Supplementary Methods

Target Selection and Reagent Design

The genes targeted for the RIGHT10K project largely reflected a combination of the 21 gene-drug pairs already identified and implemented in the Mayo Clinic's Personal Genomics Laboratory (PGL) and the next generation of the Human Genome Sequencing Center at Baylor College of Medicine's (BCM-HGSC) PGRN-seq custom oligonucleotide capture reagent. The original PGRN-seq panel was developed by the Deep Sequencing Resource of the Pharmacogenomics Research Network (PGRN) as described previously[1] and used by the eMERGE Network's PGx Project[2]. For this project, a new capture reagent for pharmacogenomics analysis based on prior methods developed for 'gene panels'[3], incorporated our prior experience to target 77 genes including all genes for which the Clinical Pharmacogenetics Implementation Consortium (CPIC) clinical guidelines (https://cpicpgx.org/guidelines/) existed at the time of design. Briefly, the new design eliminated the non-coding flanking sequences targeted in previous iterations (exceptions include the entire *CYP2D6* locus including its pseudogenes, intron 6 of *CYP3A4* and both exons and introns for *CYP2A6* and *CYP2A7*) to optimize multiplex levels and control cost per sample. The probe set also targeted the Affymetrix DMET Plus (Affymetrix/Thermo Fisher Scientific, Santa Clara, CA) and Illumina VeraCode ADME (Illumina, San Diego, CA) array genotyping targets to maintain backward compatibility with the majority of genotyping assays together with twelve tag SNVs for specific *HLA-A* and *HLA-B* alleles, fourteen SNVs specified by the PGRN and ninety-six SNV site probes from a Fluidigm SNP Trace panel[4] used to ensure correct sample

identification through the sequencing process. The final design capture space totaled 458,121 bp and 98.8% of design targets were covered with probes.

Sequencing Sample Flow

DNA was extracted from whole blood at the Mayo Clinic and was then transferred to the Baylor College of Medicine's Human Genome Sequencing Center Clinical Laboratory (BCM-HGSC-CL). The paired-end pre-capture library procedure included genomic DNA fragmented by sonication and ligation to Illumina paired-end adapters. The adapter-ligated DNA was PCR-amplified using primers containing sequencing barcodes (indexes) to enable sample multiplexing. For the target enrichment capture procedure, the pre-capture library was enriched by solution hybridization to PGx-Seq biotinylated probes (Roche NimbleGen, Madison, WI) using a 47-plex format[5]. Sequencing was performed using the Illumina HiSeq 2500 platform (Illumina, San Diego, CA) using a 94-plex format generating 101 bp paired-end reads.

Sequence Interpretation and Assay Validation

After extensive development and pilot testing of target coverage using 512 Mayo Clinic RIGHT1K samples, 310 variant sites across 13 genes were chosen for clinical implementation: *CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, DPYD, HLA-A, HLA-B, SLCO1B1, TPMT, UGT1A1*, and *VKORC1*. The BCM-HGSC-CL established CAP/CLIA clinical assay validation documenting accuracy, precision, analytical sensitivity, analytical specificity, reportable range, and reference range. As gold standard samples for pharmacogenomic testing by sequencing are

still being defined, we used a combination of samples including 1 Coriell NIST (National Institute of Standards and Technology) Gold Standard (NA12878), 419 Mayo RIGHT1K pilot samples, 51 Coriell 1000 Genomes samples and 5 clinical exome samples from the Baylor Genetics Lab eMERGE clinical samples[5]. Given the importance of the *CYP2D6* gene in this pharmacogenetics test (219 variant sites) and inherent challenges in the discrimination of *CYP2D6* pseudogene regions, we compared 414 samples from this assay to a uniquely amplified 8.4 kb amplicon of the *CYP2D6* region to determine analytical sensitivity of SNV/Indel detection within this gene. For the *HLA* genes, we used Omixon HLA Explore™ software (Omixon, Budapest, Hungary and Major, et.al[6]) and custom parsing software to identify the four allele types of interest and validated the approach using 30 Mayo Clinic RIGHT 1K clinical samples with independently derived results.

Data Management

Information Technology

Virtually every aspect of this study required information technology (IT) resources to build unique software scripts to handle the data and files. Software was also required for every step of the end-to-end process, including transferring specimens and associated deidentified subject information in a secure fashion to the BCM-HGSC-CL, NGS sequence interpretation and bam, vcf, csv file generation, as well as uploading of results files to a secure server for download by the Mayo Clinic's Personalized Genomics Laboratory (PGL). Further software was required to transfer relevant results to the Mayo EHR, establish subroutines for *CYP2D6* analysis and where

appropriate, ancillary testing, display of results for the laboratory director and management of rare alleles. Finally, new software was developed to securely transfer patient results for further annotation (as outlined below), curation of results for clinical and non-clinical pharmacogenes for research purposes and the generation of a rare allele database for analysis.

Sequence Data Management

Sequencing reads were converted to individual FASTQ files by Illumina bcl2fastq 1.8.3 software, and mapped to the GRCh37 (hg19) human genome reference using the BWA program[7]. The Mercury bioinformatics pipeline was modified to replace the iPipe pipeline management software with HgV and use of the force-calling features in the Atlas2 software for the 310 allele-defining sites so that variant no calls could be distinguished from low coverage. The BCM-HGSC-CL software program xAtlas was used to generate a .gvcf file HTSlib from samtools (https://www.htslib.org/) to produce a gVCF in the 4.1 format using the standard min30p3a non-variant binning scheme. Star Assign software was developed to extract the force-called sites from each sample's VCF file to generate concatenated diplotypes, which in turn were used to query lookup tables supplied by the Mayo Clinic to assign appropriate star alleles and their associated phenotypes as predicted drug metabolizer status for all but the HLA genes and *CYP2D6*. For *HLA*, Star Assign parsed the Omixon HLA Explore output looking for "BEST" and returned "positive" if "BEST" matched one of the four allele types of interest (*HLA-A31:01, HLA-B15:02, HLA-B57:01* and *HLA-B58:01)* and "negative" otherwise. For *CYP2D6*, the Mayo Clinic's Personalized Genomics Laboratory (PGL) used a combination of the 217 *CYP2D6* regional force call sites in the VCF together with read depth and allelic ratios derived from the BAM file for

each sample as input to CNVAR, a PGL-developed calling algorithm used to determine *CYP2D6* star alleles and identify samples having structural variation for reflex to a PCR cascade protocol for final predicted metabolizer phenotypes. This software was separately validated to CLIA specifications by comparing its output to 500 "truth set" samples previously characterized by Luminex (Luminex Corporation, Austin, TX) genotyping and/or real-time QPCR. Due to limitations in oligonucleotide capture, Illumina short read mapping and high homology between *CYP2D6* and its neighboring pseudo genes, we estimated in advance of the project that 3-5% of samples would require ancillary PCR cascade testing to determine final allele structure and phenotype prediction. However, the use of CNVAR resulted in just 72 (0.71%) of 10,077 samples needing reflex testing. The average quality control metrics of the sequencing data were: >75% of reads were aligned to target, 99.85% of target bases were covered, >99% of target bases were covered at >20X, and average coverage of target bases was >490X.

Rare Haplotype Management

For novel/rare variants, Star Assign software was used to confirm SNP/indel quality, exclude those present on a Mayo Clinic list of benign variants and then classify the remaining variants as stop-gain/stop-loss, frameshift, splice site or missense. A text report file was generated for each sample summarizing the genotype and phenotype status for each sample together with its HLA status and any novel variant positions and types. For clinical interpretation of rare variants not included in the look-up tables and flagged for manual curation by laboratory directors with PGx experience in the Mayo Clinic Personalized Genomics Laboratory, we first consulted all known databases of interest including the Pharmacogene Variation Consortium (www.pharmvar.org),

the *TPMT* nomenclature database (https://liu.se/en/research/tpmt-nomenclature-committee), and the *UGT1A1* allele nomenclature database (https://www.pharmacogenomics.pha.ulaval.ca) to determine if the variant was part of a rare haplotype not included in the look-up table and to assign the appropriate star-allele. All variants/alleles with unknown function were evaluated using a combination of *in silico* evaluation tools (Align GVGD, SIFT, Mutation Taster, PolyPhen-2, SpliceSiteFinder, MexEntScan, NNSPLICE, GeneSplicer) and a modification of the ACMG criteria for variant interpretation. Results from *in silico* evaluation tools were interpreted with caution and in conjunction with professional clinical judgment. Variants that were classified as variants of uncertain significance (VUS) and those expected to impact function were included in clinical reports, while variants classified as benign or likely benign were not reported. When a rare variant without a corresponding star-allele was reported, specialized nomenclature was developed. Specifically, if the variant was known to be on the same allele as a known star allele, it was reported using HGVS nomenclature and "with" the star allele (e.g. *3 with a heterozygous c.122A>G, p.Tyr41Cys) and when the cis/trans status was unknown a semi-colon was used (e.g. *1/*2; a heterozygous c.122A>G, p.Tyr41Cys was detected). Genotypes were used to predict phenotypes. If an VUS was identified, a range was provided that included the worst-case scenario (e.g. the VUS results in a no function allele) to the best-case scenario (e.g. the VUS does not impact function), also considering the potential ambiguity in cis/trans status (e.g. worst-case would include the VUS resulting in a no function allele in trans with the defined no function allele, when present, when cis/trans was unknown).

Delivery of Results

We used the secure cloud portal DNAnexus (Mountain View, CA) both to download sample information from the Mayo Clinic's Biobank to the BCM-HGSC-CL prior to sequence generation and to upload each sample's sequencing results in the form of a text report file, gvcf, richly annotated VCF and BAM files for download to the Mayo Clinic PGL for *CYP2D6* processing, curation of difficult genotypes or rare alleles and data archiving. Two different formats for reporting results were used: the Mayo PGL methodology to populate discrete fields in the EHR to allow the use of best practice advisories, and the OneOme (Minneapolis, MN) report, which was a user friendly report focused primarily on binning of actionable and informative pharmacogenomic drug information. Both result formats were available to study participants as part of an ongoing study designed to understand the impact of pharmacogenomic testing on the attitudes and behaviors of patients.

The PGL methodology incorporated results into the electronic health record (EHR) to populate discrete fields and to fire clinical decision support (CDS) rules in the GE EHR and best practice advisories (BPA) in the EPIC EHR. Phenotype predictions were used to populate automated comments that were gene specific for a limited number of drug comments. Phenotyping was done using published guidelines such as those found in PharmVar for all *CYP* genes (PharmVar.org), PharmGKB for *VKORC1, SLCO1B1* (PharmGKB.org), or per published guidelines for *UGT1A1*[8], *TPMT*[9], and *DPYD*[10]. Phenotype prediction was done in a way that was consistent with that described by Ji et al[11].

For OneOme methodology, after samples were genotyped and reviewed, OneOme leveraged its pharmacogenomics medication response platform to provide interpretation and custom medication reports based on genotypes for the 13 clinically relevant genes included in the "drug-gene pair" alerts that currently fire in the Mayo EHR, including rare variants. The relevant information was securely transferred to OneOme using a specified XML format that provided detailed information about the genetic variants, haplotypes, rare variations, and special comments. Each file was validated to ensure that the data were complete and formatted properly. Once validation was complete, the samples were loaded into OneOme's medication response platform for analysis. After phenotypes were determined for a sample, OneOme's laboratory director reviewed and signed off on the clinical annotation and predictions for the report, and the report was returned to Mayo via SFTP where it was uploaded into the Mayo laboratory systems (SCC SOFT) and sent to the EHR.

Regulatory Management

The testing for this project was designed to be used for clinical and research purposes and to empower clinical decision support; therefore, the testing needed to be CLIA compliant. Because testing and report generation were being performed at three sites (Baylor, Mayo Clinic, OneOme) the testing was performed under a distributive testing model. This was a special situation which required that the medical directors for each of the laboratories had to sign off on the testing and to coordinate comments and interpretations that were included in the testing reports as well as agree on quality metrics and testing standards within the framework of having a CLIA-compliant laboratory developed test. A clear delineation of which

laboratory performed specific parts of the testing and reporting was required and verbiage needed to be in the report, which stated this fact as well as CLIA numbers for the laboratories involved.

REFERENCES

1. Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, Scherer S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet Genomics.* 2016;26(4):161-168.
2. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clin Pharmacol Ther.* 2014;96(4):482-489.
3. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods.* 2007;4(11):903-905.
4. Liang-Chu MM, Yu M, Haverty PM, et al. Human biosample authentication using the high-throughput, cost-effective SNPtrace(TM) system. *PLoS One.* 2015;10(2):e0116218.
5. e MCEaabe, e MC. Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network. *Am J Hum Genet.* 2019;105(3):588-605.
6. Major E, Rigo K, Hague T, Berces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLoS One.* 2013;8(11):e78410.

7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.

8. Canu G, Minucci A, Zuppi C, Capoluongo E. Gilbert and Crigler Najjar syndromes: an update of the UDP-glucuronosyltransferase 1A1 (UGT1A1) gene mutation database. *Blood Cells Mol Dis.* 2013;50(4):273-280.

9. Appell ML, Berg J, Duley J, et al. Nomenclature for alleles of the thiopurine methyltransferase gene. *Pharmacogenet Genomics.* 2013;23(4):242-248.

10. Amstutz U, Henricks LM, Offer SM, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing: 2017 Update. *Clin Pharmacol Ther.* 2018;103(2):210-216.

11. Ji Y, Skierka JM, Blommel JH, et al. Preemptive Pharmacogenomic Testing for Precision Medicine: A Comprehensive Analysis of Five Actionable Pharmacogenomic Genes Using Next-Generation DNA Sequencing and a Customized CYP2D6 Genotyping Cascade. *J Mol Diagn.* 2016;18(3):438-445.

Supplementary Table 1. The number of PGx research protocols that have been reviewed and granted access to the Mayo-Baylor RIGHT 10K PGx data.

Disease Category	Number of Studies	Drugs or Drug Class
Cardiovascular Disease	7	Statins, anticoagulants and antihypertensives
Pain	5	Opioids and other analgesics
Gastroenterology	4	Proton pump inhibitors, immunomodulators and biologics
Infectious Disease	3	Antibiotics and antifungals
Bioethics	2	All drug classes
Population Health	2	Gene-drug association studies broadly
Psychiatry	2	Antidepressants and CNS stimulants
Diabetes	1	Metformin and other oral antidiabetic medications
Oncology	1	Cancer chemotherapy
Ophthalmology	1	Steroids and alpha-blockers
Surgery	1	Antiemetic medications
Women's Health	1	Hormone therapy