptic seizures that arise most frequently between the ages of 3 and 5 (Amrutkar and Riel-Romero, 2018). IS occurs during the first year of life and is cryptic in its presentation, with mild head bobbing and is often not detected until the seizures have caused significant neurological damage (Kossoff, 2010). IS often progress into LGS over time.

We developed a free and user-friendly tool, VarCount, to query vcf files for individuals harboring variants that qualify according to user specification. To test its function, we used VarCount to quantify rare, compound heterozygous mutations in probands from the Epi4k trio dataset and found enrichment for variants in six genes including *PRTG*, *TNC* and *MACF1*. *PRTG* codes for protogenin, a member of the immunoglobulin superfamily that is involved in axis elongation and neuronal growth during early vertebrate development (Toyoda, Nakamura, & Watanabe, 2005; Vesque, Anselme, Couve, Charnay, & Schneider-Maunoury, 2006). *TNC* and *MACF1* are also directly involved in neuronal development and/or growth. *TNC* (Tenascin-C) is an extracellular matrix glycoprotein involved in axonal growth and guidance (Jakovcevski, Miljkovic, Schachner, & Andjus, 2013). Seizures up-regulate *TNC* in the hippocampus, and in a pilocarpine epilepsy model up-regulation was shown to be mediated by TGF- β signalling (Mercado-Gomez, Landgrave-Gomez, Arriaga-Avila, Nebreda-Corona, & Guevara-Guzman, 2014). *MACF1* is a cytoskeletal crosslinking protein highly expressed in the brain and is crucial for neuron development and migration (Moffat, Ka, Jung, Smith, & Kim, 2017). *MACF1* mutations are associated with the neurological pathologies Parkinson's disease, autism, and schizophrenia (Moffat et al., 2017). Recently, highly penetrant *de novo MACF1* mutations were identified in several patients with a newly characterized lissencephaly with a complex brain malformation (Dobyns et al., 2018). This new phenotype highlights *MACF1* mutations' variable impact on disease pathogenesis. Given both the enrichment in Epi4k probands for compound heterozygous mutations in these genes as well as their known involvement in neuronal processes, we suggest that *PRTG*, *TNC*, and *MACF1* are candidate recessive epilepsy genes.

The primary publication reporting analysis of the Epi4k trio dataset was a description of *de novo* mutations in the probands (Epi et al., 2013). An analysis of compound heterozygous variants was also reported, using a minor allele frequency cutoff of 0.15%, which is lower than the cutoff used in the work presented here. In this analysis, the parents were used as internal controls, and compound heterozygous variants in 351 genes were identified, without genome-wide significance. The authors only listed five of the genes which are known to cause Mendelian disorders that include a seizure phenotype – *ASPM*, *CNTNAP2*, *GPR98*, *PCNT*, and *POMGNT1*. In our analysis using the 1000 genomes participants as controls, enrichment for compound heterozygous variants was not detected in any of these genes. Using the number of individuals with *in trans* variants in a gene (transcript) as an indicator function required at least two probands to have qualifying variants in order to detect single-test significance, with complete absence of qualifying variants in controls. It is clear from the analyses using either internal controls or the 1000 genomes as controls that a larger sample size is required to achieve genome-wide significance.

The *de novo* variants reported by the Epi4K Consortium and the Epilepsy Phenome/Genome Project (Epi et al., 2013) in the nine probands with compound heterozygous variants in *PRTG*, *TNC*, or *MACF1* are described in Table 2. Of the twelve genes with *de novo* variants identified in the nine patients, three are implicated in neurological disease. *CELSR1* is a planar cell polarity gene in which mutations are known to cause neural tube defects including spina bifida (Robinson et al., 2012). *De novo* deletions of *DIP2C* have been reported in two patients with cerebral palsy, one of whom also had ADHD, and the other had seizures in infancy (Zarrei et al., 2018). In another report, deletions including *DIP2C* and/or *ZMYND11* were identified in several patients with developmental delay including three patients with seizures (DeScipio et

al., 2012). *GLRA2* is a glycine receptor involved in neurodevelopment in which mutations are implicated in autism, (Pilorge et al., 2016; Lin et al., 2017) including a patient with comorbid epilepsy (Zhang et al., 2017).

Of the *de novo* variants reported in these genes, the nonsense variant in *CELSRI* identified in one of the probands with in trans *PRTG* variants is the most likely to be pathogenic. However, regarding their involvement in neural tube defects, mutations in *CELSRI* are thought to contribute to pathogenesis but not in a Mendelian fashion, as variants have been found to be inherited from unaffected parents or to be ineffective in functional assays (Robinson et al., 2012; Allache et al., 2012). The nonsense *CELSRI* mutation in the patient reported here may contribute to epilepsy in the presence of a genetic modifier. The *de novo* missense mutation in *DIP2C* is predicted to be deleterious (CADD = 23.6) and has a low rate of benign missense variation based on constraint metrics ($z = 5.82$). The *de novo* variant in *GLRA2* is in the 3'UTR so it is difficult to predict its impact on gene function and subsequent pathogenicity.

The compound heterozygous variants in *PRTG*, *TNC*, and *MACF1* are similarly variable in predicted pathogenicity, with CADD scores ranging from between less than one to thirty-eight. *PRTG* and *TNC* both have constraint metrics indicative of a high tolerance to both missense and loss-of-function variants, while *MACF1* is moderately intolerant of missense variants ($z = 2.63$) and extremely intolerant of loss-of-function variants (pLI = 1.0). Interestingly, aside from the 3'UTR variant in *GLRA2*, none of the *de novo* variants in the Epi4k participants with *MACF1* compound heterozygous variants are in genes associated with neurological disease or predicted with confidence to have a negative impact on gene function. This, in addition to *MACF1*'s intolerance to missense or nonsense variants, is supportive of the pathogenicity of the biallelic variants in the gene.

In summary, we present a free tool VarCount for the quantification of qualifying mutations as an indicator function per individual in the analysis of variant lists (vcf files). We used VarCount to assess enrichment of rare, coding, compound heterozygous variants in a cohort of 264 epilepsy probands and found enrichment in three genes involved in neurodevelopmental processes – *PRTG*, *TNC* and *MACF1*. A missense change at the E104 residue of PRTG was identified three times in two different probands. Significance was not maintained after correction for multiple testing, and larger cohorts or candidate gene studies using a different sample set are necessary to validate this enrichment. In the context of the *de novo* mutations also present in these patients, experimentation is necessary in order to delineate if the compound heterozygous or *de novo* mutations, or both, are pathogenic in the development of epileptic encephalopathy. *PRTG*, *TNC*, and *MACF1* are candidate recessive epilepsy genes and our work highlights that inheritance of compound heterozygous variants should not be excluded from gene discovery or diagnostic analyses of patients with epilepsy.

ACKNOWLEDGEMENTS. This work was supported by the following grants: T32GM008629 (XX), T32GM082729-01(XX), T32GM007337 (XX and XX), R01AR059703 (PJF, VBM and AGB) and R01NS098590 (AGB, PJF).

References

- Allache, R., et al. (2012). "Role of the planar cell polarity gene CELSR1 in neural tube defects and caudal agenesis." *Birth Defects Res A Clin Mol Teratol* 94(3): 176-181.
- Amrutkar, C. and R. M. Riel-Romero (2018). *Lennox Gastaut Syndrome*. StatPearls. Treasure Island (FL).
- Auton A Fau - Brooks, L. D., Brooks Ld Fau - Durbin, R. M., Durbin Rm Fau - Garrison, E. P., Garrison Ep Fau - Kang, H. M., Kang Hm Fau - Korbel, J. O., Korbel Jo Fau - Marchini, J. L., . . . Abecasis, G. R. A global reference for human genetic variation. (1476-4687 (Electronic)). doi:D - NLM: NIHMS753481
- Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol*, 18(1), 77. doi:10.1186/s13059-017-1212-4

- Callen, D. F., Ricciardelli, C., Butler, M., Stapleton, A., Stahl, J., Kench, J. G., . . . Holm, R. (2010). Co-expression of the androgen receptor and the transcription factor ZNF652 is related to prostate cancer outcome. *Oncol Rep*, 23(4), 1045-1052.
- Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2), 80-92. doi:10.4161/fly.19695
- Clifford, G. M., Rana, R. K., Franceschi, S., Smith, J. S., Gough, G., & Pimenta, J. M. (2005). Human papillomavirus genotype distribution in low-grade cervical lesions: comparison by geographic region and with cervical cancer. *Cancer Epidemiol Biomarkers Prev*, 14(5), 1157-1164. doi:10.1158/1055-9965.EPI-04-0812
- Cornet, I., Gheit, T., Franceschi, S., Vignat, J., Burk, R. D., Sylla, B. S., . . . Group, I. H. V. S. (2012). Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR. *J Virol*, 86(12), 6855-6861. doi:10.1128/JVI.00483-12
- Correa, R. G., Krajewska, M., Ware, C. F., Gerlic, M., & Reed, J. C. (2014). The NLR-related protein NWD1 is associated with prostate cancer and modulates androgen receptor signaling. *Oncotarget*, 5(6), 1666-1682. doi:10.18632/oncotarget.1850
- Covanis, A. (2012). "Epileptic encephalopathies (including severe epilepsy syndromes)." *Epilepsia* 53 Suppl 4: 114-126.
- Cox, A. J., Darbro, B. W., Laxer, R. M., Velez, G., Bing, X., Finer, A. L., . . . Ferguson, P. J. (2017). Recessive coding and regulatory mutations in FBLIM1 underlie the pathogenesis of chronic recurrent multifocal osteomyelitis (CRMO). *PLoS One*, 12(3), e0169687. doi:10.1371/journal.pone.0169687
- DeScipio, C., et al. (2012). "Subtelomeric deletion of chromosome 10p15.3: clinical findings and molecular cytogenetic characterization." *Am J Med Genet A* 158A(9): 2152-2161.
- Dobyns, W. B., et al. (2018). "MACF1 Mutations Encoding Highly Conserved Zinc-Binding Residues of the GAR Domain Cause Defects in Neuronal Migration and Axon Guidance." *Am J Hum Genet*.
- Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res*, 32(Web Server issue), W665-667. doi:10.1093/nar/gkh381
- Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., & Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12(2), 241-253.
- Epi, K. C., Epilepsy Phenome/Genome, P., Allen, A. S., Berkovic, S. F., Cossette, P., Delanty, N., . . . Winawer, M. R. (2013). De novo mutations in epileptic encephalopathies. *Nature*, 501(7466), 217-221. doi:10.1038/nature12439
- Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061-1073. doi:10.1038/nature09534
- Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56-65. doi:10.1038/nature11632
- Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74. doi:10.1038/nature15393
- Giovannucci, E., Stampfer, M. J., Krithivas, K., Brown, M., Dahl, D., Brufsky, A., . . . Kantoff, P. W. (1997). The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. *Proc Natl Acad Sci U S A*, 94(7), 3320-3323.
- Goncalves, C. M., Castro, M. A., Henriques, T., Oliveira, M. I., Pinheiro, H. C., Oliveira, C., . . . Carmo, A. M. (2009). Molecular cloning and analysis of SSc5D, a new member of the scavenger receptor cysteine-rich superfamily. *Mol Immunol*, 46(13), 2585-2596. doi:10.1016/j.molimm.2009.05.006
- Haiman, C. A., Chen, G. K., Blot, W. J., Strom, S. S., Berndt, S. I., Kittles, R. A., . . . Henderson, B. E. (2011). Genome-wide association study of prostate cancer in men of

- African ancestry identifies a susceptibility locus at 17q21. *Nat Genet*, 43(6), 570-573. doi:10.1038/ng.839
- Ho, L., Chan, S. Y., Burk, R. D., Das, B. C., Fujinaga, K., Icenogle, J. P., . . . et al. (1993). The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. *J Virol*, 67(11), 6413-6423.
- Jakovcevski, I., Miljkovic, D., Schachner, M., & Andjus, P. R. (2013). Tenascins and inflammation in disorders of the nervous system. *Amino Acids*, 44(4), 1115-1127. doi:10.1007/s00726-012-1446-0
- Jelen, M. M., Chen, Z., Kocjan, B. J., Burt, F. J., Chan, P. K., Chouhy, D., . . . Poljak, M. (2014). Global genomic diversity of human papillomavirus 6 based on 724 isolates and 190 complete genome sequences. *J Virol*, 88(13), 7307-7316. doi:10.1128/JVI.00621-14
- Jelen, M. M., Chen, Z., Kocjan, B. J., Hosnjak, L., Burt, F. J., Chan, P. K., . . . Poljak, M. (2016). Global Genomic Diversity of Human Papillomavirus 11 Based on 433 Isolates and 78 Complete Genome Sequences. *J Virol*, 90(11), 5503-5513. doi:10.1128/JVI.03149-15
- Konecny, R., Baker, N. A., & McCammon, J. A. (2012). iAPBS: a programming interface to Adaptive Poisson-Boltzmann Solver (APBS). *Comput Sci Discov*, 5(1). doi:10.1088/1749-4699/5/1/015005
- Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A., & Neale, B. M. (2016). Discovery of rare variants for complex phenotypes. *Hum Genet*, 135(6), 625-634. doi:10.1007/s00439-016-1679-1
- Kossoff, E. H. (2010). "Infantile spasms." *Neurologist* 16(2): 69-75.
- Lauring, A. S., Frydman, J., & Andino, R. (2013). The role of mutational robustness in RNA virus evolution. *Nat Rev Microbiol*, 11(5), 327-336. doi:10.1038/nrmicro3003
- Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., . . . MacArthur, D. (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*. doi:10.1101/030338
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285-291. doi:10.1038/nature19057
- Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., . . . Wang, J. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet*, 42(11), 969-972. doi:10.1038/ng.680
- Lin, M. S., et al. (2017). "alpha2-glycine receptors modulate adult hippocampal neurogenesis and spatial memory." *Dev Neurobiol* 77(12): 1430-1441.
- Martinez, V. G., Moestrup, S. K., Holmskov, U., Mollenhauer, J., & Lozano, F. (2011). The conserved scavenger receptor cysteine-rich superfamily in therapy and diagnosis. *Pharmacol Rev*, 63(4), 967-1000. doi:10.1124/pr.111.004523
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9), 1297-1303. doi:10.1101/gr.107524.110
- Mercado-Gomez, O., Landgrave-Gomez, J., Arriaga-Avila, V., Nebreda-Corona, A., & Guevara-Guzman, R. (2014). Role of TGF-beta signaling pathway on Tenascin C protein upregulation in a pilocarpine seizure model. *Epilepsy Res*, 108(10), 1694-1704. doi:10.1016/j.eplepsyres.2014.09.019
- Miro-Julia, C., Rosello, S., Martinez, V. G., Fink, D. R., Escoda-Ferran, C., Padilla, O., . . . Lozano, F. (2011). Molecular and functional characterization of mouse S5D-SRCRB: a new group B member of the scavenger receptor cysteine-rich superfamily. *J Immunol*, 186(4), 2344-2354. doi:10.4049/jimmunol.1000840
- Miwa, H., Go, M. F., & Sato, N. (2002). H. pylori and gastric cancer: the Asian enigma. *Am J Gastroenterol*, 97(5), 1106-1112. doi:10.1111/j.1572-0241.2002.05663.x

- Moffat, J. J., Ka, M., Jung, E. M., Smith, A. L., & Kim, W. Y. (2017). The role of MACF1 in nervous system development and maintenance. *Semin Cell Dev Biol*, *69*, 9-17. doi:10.1016/j.semcdb.2017.05.020
- Moshfegh, Y., Velez, G., Li, Y., Bassuk, A. G., Mahajan, V. B., & Tsang, S. H. (2016). BESTROPHIN1 mutations cause defective chloride conductance in patient stem cell-derived RPE. *Hum Mol Genet*, *25*(13), 2672-2680. doi:10.1093/hmg/ddw126
- O'Brien, T. G., Guo, Y., Visvanathan, K., Sciulli, J., McLaine, M., Helzlsouer, K. J., & Watkins-Bruner, D. (2004). Differences in ornithine decarboxylase and androgen receptor allele frequencies among ethnic groups. *Mol Carcinog*, *41*(2), 120-123. doi:10.1002/mc.20047
- Oliveros, J. C. (2015). Venny. An interactive tool for comparing lists with Venn's diagrams.
- Peprah, E., Xu, H., Tekola-Ayele, F., & Royal, C. D. (2015). Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics*, *18*(1), 40-51. doi:10.1159/000367962
- Pilorge, M., et al. (2016). "Genetic and functional analyses demonstrate a role for abnormal glycinergic signaling in autism." *Mol Psychiatry* *21*(7): 936-945.
- Polley, S., Louzada, S., Forni, D., Sironi, M., Balaskas, T., Hains, D. S., . . . Hollox, E. J. (2015). Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy. *Proc Natl Acad Sci U S A*, *112*(16), 5105-5110. doi:10.1073/pnas.1416531112
- Rand, K. A., Rohland, N., Tandon, A., Stram, A., Sheng, X., Do, R., . . . Haiman, C. A. (2016). Whole-exome sequencing of over 4100 men of African ancestry and prostate cancer risk. *Hum Mol Genet*, *25*(2), 371-381. doi:10.1093/hmg/ddv462
- Robinson, A., et al. (2012). "Mutations in the planar cell polarity genes CELSR1 and SCRIB are associated with the severe neural tube defect craniorachischisis." *Hum Mutat* *33*(2): 440-447.
- Salter, J. D., Bennett, R. P., & Smith, H. C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem Sci*, *41*(7), 578-594. doi:10.1016/j.tibs.2016.05.001
- Sanjak, J. S., Long, A. D., & Thornton, K. R. (2017). A Model of Compound Heterozygous, Loss-of-Function Alleles Is Broadly Consistent with Observations from Complex-Disease GWAS Datasets. *PLoS Genet*, *13*(1), e1006573. doi:10.1371/journal.pgen.1006573
- Stelzer, G., Inger, A., Olender, T., Iny-Stein, T., Dalah, I., Harel, A., . . . Lancet, D. (2009). GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation. *OMICS*, *13*(6), 477-487. doi:10.1089/omi.2009.0069
- Toral, M. A., Velez, G., Boudreault, K., Schaefer, K. A., Xu, Y., Saffra, N., . . . Mahajan, V. B. (2017). Structural modeling of a novel SLC38A8 mutation that causes foveal hypoplasia. *Mol Genet Genomic Med*, *5*(3), 202-209. doi:10.1002/mgg3.266
- Toyoda, R., Nakamura, H., & Watanabe, Y. (2005). Identification of protogenin, a novel immunoglobulin superfamily gene expressed during early chick embryogenesis. *Gene Expr Patterns*, *5*(6), 778-785. doi:10.1016/j.modgep.2005.04.001
- Vesque, C., Anselme, I., Couve, E., Charnay, P., & Schneider-Maunoury, S. (2006). Cloning of vertebrate Protogenin (Prtg) and comparative expression analysis during axis elongation. *Dev Dyn*, *235*(10), 2836-2844. doi:10.1002/dvdy.20898
- Vieira, V. C., Leonard, B., White, E. A., Starrett, G. J., Temiz, N. A., Lorenz, L. D., . . . Harris, R. S. (2014). Human papillomavirus E6 triggers upregulation of the antiviral and cancer genomic DNA deaminase APOBEC3B. *MBio*, *5*(6). doi:10.1128/mBio.02234-14
- W, D. The PyMOL Molecular Graphics System.
- Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., . . . Abecasis, G. R. Ancestry estimation and control of population stratification for sequence-based association studies. (1546-1718 (Electronic)). doi:D - NLM: NIHMS588246

- D - NLM: PMC4084909 EDAT- 2014/03/19 06:00 MHDA- 2014/05/20 06:00 CRDT- 2014/03/18 06:00 PHST- 2013/08/08 [received] PHST- 2014/02/21 [accepted] AID - ng.2924 [pii] AID - 10.1038/ng.2924 [doi] PST - ppublish
- Warren, C. J., Westrich, J. A., Doorslaer, K. V., & Pyeon, D. (2017). Roles of APOBEC3A and APOBEC3B in Human Papillomavirus Infection and Disease Progression. *Viruses*, 9(8). doi:10.3390/v9080233
- Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci*, 86, 291-2937. doi:10.1002/cpps.20
- Zaidi, S. F. (2016). Helicobacter pylori associated Asian enigma: Does diet deserve distinction? *World J Gastrointest Oncol*, 8(4), 341-350. doi:10.4251/wjgo.v8.i4.341
- Zarrei, M., et al. (2018). "De novo and rare inherited copy-number variations in the hemiplegic form of cerebral palsy." *Genet Med* 20(2): 172-180.
- Zhang, Y., et al. (2017). "Structure-Function Analysis of the GlyR alpha2 Subunit Autism Mutation p.R323L Reveals a Gain-of-Function." *Front Mol Neurosci* 10: 158.
- Zhong, K., Karssen, L. C., Kayser, M., & Liu, F. (2016). CollapsABEL: an R library for detecting compound heterozygote alleles in genome-wide association studies. *BMC Bioinformatics*, 17, 156. doi:10.1186/s12859-016-1006-9

Table 1. Rare (< 0.5% and 1.0% MAF) Compound Heterozygous Mutations in Epi4k participants

Gene	Transcript	Sex	Phen ^a	Anc ^b	Chr:bp*	dbSNP ID	REF	ALT	Exac ALL AF ^c	ExAC NFE AF	Peptide change	Epi4k #	1kg #	p-value	Epi4k #	1kg #	p-value
												EU(y/n): All(y/n)	EU(y/n): All(y/n)	EU:All	EU(y/n): All(y/n)	EU(y/n): All(y/n)	EU:All
PRTG	ENST000 00389286	F	IS	EU	15:56032666	rs373423650	T	C	9.94E-05	1.65E-04	E104G	3/204: 3/261	0/503: 0/2504	0.0245 0.00086	2/205: 2/262	0/503: 0/2504	0.0847: 0.0091
					15:56032665	rs185716584	C	G	0.003022	0.004315	E104D						
		M	LGS	EU	15:55965698	rs372777171	C	A	8.28E-06	1.50E-05	E575*						
					15:56032665	rs185716584	C	G	0.003022	0.004315	E104D						
					15:56035093	rs35718474	G	A	0.003033	0.006701	P13S						
15:55916703	rs148011047	T	G	0.002611	0.003608	K977T											
OSBP2	ENST000 00332585	M	IS	EU	22:31091324	rs200118898	C	G	1.09E-04	1.98E-04	S143W	2/205: 2/262	0/503: 0/2504	0.0847: 0.0091	2/205: 2/262	0/503: 0/2504	0.292: 0.0091
					22:31137288	rs201298398	G	A	3.20E-04	5.19E-04	R262Q						
		M	LGS	E Asi	22:31091239	rs576508023	G	C	8.42E-06	0	G115R						
					22:31283452	NA	C	T	1.66E-05	0	R383W						
ABCC11	ENST000 00394747	F	IS	EU	16:48242388	rs529824818	C	T	8.24E-06	0	C543Y	2/205: 2/262	0/503: 2/2502	0.0847: 0.0478	2/205: 2/262	0/503: 1/2502	0.0847: 0.0255
					16:48226471	NA	C	A	NA	NA	R889M						
		M	LGS	EU	16:48234362	NA	C	A	NA	NA	G636V						
TNC	ENST000 00345230	M	IS	EU	9:117840315	rs200401362	C	T	8.24E-06	0	G861R	2/205: 3/261	0/503: 4/2500	0.0847: 0.022	1/206: 2/262	0/503: 2/2502	0.29: 0.0478
					9:117849499	rs117058692	C	G	7.77E-04	0.001248	G171R						
		M	IS	C/S Asia	9:117853183	rs143586851	C	T	2.23E-04	3.90E-04	A39T						
					9:117849402	rs149986851	C	A	2.06E-04	3.75E-04	G203V						
		M	IS	EU	9:117835900	rs139280264	G	A	8.51E-04	0.001247	R1066C						
					9:117848284	rs371055558	C	T	2.54E-05	3.07E-05	G576S						
9:117849382	rs144032672	C	T	0.001336	0.002026	G210S											
MACF1	ENST000 00289893	M	IS	EU	1:39900231	NA	G	A	1.65E-05	3.00E-05	R4344Q	3/204: 3/261	1/502: 11/2493	0.0769: 0.142	3/204: 3/261	0/503: 6/2498	0.0245 0.0478
					1:39853797	rs141949859	G	T	6.34E-04	1.12E-03	V3535F						
		M	LGS	EU	1:39852984	rs145271544	G	T	NA	NA	A3264S						
					1:39951304	rs138819868	T	G	2.46E-03	3.85E-03	F5885L						
					1:39800136	NA	A	G	8.28E-06	0	I1066V						
1:39910474	NA	G	C	8.264	1.50E-05	W4967C											
STAB1	ENST000 00321725	M	IS	ME	3:52549439	rs145751447	G	A	9.17E-05	0	D1289N	0/207: 2/262	1/502: 6/2498	0.99: 0.173	0/207: 2/262	1/502: 2/2502	0.99: 0.0478
					3:52554531	rs147953260	G	A	0.001512	0.002524	R1872H						
		F	IS	C/S Asia	3:52540204	rs189303343	A	G	4.05E-05	3.66E-05	I590V						
					3:52557268	NA	C	T	NA	NA	R2351W						

^a Phen=phenotype, ^bAnc=ancestry, ^cAF=allele frequency,

^dy/n corresponds to yes/no counts of individuals with qualifying mutations

*bp (base pair position) in hg19/Build37

Ancestries: EU=European, E Asia = East Asia, C/S Asia = Central/South Asia, ME = Middle East

p-values in bold are the most significant for the specific analysis

Table 2. *De novo* variants in Epi4k probands with CH variants in PRTG, TNC, or MACF1

Gene	missense z-score ^a	pLI ^b	CH variants	Sex	Phenotype	gene w/ de novo	variant type	Polyphen-2 ^c	CADD ^d	missense z-score ^a	pLI ^b	
<i>PRTG</i>	-0.92	0.02	E104G, E104D	F	IS	<i>HSF2</i>	missense	P	24	1.77	0.93	
<i>PRTG</i>			E104D, E575*	M	LGS	<i>CELSR1</i>	stop-gained	NA	35	4.42	1.00	
<i>PRTG</i>			P13S, K977T	M	LGS	<i>FAM102A</i>	missense	D	25.1	1.80	0.43	
						<i>USP42</i>	3' UTR	NA		-0.69	1.00	
<i>TNC</i>	-0.14	0.00	G861R, G171R, A39T	M	IS	<i>DIP2C</i>	missense	B	23.6	5.82	1.00	
							<i>IFT172</i>	splice donor	NA	23.9	0.22	0.00
<i>TNC</i>			G203V, R1066C	M	IS	NA	-	-	-	-	-	-
<i>TNC</i>			G576S, G210S	M	IS	NA	-	-	-	-	-	
<i>MACF1</i>	2.63	1.00	R4344Q, V3535F	M	IS	<i>FAM19A2</i>	5' UTR	NA		1.11	0.55	
						<i>GLRA2</i>	3' UTR	NA		3.28	0.91	
<i>MACF1</i>			A3264S, F5885L	M	IS	<i>LRRC8D</i>	3' UTR	NA		2.70	0.87	
						<i>SNX30</i>	missense	B	17.9	0.90	0.97	
						<i>WDFY2</i>	synonymous	NA		1.15	0.00	
<i>MACF1</i>	I1066V, W4967C	M	LGS	<i>FAM227A</i>	missense	P	9.4	NA	NA			

Ancestries: EU=European, C/S Asia = Central/South Asia; Phenotypes: IS = Infantile spasms, LGS = Lennox-Gastaut syndrome

^az-score is a measure of tolerance to missense variants, based on ratio of expected to identified; ^bpLI is the probability that a gene is intolerant to loss-of-function variants;

^cPolyphen2 – prediction of a missense mutation's impact on protein structure and function: B=benign, P=possibly damaging, D=damaging (Adzhubei et al., 2010); ^dCADD = phred-scaled score of Combined Annotation Dependent Depletion, a measure of the deleteriousness of a SNP or INDEL (Kircher et al., 2014).

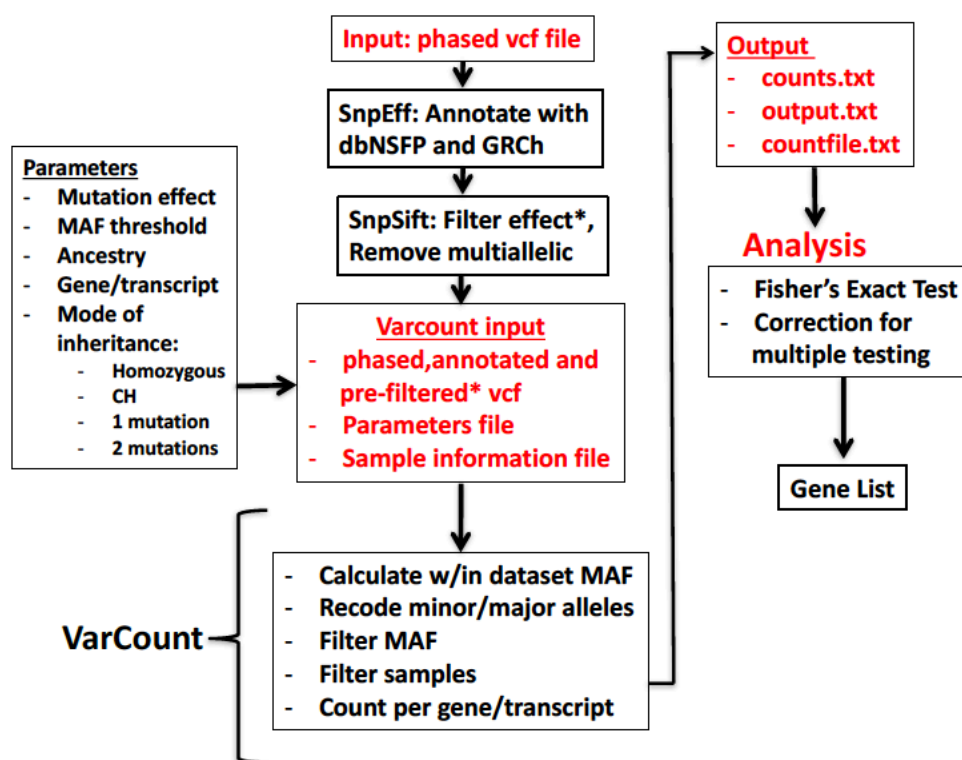


Figure 1. Flow diagram for the processing and analysis of variant lists. Vcf files are annotated and filtered using SnpSift/SnpEff. Final vcf along with parameter and sample information files are input to VarCount. The input files are processed to recode minor and major alleles when the MAF > 0.5 and to count the number of individuals with mutations qualifying based on information in the parameter file. The final output lists for every transcript or gene, the number of individuals with qualified mutations in that locus (counts.txt), which individuals have the mutation(s) (countfile.txt), and which mutations are harbored by each individual (output.txt).

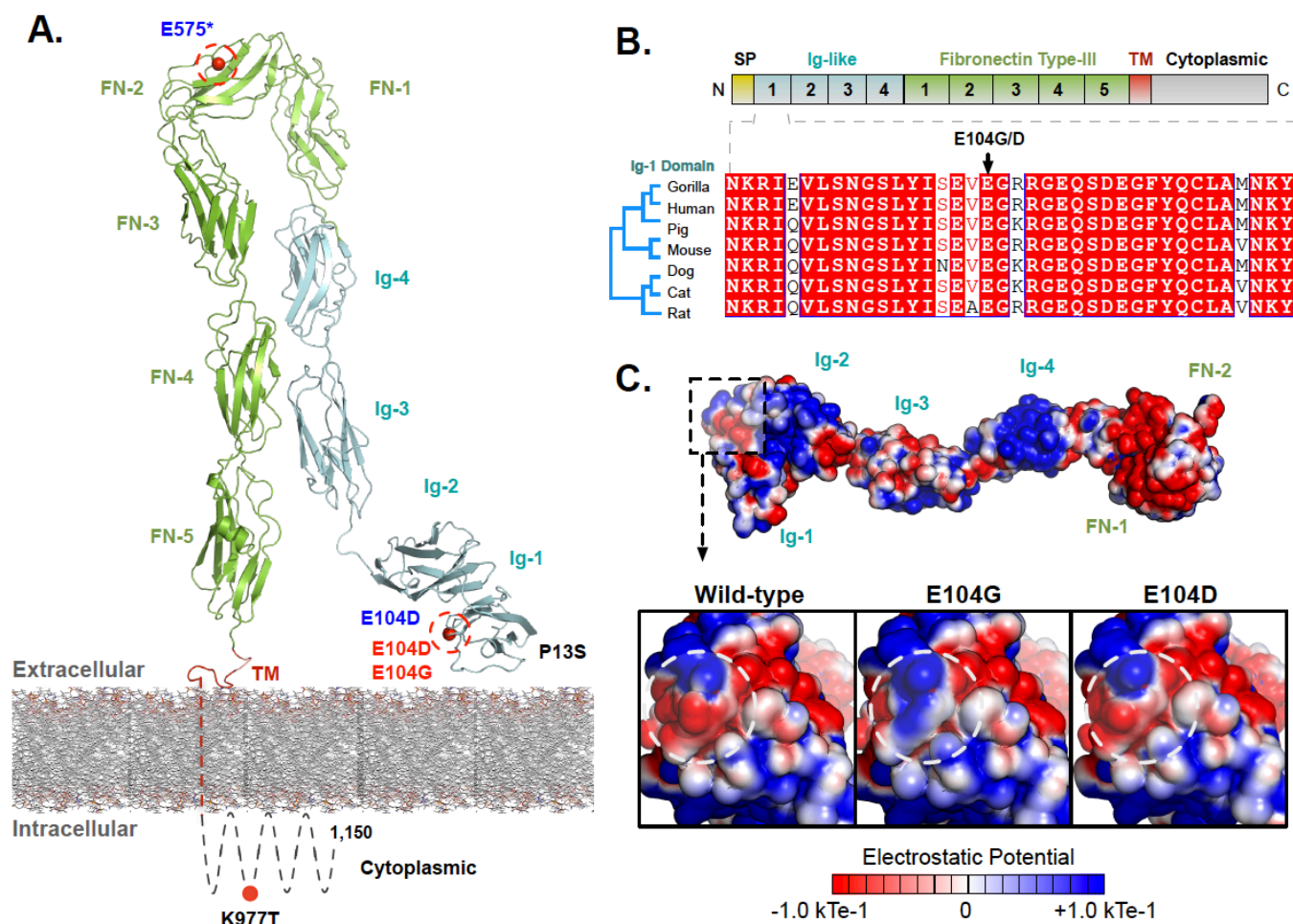


Figure 2. PRTG compound heterozygous mutations in Epi4k probands. (A) Protein diagram indicating mutation locations in each child. The three pairs of *in trans* mutations are indicated in red, blue and black. The mutations in red and blue were found using a 0.5% MAF threshold and the mutations in black were found using a 1% MAF threshold. (B) Bottom: electrostatic potential surface of PRTG calculated in APBS. Top: Close-up of the PRTG electrostatic potential surface at the site of mutation. The p.Glu104Gly mutation leads to a loss of negative charge, which may disrupt interactions with putative PRTG binding partners. The p.Glu104Asp mutation does not lead to a change in charge or electrostatic potential.

Supplementary Table 1: Ancestry and Phenotype Information for Epi4k probands

proband ID	TRACE ^a ancestry*	Reported Ancestry	IS/LGS ^b	Sex	proband ID	TRACE ancestry	Reported Ancestry	IS/LGS	Sex
isnd21451b1	C/S Asia	Asian	IS	F	isnd35150fl1	EU	White	IS	M
isnd21751e1	EU	White	IS	M	isnd35151fo1	EU	White	IS	M
isnd22993f1	EU	White	IS	M	isnd35197fn1	EU	White	IS	F
isnd23231g1	EU	Other	IS	M	isnd35351fz1	America	White	IS	M
isnd23465h1	EU	White	IS	M	isnd35575et1	EU	White	IS	F
isnd24005j1	EU	White	IS	M	isnd35845ek1	ME	Other	IS	M
isnd24104l1	C/S Asia	Asian	IS	M	isnd35907fs1	Africa	Other	IS	M
isnd24188n1	EU	White	IS	F	isnd35929ge1	EU	White	IS	M
isnd24217o1	EU	White	IS	M	isnd35951fu1	EU	White	IS	M
isnd24290af1	ME	Hispanic	IS	M	isnd36066fd1	EU	White	IS	F
isnd24346p1	ME	White	IS	M	isnd36158fv1	EU	White	IS	F
isnd24470t1	EU	White	IS	M	isnd36206fw1	C/S Asia	White	IS	F
isnd24539d1	America	Hispanic	IS	M	isnd36211dg1	EU	White	IS	F
isnd24704r1	EU	White	IS	F	isnd36367fx1	C/S Asia	Asian	IS	F
isnd24782s1	America	Hispanic	IS	M	isnd36387fy1	EU	White	IS	F
isnd25070u1	EU	Other	IS	F	isnd36561ga1	EU	White	IS	M
isnd25181w1	EU	White	IS	M	isnd36610ge1	EU	White	IS	M
isnd25582x1	EU	White	IS	F	lgsnd22752gg1	EU	Other	LGS	M
isnd25606v1	EU	White	IS	F	lgsnd23319gi1	America	Hispanic	LGS	M
isnd25793ac1	EU	White	IS	M	lgsnd23543gr1	EU	White	LGS	M
isnd25839z1	Africa	African American	IS	M	lgsnd23813gv1	EU	White	LGS	M
isnd26087ad1	America	Hispanic	IS	F	lgsnd23828gl1	EU	White	LGS	M
isnd26900ah1	EU	White	IS	F	lgsnd24053gj1	EU	White	LGS	F
isnd26970ai1	EU	White	IS	F	lgsnd24065go1	EU	White	LGS	M
isnd26974aj1	EU	White	IS	F	lgsnd24070gk1	EU	White	LGS	M
isnd27062aa1	EU	White	IS	F	lgsnd24191jw1	EU	White	LGS	M
isnd27253al1	EU	White	IS	F	lgsnd24447gm1	EU	White	LGS	F
isnd27474am1	EU	White	IS	F	lgsnd24471gn1	EU	White	LGS	M
isnd27521bi1	EU	Other	IS	M	lgsnd24646gp1	EU	White	LGS	F
isnd27732an1	Africa	Other	IS	M	lgsnd24762gq1	C/S Asia	White	LGS	F
isnd27841ay1	Africa	African American	IS	M	lgsnd25442gs1	EU	White	LGS	F
isnd27935ar1	EU	White	IS	M	lgsnd25544gt1	EU	White	LGS	F
isnd27949ag1	EU	White	IS	M	lgsnd25992hc1	EU	Other	LGS	M
isnd28478au1	EU	White	IS	F	lgsnd26319gy1	EU	White	LGS	F
isnd28661av1	EU	White	IS	M	lgsnd27109hd1	EU	White	LGS	F
isnd28699ax1	EU	White	IS	M	lgsnd27155he1	EU	White	LGS	F
isnd28895fp1	EU	White	IS	M	lgsnd27345hh1	EU	White	LGS	F
isnd28982be1	EU	White	IS	M	lgsnd27497hi1	EU	White	LGS	M
isnd29057be1	EU*	White	IS	F	lgsnd27543hn1	E Asia	Asian	LGS	F
isnd29126bj1	EU	White	IS	M	lgsnd27594hj1	EU	Other	LGS	M
isnd29199bk1	C/S Asia	White	IS	M	lgsnd27637hk1	EU	White	LGS	M
isnd29258bl3	EU	White	IS	M	lgsnd27682hl1	EU	White	LGS	M
isnd29267bm1	EU	Other	IS	M	lgsnd27753ha1	EU	White	LGS	M
isnd29292ca1	EU/CS	White	IS	F	lgsnd27952in1	EU	White	LGS	F
isnd29305co1	EU	White	IS	M	lgsnd28027hw1	EU	White	LGS	M
isnd29319bo1	EU	White	IS	F	lgsnd28181hp1	EU	White	LGS	F
isnd29352br1	EU	White	IS	F	lgsnd28245hq1	EU	White	LGS	M
isnd29366bs1	EU	White	IS	F	lgsnd28402ip1	EU	White	LGS	M
isnd29377bt1	ME C/S Asia	Other	IS	F	lgsnd28432hu1	EU	White	LGS	M

^aAncestry determined by LASER/TRACE software, ^bIS = Infantile Spasms, LGS = Lennox Gastaut Syndrome

*Ancestries: C/S Asia = Central/South Asia, E Asia = East Asia, EU = European, ME = ME

Supplementary Table 1 Continued: Ancestry and Phenotype Information for Epi4k probands

proband ID	TRACE ^a ancestry*	Reported Ancestry	IS/LGS ^b	Sex	proband ID	TRACE ancestry	Reported Ancestry	IS/LGS	Sex
isnd29378bu1	EU	Other	IS	M	lgsnd28509hr1	EU	White	LGS	M
isnd29383bv1	C/S Asia	White	IS	M	lgsnd28633ht1	EU	White	LGS	M
isnd29429es1	EU	White	IS	M	lgsnd28840hf1	EU	White	LGS	M
isnd29514bw1	EU	White	IS	F	lgsnd28866hs1	EU	White	LGS	F
isnd29556k1	EU/CS Asia	White	IS	F	lgsnd28881hv1	C/S Asia	Other	LGS	F
isnd29711ao1	EU	White	IS	F	lgsnd28949hx1	EU	White	LGS	M
isnd29810bq1	EU	White	IS	F	lgsnd29055hz1	EU	Other	LGS	M
isnd29844az1	EU	White	IS	F	lgsnd29058ia1	EU	White	LGS	F
isnd29865by1	EU	White	IS	M	lgsnd29125iz1	E Asia	Asian	LGS	M
isnd29900bg1	EU	White	IS	F	lgsnd29146ic1	EU	White	LGS	M
isnd30071cf1	ME	Other	IS	M	lgsnd29196ig1	EU	White	LGS	F
isnd30086ak1	EU	White	IS	F	lgsnd29374gh1	EU	White	LGS	M
isnd30090dk1	EU	Other	IS	M	lgsnd29394ie1	EU	Other	LGS	F
isnd30279ci1	EU	White	IS	M	lgsnd29446if1	ME	White	LGS	M
isnd30280cj1	EU	White	IS	F	lgsnd29528ih1	EU	White	LGS	M
isnd30302ck1	EU	White	IS	M	lgsnd29554gw1	America	Hispanic	LGS	M
isnd30373dm1	EU	White	IS	M	lgsnd29789ii1	EU	White	LGS	M
isnd30377cm1	EU	White	IS	M	lgsnd29838ij1	EU	Hispanic	LGS	F
isnd30384cn1	EU	White	IS	M	lgsnd29864it1	EU	White	LGS	M
isnd30431cp1	C/S Asia	White	IS	F	lgsnd29904ik1	Africa	Other	LGS	M
isnd30439cv1	EU	White	IS	M	lgsnd29958il1	EU	White	LGS	M
isnd30441cr1	C/S Asia	White	IS	M	lgsnd30052je1	EU	White	LGS	M
isnd30474ch1	EU/CS	Other	IS	F	lgsnd30133iq1	EU	White	LGS	M
isnd30482dy1	EU	White	IS	F	lgsnd30216iy1	EU	White	LGS	F
isnd30485cs1	EU	White	IS	M	lgsnd30241ir1	EU	White	LGS	M
isnd30552ct1	EU	White	IS	F	lgsnd30378ix1	EU	White	LGS	M
isnd30575cz1	EU	White	IS	F	lgsnd30383iv1	ME	White	LGS	M
isnd30610cg1	EU	White	IS	M	lgsnd30631iw1	EU	White	LGS	M
isnd30629gb1	EU	Other	IS	F	lgsnd30729js1	EU	White	LGS	F
isnd30679cu1	EU	White	IS	M	lgsnd30798ja1	EU	White	LGS	F
isnd30831dl1	EU	White	IS	M	lgsnd30864jf1	EU	White	LGS	M
isnd30880ex1	EU	White	IS	M	lgsnd30965jg1	E Asia	Asian	LGS	M
isnd30915as1	EU	White	IS	M	lgsnd31059jh1	EU	White	LGS	M
isnd31115da1	EU	White	IS	F	lgsnd31063ji1	EU	White	LGS	F
isnd31120cq1	EU	Other	IS	M	lgsnd31153jj1	EU	White	LGS	F
isnd31134ba1	EU	White	IS	M	lgsnd31159jd1	EU	White	LGS	F
isnd31192db1	EU	White	IS	M	lgsnd31244jk1	Africa	Other	LGS	M
isnd31228dc1	EU	White	IS	M	lgsnd31529id1	C/S Asia	Other	LGS	M
isnd31241dr1	EU	White	IS	F	lgsnd31533jo1	EU	Other	LGS	M
isnd31305dd1	E Asia	Asian	IS	F	lgsnd31574jm1	EU	White	LGS	M
isnd31308de1	C/S Asia	Asian	IS	F	lgsnd31650jv1	EU	White	LGS	M
isnd31364df1	EU	White	IS	M	lgsnd31664jn1	EU	White	LGS	M
isnd31602ds1	EU	White	IS	M	lgsnd31867jq1	EU	White	LGS	F
isnd31635di1	EU	White	IS	M	lgsnd31894jz1	EU	White	LGS	M
isnd31702dh1	EU	White	IS	M	lgsnd31959kb1	EU	White	LGS	M
isnd31770cy1	America	Hispanic	IS	F	lgsnd31961jp1	EU	White	LGS	M
isnd31821dn1	EU	White	IS	M	lgsnd32224jb1	E Asia	Other	LGS	M
isnd31831do1	EU	White	IS	F	lgsnd32239kg1	EU	White	LGS	M
isnd31899dq1	EU	White	IS	F	lgsnd32265jt1	EU	White	LGS	M

^aAncestry determined by LASER/TRACE software, ^bIS = Infantile Spasms, LGS = Lennox Gastaut Syndrome

*Ancestries: C/S Asia = Central/South Asia, E Asia = East Asia, EU = European, ME = ME

Supplementary Table 1 Continued: Ancestry and Phenotype Information for Epi4k probands

proband ID	TRACE ^a ancestry*	Reported Ancestry	IS/LGS ^b	Sex	proband ID	TRACE ancestry	Reported Ancestry	IS/LGS	Sex
isnd32065dv1	EU	Other	IS	M	lgsnd32289ju1	EU	Other	LGS	M
isnd32121dj1	EU	White	IS	F	lgsnd32340kk1	EU	White	LGS	M
isnd32131ce1	CS/E Asia	White	IS	F	lgsnd32497jr1	EU	White	LGS	M
isnd32132dw1	EU	White	IS	M	lgsnd32552jy1	EU	White	LGS	M
isnd32241ec1	EU	White	IS	F	lgsnd32562jx1	EU	White	LGS	F
isnd32398dz1	C/S Asia	Other	IS	F	lgsnd32630ka1	ME	White	LGS	F
isnd32464eb1	EU	White	IS	F	lgsnd32670kl1	EU	White	LGS	M
isnd32641du1	EU	White	IS	M	lgsnd32727kd1	EU	White	LGS	F
isnd32671ef1	EU	White	IS	M	lgsnd32763ke1	EU	White	LGS	M
isnd32722cd1	EU	White	IS	F	lgsnd32802kf1	EU	White	LGS	F
isnd32757ee1	EU	White	IS	M	lgsnd32879kj1	EU	White	LGS	F
isnd33197eg1	Africa	African American	IS	F	lgsnd32890jl1	EU	White	LGS	M
isnd33296eh1	EU	White	IS	M	lgsnd33014iu1	Africa	African American	LGS	M
isnd33322el1	E Asia	Asian	IS	M	lgsnd33064kh1	EU	White	LGS	F
isnd33342ej1	EU	White	IS	M	lgsnd33323kc1	EU	White	LGS	M
isnd33520bn1	Africa	African American	IS	F	lgsnd33346km1	EU	White	LGS	M
isnd33651ed1	EU	White	IS	M	lgsnd33590kn1	C/S Asia	White	LGS	F
isnd34077er1	EU	White	IS	F	lgsnd33706ko1	EU	White	LGS	F
isnd34116ff1	C/S Asia	White	IS	F	lgsnd33762kq1	EU	Other	LGS	F
isnd34128ep1	C/S Asia	White	IS	F	lgsnd34131kr1	EU	White	LGS	F
isnd34144ew1	America	Hispanic	IS	M	lgsnd34164ks1	EU	White	LGS	F
isnd34170dt1	EU	White	IS	M	lgsnd34306ki1	EU	White	LGS	M
isnd34274ex1	EU	White	IS	M	lgsnd34424kv1	EU	White	LGS	M
isnd34304ey1	EU	White	IS	F	lgsnd34500kx1	EU	White	LGS	F
isnd34338ez1	EU	White	IS	M	lgsnd34528kz1	EU	White	LGS	F
isnd34401fc1	EU	White	IS	F	lgsnd34593kt1	EU	White	LGS	M
isnd34404fa1	Africa	African American	IS	F	lgsnd34816ku1	ME	Other	LGS	M
isnd34430aq1	EU	White	IS	M	lgsnd35136ky1	EU	White	LGS	M
isnd34548fb1	Africa	African American	IS	M	lgsnd35495kp1	EU	White	LGS	M
isnd34680fj1	C/S Asia	Other	IS	F	lgsnd35817lb1	EU	Other	LGS	M
isnd34750fe1	EU	White	IS	F	lgsnd36163lc1	EU	White	LGS	M
isnd34962fh1	EU	White	IS	F	lgsnd36210jc1	Africa	African American	LGS	M
isnd34968fg1	EU	White	IS	F	lgsnd36440le1	EU	White	LGS	M
isnd35054fi1	EU	Other	IS	F	lgsnd36798lg1	Africa	African American	LGS	F

^aAncestry determined by LASER/TRACE software, ^bIS = Infantile Spasms, LGS = Lennox Gastaut Syndrome

*Ancestries: C/S Asia = Central/South Asia, E Asia = East Asia, EU = European, ME = ME

Supplementary Table 2. Epi4k and 1000 genomes CH and homozygous counts and p-values, all ancestries, 1%MAF, top 50

Gene	TRID ^a	Epi4k 1% yes ^b	Epi4k 1% no	1kg ^c 1% yes	1kg 1% no	pval	Pval bh ^d	Pval bonf ^e	OR ^f
PRTG	ENST00000389286	3	261	0	2504	0.000858684	1	1	Inf
DNAJC4	ENST00000321460	2	262	0	2504	0.009065347	1	1	Inf
DNAJC4	ENST00000321685	2	262	0	2504	0.009065347	1	1	Inf
OSBP2	ENST00000332585	2	262	0	2504	0.009065347	1	1	Inf
OSBP2	ENST00000382310	2	262	0	2504	0.009065347	1	1	Inf
OSBP2	ENST00000446658	2	262	0	2504	0.009065347	1	1	Inf
MUC16	ENST00000397910	11	253	215	2289	0.01244148	1	1	0.462993427
FLG	ENST00000368799	4	260	113	2391	0.015655138	1	1	0.325614508
TNC	ENST00000345230	3	261	4	2500	0.022454859	1	1	7.172904306
TNC	ENST00000423613	3	261	4	2500	0.022454859	1	1	7.172904306
TNC	ENST00000537320	3	261	4	2500	0.022454859	1	1	7.172904306
C9orf114	ENST00000361256	2	262	1	2503	0.025478672	1	1	19.0610251
TNC	ENST00000346706	3	261	5	2499	0.033434418	1	1	5.737893801
PKD1	ENST00000262304	1	263	59	2445	0.040946041	1	1	0.15762227
PKD1	ENST00000423118	1	263	59	2445	0.040946041	1	1	0.15762227
TNC	ENST00000340094	3	261	6	2498	0.04668945	1	1	4.780720266
TNC	ENST00000341037	3	261	6	2498	0.04668945	1	1	4.780720266
TNC	ENST00000350763	3	261	6	2498	0.04668945	1	1	4.780720266
ANTXRL	ENST00000447511	3	261	6	2498	0.04668945	1	1	4.780720266
TNC	ENST00000535648	3	261	6	2498	0.04668945	1	1	4.780720266
TNC	ENST00000542877	3	261	6	2498	0.04668945	1	1	4.780720266
ABCC11	ENST00000353782	2	262	2	2502	0.047765771	1	1	9.533392929
ABCC11	ENST00000356608	2	262	2	2502	0.047765771	1	1	9.533392929
ABCC11	ENST00000394747	2	262	2	2502	0.047765771	1	1	9.533392929
ABCC11	ENST00000394748	2	262	2	2502	0.047765771	1	1	9.533392929
PTCHD3	ENST00000438700	2	262	2	2502	0.047765771	1	1	9.533392929
MUC17	ENST00000306151	1	263	51	2453	0.056611754	1	1	0.182941761
NLRX1	ENST00000409265	2	262	3	2501	0.074665107	1	1	6.355667514
SCN10A	ENST00000449082	2	262	3	2501	0.074665107	1	1	6.355667514
NLRX1	ENST00000525863	2	262	3	2501	0.074665107	1	1	6.355667514
FUK	ENST00000571514	2	262	3	2501	0.074665107	1	1	6.355667514
MUC17	ENST00000379439	1	263	48	2456	0.082911884	1	1	0.194612284
AHNAK2	ENST00000333244	3	261	80	2424	0.083759229	1	1	0.348364788
CX3CL1	ENST00000006053	1	263	0	2504	0.095375723	1	1	Inf
TSPAN32	ENST00000182290	1	263	0	2504	0.095375723	1	1	Inf
CEACAM21	ENST00000187608	1	263	0	2504	0.095375723	1	1	Inf
C19orf26	ENST00000215376	1	263	0	2504	0.095375723	1	1	Inf
RNF215	ENST00000215798	1	263	0	2504	0.095375723	1	1	Inf
UBFD1	ENST00000219638	1	263	0	2504	0.095375723	1	1	Inf
KIAA1199	ENST00000220244	1	263	0	2504	0.095375723	1	1	Inf
TGFB1	ENST00000221930	1	263	0	2504	0.095375723	1	1	Inf
CRX	ENST00000221996	1	263	0	2504	0.095375723	1	1	Inf
PRDM2	ENST00000235372	1	263	0	2504	0.095375723	1	1	Inf
NT5C1A	ENST00000235628	1	263	0	2504	0.095375723	1	1	Inf
DHDSS	ENST00000236342	1	263	0	2504	0.095375723	1	1	Inf
CAMSAP2	ENST00000236925	1	263	0	2504	0.095375723	1	1	Inf
CCDC92	ENST00000238156	1	263	0	2504	0.095375723	1	1	Inf
VIL1	ENST00000248444	1	263	0	2504	0.095375723	1	1	Inf
STRIP2	ENST00000249344	1	263	0	2504	0.095375723	1	1	Inf

^aTRID = Transcript ID, ^bCount of subjects with CH or homozygous mutations, ^c1kg = 1000 genomes, ^dbh = Benamini Hochberg correction, ^ebonf = Bonferroni correction, ^fOR = Odds ratio

Supplementary Table 3. Epi4k and 1000 genomes CH and homozygous counts and p-values, all ancestries, 0.5%MAF, top 50

Gene	TRID ^a	Epi4k 0.5% yes ^b	Epi4k 0.5% no	1kg ^c 0.5% yes	1kg 1% no	pval	Pval bh ^d	Pval bonf ^e	OR ^f
OSBP2	ENST00000332585	2	262	0	2504	0.009065347	1	1	Inf
OSBP2	ENST00000382310	2	262	0	2504	0.009065347	1	1	Inf
PRTG	ENST00000389286	2	262	0	2504	0.009065347	1	1	Inf
OSBP2	ENST00000446658	2	262	0	2504	0.009065347	1	1	Inf
ABCC11	ENST00000353782	2	262	1	2503	0.025478672	1	1	19.0610251
ABCC11	ENST00000356608	2	262	1	2503	0.025478672	1	1	19.0610251
ABCC11	ENST00000394747	2	262	1	2503	0.025478672	1	1	19.0610251
ABCC11	ENST00000394748	2	262	1	2503	0.025478672	1	1	19.0610251
MACF1	ENST00000289893	3	261	6	2498	0.04668945	1	1	4.780720266
STAB1	ENST00000321725	2	262	2	2502	0.047765771	1	1	9.533392929
TNC	ENST00000345230	2	262	2	2502	0.047765771	1	1	9.533392929
TNC	ENST00000423613	2	262	2	2502	0.047765771	1	1	9.533392929
TNC	ENST00000537320	2	262	2	2502	0.047765771	1	1	9.533392929
TNC	ENST00000340094	2	262	3	2501	0.074665107	1	1	6.355667514
TNC	ENST00000341037	2	262	3	2501	0.074665107	1	1	6.355667514
TNC	ENST00000346706	2	262	3	2501	0.074665107	1	1	6.355667514
TNC	ENST00000350763	2	262	3	2501	0.074665107	1	1	6.355667514
TNC	ENST00000535648	2	262	3	2501	0.074665107	1	1	6.355667514
TNC	ENST00000542877	2	262	3	2501	0.074665107	1	1	6.355667514
MACF1	ENST00000372915	3	261	8	2496	0.079581706	1	1	3.583524267
MACF1	ENST00000564288	3	261	8	2496	0.079581706	1	1	3.583524267
MACF1	ENST00000567887	3	261	8	2496	0.079581706	1	1	3.583524267
CX3CL1	ENST00000006053	1	263	0	2504	0.095375723	1	1	Inf
LLGL2	ENST00000167462	1	263	0	2504	0.095375723	1	1	Inf
CEACAM21	ENST00000187608	1	263	0	2504	0.095375723	1	1	Inf
UBFD1	ENST00000219638	1	263	0	2504	0.095375723	1	1	Inf
KIAA1199	ENST00000220244	1	263	0	2504	0.095375723	1	1	Inf
TFR2	ENST00000223051	1	263	0	2504	0.095375723	1	1	Inf
TIMELESS	ENST00000229201	1	263	0	2504	0.095375723	1	1	Inf
PRDM2	ENST00000235372	1	263	0	2504	0.095375723	1	1	Inf
DHDDS	ENST00000236342	1	263	0	2504	0.095375723	1	1	Inf
CCDC92	ENST00000238156	1	263	0	2504	0.095375723	1	1	Inf
AHDC1	ENST00000247087	1	263	0	2504	0.095375723	1	1	Inf
VIL1	ENST00000248444	1	263	0	2504	0.095375723	1	1	Inf
STRIP2	ENST00000249344	1	263	0	2504	0.095375723	1	1	Inf
E2F8	ENST00000250024	1	263	0	2504	0.095375723	1	1	Inf
CHRNA10	ENST00000250699	1	263	0	2504	0.095375723	1	1	Inf
KRT12	ENST00000251643	1	263	0	2504	0.095375723	1	1	Inf
LPIN1	ENST00000256720	1	263	0	2504	0.095375723	1	1	Inf
RALGAPA1	ENST00000258840	1	263	0	2504	0.095375723	1	1	Inf
ARHGAP20	ENST00000260283	1	263	0	2504	0.095375723	1	1	Inf
BCAR3	ENST00000260502	1	263	0	2504	0.095375723	1	1	Inf
KHK	ENST00000260599	1	263	0	2504	0.095375723	1	1	Inf
ADCY3	ENST00000260600	1	263	0	2504	0.095375723	1	1	Inf
EXOC6	ENST00000260762	1	263	0	2504	0.095375723	1	1	Inf
PZP	ENST00000261336	1	263	0	2504	0.095375723	1	1	Inf
TRPV4	ENST00000261740	1	263	0	2504	0.095375723	1	1	Inf
TELO2	ENST00000262319	1	263	0	2504	0.095375723	1	1	Inf
CCM2L	ENST00000262659	1	263	0	2504	0.095375723	1	1	Inf

^aTRID = Transcript ID, ^bCount of subjects with CH or homozygous mutations, ^c1kg = 1000 genomes, ^dbh = Benamini Hochberg correction, ^ebonf = Bonferroni correction, ^fOR = Odds ratio

Supplementary Table 4. Epi4k and 1000 genomes CH and homozygous counts and p-values, EU ancestry, 1%MAF, top 50

Gene	TRID ^a	Epi4k 1% yes ^b	Epi4k 1% no	1kg ^c 1% yes	1kg 1% no	pval	Pval bh ^d	Pval bonf ^e	OR ^f
PRTG	ENST00000389286	3	204	0	503	0.024527517	1	1	Inf
TTN	ENST00000342992	28	179	43	460	0.053560126	1	1	1.67202516
OBSCN	ENST00000366707	4	203	2	501	0.063150347	1	1	4.923199244
OBSCN	ENST00000422127	7	200	6	497	0.063251543	1	1	2.894171878
MACF1	ENST00000289893	3	204	1	502	0.076878328	1	1	7.358758844
SYNE2	ENST00000357395	3	204	1	502	0.076878328	1	1	7.358758844
MACF1	ENST00000372915	3	204	1	502	0.076878328	1	1	7.358758844
SYNE2	ENST00000394768	3	204	1	502	0.076878328	1	1	7.358758844
MUC2	ENST00000441003	3	204	1	502	0.076878328	1	1	7.358758844
SYNE2	ENST00000555002	3	204	1	502	0.076878328	1	1	7.358758844
MACF1	ENST00000564288	3	204	1	502	0.076878328	1	1	7.358758844
MACF1	ENST00000567887	3	204	1	502	0.076878328	1	1	7.358758844
TNC	ENST00000340094	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000341037	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000345230	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000346706	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000350763	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000353782	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000356608	2	205	0	503	0.084709668	1	1	Inf
SYNE1	ENST00000356820	2	205	0	503	0.084709668	1	1	Inf
C9orf114	ENST00000361256	2	205	0	503	0.084709668	1	1	Inf
ZZEF1	ENST00000381638	2	205	0	503	0.084709668	1	1	Inf
SDK2	ENST00000388726	2	205	0	503	0.084709668	1	1	Inf
SDK2	ENST00000392650	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000394747	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000394748	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000423613	2	205	0	503	0.084709668	1	1	Inf
OBSCN	ENST00000441106	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000535648	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000537320	2	205	0	503	0.084709668	1	1	Inf
TNC	ENST00000542877	2	205	0	503	0.084709668	1	1	Inf
TTN	ENST00000591111	28	179	46	457	0.104033319	1	1	1.553012662
OBSCN	ENST00000570156	7	200	7	496	0.132113123	1	1	2.476401909
TTN	ENST00000589042	28	179	50	453	0.186519554	1	1	1.41648123
TTN	ENST00000342175	18	189	30	473	0.191266503	1	1	1.5006686
TTN	ENST00000359218	18	189	30	473	0.191266503	1	1	1.5006686
TTN	ENST00000460472	18	189	30	473	0.191266503	1	1	1.5006686
MACF1	ENST00000317713	2	205	1	502	0.205073971	1	1	4.884989822
DNAH11	ENST00000328843	2	205	1	502	0.205073971	1	1	4.884989822
PCNT	ENST00000359568	2	205	1	502	0.205073971	1	1	4.884989822
MACF1	ENST00000361689	2	205	1	502	0.205073971	1	1	4.884989822
MDN1	ENST00000369393	2	205	1	502	0.205073971	1	1	4.884989822
MACF1	ENST00000372925	2	205	1	502	0.205073971	1	1	4.884989822
DNAH11	ENST00000409508	2	205	1	502	0.205073971	1	1	4.884989822
MDN1	ENST00000428876	2	205	1	502	0.205073971	1	1	4.884989822
MACF1	ENST00000539005	2	205	1	502	0.205073971	1	1	4.884989822
MACF1	ENST00000545844	2	205	1	502	0.205073971	1	1	4.884989822
HIVEP2	ENST00000012134	1	206	0	503	0.291549296	1	1	Inf
TNS1	ENST00000171887	1	206	0	503	0.291549296	1	1	Inf

^aTRID = Transcript ID, ^bCount of subjects with CH or homozygous mutations, ^c1kg = 1000 genomes, ^dbh = Benamini Hochberg correction, ^ebonf = Bonferroni correction, ^fOR = Odds ratio

Supplementary Table 5. Epi4k and 1000 genomes CH and homozygous counts and p-values, EU ancestry, 0.5%MAF, top 50

Gene	TRID ^a	Epi4k 0.5% yes ^b	Epi4k 0.5% no	1kg ^c 0.5% yes	1kg 1% no	pval	Pval bh ^d	Pval bonf ^e	OR ^f
MACF1	ENST00000289893	3	204	0	503	0.024527517	1	1	Inf
MACF1	ENST00000372915	3	204	0	503	0.024527517	1	1	Inf
MACF1	ENST00000564288	3	204	0	503	0.024527517	1	1	Inf
MACF1	ENST00000567887	3	204	0	503	0.024527517	1	1	Inf
OBSCN	ENST00000366707	4	203	2	501	0.063150347	1	1	4.923199244
MACF1	ENST00000317713	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000353782	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000356608	2	205	0	503	0.084709668	1	1	Inf
MACF1	ENST00000361689	2	205	0	503	0.084709668	1	1	Inf
MACF1	ENST00000372925	2	205	0	503	0.084709668	1	1	Inf
PRTG	ENST00000389286	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000394747	2	205	0	503	0.084709668	1	1	Inf
ABCC11	ENST00000394748	2	205	0	503	0.084709668	1	1	Inf
OBSCN	ENST00000441106	2	205	0	503	0.084709668	1	1	Inf
MACF1	ENST00000539005	2	205	0	503	0.084709668	1	1	Inf
MACF1	ENST00000545844	2	205	0	503	0.084709668	1	1	Inf
MYO15A	ENST00000205890	2	205	1	502	0.205073971	1	1	4.884989822
DNAH2	ENST00000389173	2	205	1	502	0.205073971	1	1	4.884989822
PKD1L2	ENST00000525539	2	205	1	502	0.205073971	1	1	4.884989822
PKD1L2	ENST00000533478	2	205	1	502	0.205073971	1	1	4.884989822
DNAH2	ENST00000572933	2	205	1	502	0.205073971	1	1	4.884989822
TTN	ENST00000342992	17	190	28	475	0.234492	1	1	1.516901915
HIVEP2	ENST00000012134	1	206	0	503	0.291549296	1	1	Inf
TNS1	ENST00000171887	1	206	0	503	0.291549296	1	1	Inf
CEACAM21	ENST00000187608	1	206	0	503	0.291549296	1	1	Inf
TGM6	ENST00000202625	1	206	0	503	0.291549296	1	1	Inf
LAMB1	ENST00000222399	1	206	0	503	0.291549296	1	1	Inf
TFR2	ENST00000223051	1	206	0	503	0.291549296	1	1	Inf
DHDDS	ENST00000236342	1	206	0	503	0.291549296	1	1	Inf
CCDC92	ENST00000238156	1	206	0	503	0.291549296	1	1	Inf
TICAM1	ENST00000248244	1	206	0	503	0.291549296	1	1	Inf
VIL1	ENST00000248444	1	206	0	503	0.291549296	1	1	Inf
E2F8	ENST00000250024	1	206	0	503	0.291549296	1	1	Inf
DHX29	ENST00000251636	1	206	0	503	0.291549296	1	1	Inf
KRT12	ENST00000251643	1	206	0	503	0.291549296	1	1	Inf
PLXNA1	ENST00000251772	1	206	0	503	0.291549296	1	1	Inf
AKAP12	ENST00000253332	1	206	0	503	0.291549296	1	1	Inf
DNHD1	ENST00000254579	1	206	0	503	0.291549296	1	1	Inf
LPIN1	ENST00000256720	1	206	0	503	0.291549296	1	1	Inf
RABEPK	ENST00000259460	1	206	0	503	0.291549296	1	1	Inf
ARHGAP20	ENST00000260283	1	206	0	503	0.291549296	1	1	Inf
TRPV4	ENST00000261740	1	206	0	503	0.291549296	1	1	Inf
NUP153	ENST00000262077	1	206	0	503	0.291549296	1	1	Inf
FBN2	ENST00000262464	1	206	0	503	0.291549296	1	1	Inf
PDE4C	ENST00000262805	1	206	0	503	0.291549296	1	1	Inf
KIF18A	ENST00000263181	1	206	0	503	0.291549296	1	1	Inf
CLTCL1	ENST00000263200	1	206	0	503	0.291549296	1	1	Inf
LRP2	ENST00000263816	1	206	0	503	0.291549296	1	1	Inf
TECTA	ENST00000264037	1	206	0	503	0.291549296	1	1	Inf

^aTRID = Transcript ID, ^bCount of subjects with CH or homozygous mutations, ^c1kg = 1000 genomes, ^dbh = Benamini Hochberg correction, ^ebonf = Bonferroni correction, ^fOR = Odds ratio

Supplementary Dataset 1 contains counts of individuals from the Epi4k and 1000 genomes dataset with compound heterozygous or homozygous mutations in each transcript. Results files with counts and p-values for all transcripts are included as well. Dataset S1 is available for viewing on dropbox @ <https://www.dropbox.com/sh/yu46zg8bjsafgt9/AACVgJWAj-26mrgsLivJublfa?dl=0>