Pharmacogenomic incidental findings in 308 families: The NIH Undiagnosed Disease Program experience

Running title: Reporting pharmacogenomic incidental findings

Elizabeth M.J. Lee, Karen Xu, Emma Mosbrook, Amanda Links, Jessica Guzman, David R. Adams, MD, PhD, Elise Flynn, Elise Valkanas, Camillo Toro, MD, Cynthia J. Tifft, MD, PhD, Cornelius F. Boerkoel, MD, PhD, William A. Gahl, MD, PhD, Murat Sincan, MD

NIH Undiagnosed Diseases Program, Common Fund, Office of the Director, NIH; and National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

Correspondence: Murat Sincan, M.D.

Address: BG 10-CRC RM 5-2535 MSC 1470, 10 Center Drive, Bethesda, MD 20892

Phone: 301-827-1647

Fax: 301-480-0498

Email: sincanm@mail.nih.gov

MATERIALS AND METHODS

Subject Cohort

Family members gave informed consent or assent under protocol 76-HG-0238, “Diagnosis and Treatment of Patients with Inborn Errors of Metabolism and Other Genetic Disorders,” approved by the NHGRI Institutional Review Board. The SNP chip data were derived from a cohort of 308 families consisting of 355 affected individuals, 278 unaffected siblings, 459 unaffected parents, and 9 other unaffected family members; this cohort included 564 females and 537 males. The exome sequence data were from a subset of this cohort; 158 families contained 182 affected individuals, 150 unaffected siblings, 313 unaffected parents, 326 females and 319 males. The average and median ages of the subjects at time of SNP chip analysis were 36.1 (SD 22.4) and 36.0 years, respectively. The average and median ages at the time of exome analysis were 35.2 (SD 22.4) and 36.0 years, respectively. Some subjects were deceased at the time of study, and for those subjects, projected age at time of sequencing was used, since it is anticipated that incidental findings will only be sought in living subjects. Self-reported ancestry was White/European (75.3%), Black/African American (2.5%), American Indian or Alaskan Native (0.4%), Asian (2.1%), Multiracial (5.3%), and Unknown (14.4%). These families included all those admitted to the NIH Undiagnosed Diseases Program and who had SNP chip or exome analyses. The SNP chip genotyping and exome sequencing were performed on a research basis between 2009 and 2014, not in a CLIA-certified fashion.

DNA Extraction

Genomic DNA was extracted from patients' peripheral whole blood using the Gentra Puregene Blood kit (Qiagen) per the manufacturer's protocol.

SNP Chip Analysis

The extracted DNA was submitted to the NHGRI core lab and run on the SNP oligo array Human OmniExpressExome v1.2 (Illumina Corp, San Diego CA). Each family was analyzed individually following the UDP's standard operating procedure. Call rates were >98% before quality control processing, and were typically >99.7%.

Exome Sequencing

Genomic DNA was submitted for exome sequencing using the Illumina TruSeq exome capture kit (Illumina, Inc., San Diego, US), which targets roughly 60 million bases consisting of the Consensus Coding Sequence (CCDS) annotated gene set as well as some structural RNAs. Captured DNA was sequenced on the Illumina HiSeq platform until coverage was sufficient to call high quality genotypes at 85% or more of targeted bases.

Alignment and Genotype Calling

BEAGLE software version 4 and the 1000 Genome's HapMap data were used to generate a phased and imputed Variant Call Format (VCF) file from SNP chip data of the parents and offspring [1]. The VCF file was then used by AlleleSeq version 0.2.3 [2] to modify the human reference and create a maternal reference and a paternal reference, which are concatenated together to generate a parental reference. Patient short reads from exome sequencing were aligned to all three reference sequences with Novoalign version 2.08.03 and were lifted back over to the standard human reference using custom Java code. BAM files were recalibrated and genotyped by HaplotypeCaller according to GATK Best Practices using GATK v2.5-2 [3,4].

Variant Annotation

Variants were annotated using transcript data from UCSC Genome Browser Database with Annovar [5]. The VCF files of SNP and exome sequencing variants in the study cohort were also annotated using the ClinVar database. The annotations taken from ClinVar included the dbSNP reference numbers, PubMed links, and PharmGKB annotation. The relationship dataset from PharmGKB was used to annotate each variant with the associated medication names (Figure 1).

Pharmacological Analysis

Using the annotated data, we identified patient SNPs with pharmacological implications, or pharmacogenomic incidental findings. The number of variants with pharmacological implications that matched a patient's genotype was calculated per patient for both SNP genotypes and exome variants. The number of medications with potential implications was also calculated per patient.

Medication lists were compiled for study participants by extracting medication records from the NIH Biomedical Translational Research Information System (BTRIS) and from the NIH UDP Integrated Collaboration System (UDPICS). This medication list was then intersected with the list of medications associated with each participant's sequence variants. For each intersection, the validity of the medication match, subject genotype, subject race and subject sex were manually curated against the PharmGKB database and supporting publications.

Phenotype Analysis

If available, patient medical records were reviewed to determine if a patient had a medical history or phenotypic characteristics that correlated with the pharmacogenetic finding.

References

1. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics.* Feb 2009;84(2):210-223.

2. Rozowsky J, Abyzov A, Wang J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology.* 2011;7:522.

3. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics.* May 2011;43(5):491-498.

4. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research.* Sep 2010;20(9):1297-1303.

5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research.* Sep 2010;38(16):e164.