Genetic Sequencing Methods in N08CB

Genetic sequencing

Genomic DNA was isolated from peripheral blood leukocytes and sequencing libraries were prepared using a TruSeq reagent kit (Illumina). DNA was subjected to acoustic shearing (Covaris), and barcoded adapters were ligated to the DNA fragments. Library quality was confirmed by electrophoresis on a Bioanalyzer (Agilent). Target enrichment was carried out using the recommended standard protocol and SureSelect oligonucleotide capture reagents (Agilent). Libraries of cases and controls were processed in random order and randomly assigned to pools to preclude the possibility of a batch effect. Sequencing was performed on a HiSeq 2000 sequencing system (Illumina) in paired-end mode to a length of 151bp x 2.

Raw sequencing reads were trimmed to remove adapter sequences and low-quality (Phred < 15) bases using BBDuk (https://sourceforge.net/projects/bbmap/). Reads were aligned to the human reference genome GRCh37 (hg19) using BWA version 0.7.8 (http://bio-bwa.sourceforge.net/). PCR duplicates were marked using Picard version 2.4.1 (http://broadinstitute.github.io/picard/). All of the remaining sequence processing steps were performed using Genome Analysis Toolkit (GATK) version 3.8 (https://software.broadinstitute.org/gatk/). Reads were realigned around indels using the Mills & 1000G Gold Standard and 1000G Phase 1 indels as references. Base quality score recalibration was performed with the same reference indels plus dbSNP build 138 as a reference for known SNPs.

Variant calling

Variant calling and variant quality score recalibration (VQSR) were performed using GATK with the GATK-recommended parameters for exome capture sequencing and with quality filtering set to achieve a predicted variant-detection sensitivity of 99%. Variants passing quality filtering were separated into SNV and indels, with only bi-allelic SNV used in the statistical analyses.

Additional quality control

Variants in non-autosomal regions of the genome were excluded. Autosomal variants were tested for deviation from Hardy-Weinberg equilibrium (a one-sided test to detect excess heterozygotes only) and were excluded based on a false discovery rate threshold of 0.01, according to the Benjamini-Hochberg procedure.

Orthogonal validation of genotypes

For the patients in this study, several SNV were genotyped using an orthogonal method: the mass-spectrometry based MassARRAY (Sequenom). Four of those assayed SNV (rs34587622, rs9038, rs17722209, and rs6875902) were among the SNV in CMT genes sequenced for this study. For

one SNV (rs34587622), the MassARRAY data were deemed unreliable because of a high rate of failed assays (18/157) and severe violation of Hardy-Weinberg equilibrium among the genotypes ($p = 7 \times 10^{-17}$). The genotypes for the remaining three SNV were used to validate the sequencing genotypes.

Variant annotation

The predicted effects of the final set of SNV on genetic transcripts were obtained using Ensembl VEP build 91. For statistical analysis, only SNV predicted to affect the coding region in a protein-coding transcript were considered. These consisted of SNV annotated by VEP with a predicted Consequence among the following: splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost, start_lost, missense_variant, or synonymous_variant. SNV with any of the listed consequences except synonymous_variant were counted as non-synonymous (protein-altering).