ARTICLE

# Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network

The eMERGE Consortium*,*

The advancement of precision medicine requires new methods to coordinate and deliver genetic data from heterogeneous sources to physicians and patients. The eMERGE III Network enrolled >25,000 participants from biobank and prospective cohorts of predominantly healthy individuals for clinical genetic testing to determine clinically actionable findings. The network developed protocols linking together the 11 participant collection sites and 2 clinical genetic testing laboratories. DNA capture panels targeting 109 genes were used for testing of DNA and sample collection, data generation, interpretation, reporting, delivery, and storage were each harmonized. A compliant and secure network enabled ongoing review and reconciliation of clinical interpretations, while maintaining communication and data sharing between clinicians and investigators. A total of 202 individuals had positive diagnostic findings relevant to the indication for testing and 1,294 had additional/secondary findings of medical significance deemed to be returnable, establishing data return rates for other testing endeavors. This study accomplished integration of structured genomic results into multiple electronic health record (EHR) systems, setting the stage for clinical decision support to enable genomic medicine. Further, the established processes enable different sequencing sites to harmonize technical and interpretive aspects of sequencing tests, a critical achievement toward global standardization of genomic testing. The eMERGE protocols and tools are available for widespread dissemination.

## Introduction

The identification, interpretation, and return of actionable clinical genetic findings is an increasing focus of precision medicine. There is also growing awareness that the discovery of genes underlying human diseases is dependent upon access to samples from carefully phenotyped individuals with (and without) clinical conditions. As clinical visits provide the ideal opportunity to record patient phenotypes, with appropriate consent, the medical care of specific patient groups can drive the accumulation of clinical data and knowledge of the genetic underpinnings of disease and the penetrance of DNA risk variants. This "virtuous cycle" of data flow from the bench to the bedside and back to the bench will be a key driver of progress in genetic and genomic translation.

While conceptually straightforward, there are many challenges that must be overcome for integrating clinical and research agendas across global populations. Clinical visits are often brief, focused upon measurement related to specific symptoms and constrained by fiscal and practical concerns. On the other hand, ascertainment for research is often open ended, longitudinal, and accompanied by rigorous consent procedures. The types of data that are recorded for each purpose can be different in both depth and quality. As a result, ideal research and clinical records often diverge.

The current phase (III) of the United States National Institute of Health's Electronic Medical Records and Genomics (eMERGE) program (see Web Resources) aims to study and improve these processes for coordinated delivery of clinical and research data, in a multi-center network, while providing actionable genetic results derived from a next-generation sequencing platform to eMERGE research participants. In previous phases, the network sampled data from large collections of volunteers (>100,000) for research and discovery purposes, as well as to establish parameters that might influence clinical data reporting. In the current phase, multiple clinical collection sites with access to predominantly healthy participants, who are willing to undergo genetic testing and to have their results returned by their physicians, were identified. The network used the opportunity to build upon experience with participant consent, to obtain clinical data from the EHR, and to return genetic testing results.[1] The program addressed challenges arising from the heterogeneity of collection sites and tools used to collect patients' and participants' data. Points of standardization were established (Table 1) and overcame obstacles of use of different instruments and molecular reagents at different sites.

Addressing these challenges advanced precision medical care by standardizing methods for phenotyping, sequencing, and genetic variant interpretation. Further, the harmonized flow, storage, and management of data provided a cohesive vehicle to access data to facilitate research while maintaining respect for patient privacy (e.g., HIPAA laws) and the ability to return important clinical findings to individuals.

## Subjects and Methods

More details of certain methods are included in the Supplemental Subjects and Methods.

*Correspondence: agibbs@bcm.edu (Richard A. Gibbs), hrehm@mgh.harvard.edu (Heidi L. Rehm)
https://doi.org/10.1016/j.ajhg.2019.07.018.

**Table 1. Items Harmonized across the Two Sequencing Centers**

| Item | Challenge | Comments |
|---|---|---|
| Collection sites | sample type | agreed to blood[a,b] |
| | sample quality | minimal quantity specified[a] |
| | intake formats | standard tables supplied to sites |
| | phenotypes | not shared unless indication for testing |
| | patient ID structure | naming conventions |
| | indications for testing | selected 40 "hard coded" |
| Assay development | gene targets | selected by consensus |
| | capture strategy | agreed exons (+/− 15 bases)/SNPs; capture probes spanned min 100 bases |
| | capture reagents | two platforms supported (Nimblegen and Illumina Rapid Capture) |
| | Sanger validation | rare variants always Sanger validated; for common SNVs, stopped validation after 5 confirmations |
| | CNV validation | all CNVs by orthogonal technology |
| Validation/proficiency | technical performance/coverage | min standards (200×; 95% coverage, etc.) |
| | ongoing proficiency | interlaboratory exchange or eMERGE samples and use of standard CAP NGS PT |
| Primary analysis | CNV calling parameters | 3+ exons |
| | pharmacogenomics | report variants and inferred diplotypes |
| Variant classification | initial harmonization | required harmonization of all medically significant differences observed 5 or more times in tested genes |
| | ongoing classifications | required consensus between labs or elevation to Clinical Annotation WG for network consensus |
| Report content[a] | consensus content | 67 genes and 14 SNVs |
| | site-specific genes and SNVs | see Figure 4 and Table S7 |
| | updates | variant reclassifications provided |
| Data delivery | physician clinical reports | PDFs, consumable xml structure; GeneInsight |
| | network access to interpreted variants and de-identified reports | GeneInsight de-identified case repository, DNAnexus Commons |
| | community data sharing | dbGaP and ClinVar submissions |
| Progress reporting | specimen progress | sequencing and reporting timelines |
| | aggregate statistic reporting | rates of secondary findings; detection rates for indications |

[a]Exceptions contributed to extended TAT
[b]BCM-HGSC accepted saliva from some sites for a predetermined number of samples

## eMERGEseq Panel Overview

### *Panel Design and Content*

A gene panel comprising a total of 109 genes and 1,551 SNV sites was developed with input from eMERGE site investigators. The design process considered potential actionability of findings and local research interests, as well as gene size. The 109 genes included 56 based upon the American College of Medical Genetics and Genomics (ACMG) actionable finding list.[2] Additionally, each site nominated 6 genes relevant to their specific aims, including discovery-focused genes with varying degrees of evidence for association with clinical phenotypes in need of further study. All nominated genes apart from titin (*TTN* [MIM: 188840]), which was excluded due to its large size, were included in the final panel design for a total of 109 genes. Further, eMERGEseq content included several categories of single-nucleotide variants (SNVs): (1) ancestry informative markers and QC/fingerprinting loci (n = 425), (2) a suite of SNVs selected to inform HLA type (n = 272), (3) pathogenic SNVs in genes not included on the panel for which return of results was planned (n = 14), (4) pathogenic or likely pathogenic SNVs in genes not included on the panel for which return of results was not planned (n = 55; for some, penetrance is poorly understood), (5) SNVs related to site-specific discovery efforts (n = 718), and (6) pharmacogenomic variants (n = 125), selected based on potential actionability, allele frequency, and space available on the platform. A summary of all eMERGEseq content can be found in Tables 2 and 3, with additional details provided in Table S1. All sequence and SNV data are shared across the network for research, and a subset of the content, namely the clinically actionable variants associated with disease

**Table 2. List of 109 eMERGE Genes, PGx, and Actionable SNVs**

| Disease Category | Gene |
|---|---|
| Cancer susceptibility and tumor diseases | *APC*, *BLM* (rs113993962), *BMPR1A*, *BRCA1*, *BRCA2*, *CHEK2*, *MEN1*, *MLH1*, *MSH2* (including rs193922376), *MSH6*, *MUTYH*, *NF2*, *PALB2*, *PMS2*, *POLD1*, *POLE*, *PTEN*, *RB1*, *RET*, *SDHAF2*, *SDHB*, *SDHC*, *SDHD*, *SMAD4*, *STK11*, *TP53*, *TSC1*, *TSC2*, *VHL*, *WT1* |
| Cardiac diseases | *ACTA2*, *ACTC1*, *ANK2*, *CACNA1C*, *DSC2*, *DSG2*, *DSP*, *GLA*, *KCNE1*, *KCNH2*, *KCNJ2*, *KCNQ1*, *LMNA*, *MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *PKP2*, *PRKAG2*, *RYR2*, *SCN5A*, *TMEM43*, *TNNI3*, *TNNT2*, *TPM1* |
| Cholesterol and lipid disorders | *ANGPTL3*, *ANGPTL4*, *APOA5*, *APOB*, *APOC3*, *LDLR*, *PCSK9*, *PLTP*, *SLC25A40* |
| Endocrine disorders | *CYP21A2* (rs6467), *HNF1A*, *HNF1B*, *MC4R*, *PON1* |
| Connective tissue disorders | *COL3A1*, *COL5A1*, *FBN1*, *MYH11*, *MYLK*, *SMAD3*, *SLC2A10*, *TGFBR1*, *TGFBR2* |
| Neuromuscular diseases | *CACNA1A*, *CACNA1B*, *CACNA1S*, *RYR1* |
| Inborn errors of metabolism | *ACADM* (rs77931234), *ALDOB* (rs77931234), *BCKDHB* (rs386834233, rs79761867), *FAH* (rs80338898), *G6PC* (rs1801175), *CPT2* (rs397509431), *OTC*, *MTHFR* |
| Immunological/inflammatory disorders | *IL33*, *IL4*, *MEFV* (rs28940579, rs61752717), *TNF*, *TYK2* |
| Neurological/psychiatric disorders | *APOE*, *ATM*, *ATP1A2*, *GRM1*, *GRM2*, *GRM5*, *GRM7*, *GRM8*, *NTRK1*, *SC1NA*, *SCN9A*, *TTR* |
| Respiratory disorders/hypertension | *BPMR2*, *CFTR*, *CORIN*, *SERPINA1* |
| Renal disorders | *CFH*, *UMOD* |
| Skeletal disorders | *TCIRG1*, *VDR* |
| Other | *F5* (clotting disorder; rs6025), *FLG* (dermatological), *HFE* (iron storage disorder; rs1800562), *TCF4* (Pitt-Hopkins syndrome), *TSLP* (association with many complex disorders) |
| PGx SNVs | *CYP2C9* (rs1799853, rs1057910), *CYP2C19* (rs12248560, rs28399504, rs41291556, rs4244285, rs4986893, rs56337013, rs72552267, rs72558186), *TPMT* (rs1142345, rs1800460, rs1800462, rs1800584), *SLCO1B1* (rs4149056), *IFNL3/IFNL4* (aka IL28B; rs12979860), *VKORC1* (rs9923231), *DPYD* (rs67376798, rs3918290, rs55886062) |

ACMG56 genes are indicated in *italics* without underlining, consensus site non-PGx TOP-6 genes are *underlined*, non-consensus TOP-6 genes are *double – underlined*, and actionable SNVs are indicated by their rs number.

or drug response, are included in clinical reports for return to the participants.

## Panel Sequencing

### *Reagents*

The gene and SNV list was used to direct construction of targeted capture platforms at two sequencing centers (SCs): The Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC), Houston TX and the Broad Institute and Partners Laboratory for Molecular Medicine, Cambridge, MA. Broad used Illumina Rapid Capture probes for this panel and the BCM-HGSC used Roche-Nimblegen methods. Each group created in-solution capture probes spanning the entire targeted regions of the eMERGEseq panel. Probes were designed to be complementary to specified exons or SNV sites with a minimum span of 100 nucleotides. Tiling was limited to exonic sequence, and analyses included ±15 intronic flanking bases (Figure S1).

### *Sample Preparation*

Clinical sites were requested to submit 2 μg of extracted DNA within a concentration range of 30–50 ng/μL. Although DNA derived from blood was the specified sample for the program, BCM-HGSC revalidated the clinical assay and accepted saliva as a DNA source for a limited number of cases due to clinical site requirements. Once received by the sequencing center, specimens were quantified using a picogreen assay, and quality was assessed by gel. Specimens with a minimum of 600 ng of DNA that did not display high levels of degradation passed sample QC and were accepted for eMERGEseq testing.

### *Ethics Approval and Consent to Participate*

All 11 sample collection sites consented participants under Institutional Review Board (IRB)-approved protocols and the two sequencing centers had IRB-approved protocols that deferred consent to the participating sites. Protocol numbers are as follows: Partners Healthcare (2015P000929), Baylor College of Medicine (#H-40455).

### *Sequencing and Primary Analysis*

Samples from DNA capture using the custom capture reagents were sequenced using standard Illumina technologies. Post-sequence processing at each site utilized preferred alignment and variant calling algorithms. The variant calling pipeline at Broad incorporates Picard deduplication, BWA alignment, and GATK variant calling for SNVs and short indels.[3] At the BCM-HGSC, alignment using BWA-MEM and variant calling using Atlas were instantiated within the Mercury Pipeline.[4]

### *Panel Fill-in*

A common set of reference samples were initially sequenced at each SC. The chosen parameters to monitor performance were coverage of targeted sequence and percentage of the targeted bases at or above 20× coverage. Both groups sequenced cohorts of control samples and identified systematically poorly covered bases as those with less than 20× coverage in >10% of tested samples. Based on this conservative threshold, both groups went through a process of enriching with more targeting probes ("fill-in"), to boost underperforming regions, prior to final validation. The reagent performance is described in Table 4, with additional details in Table S2.

**Table 3. Additional Information on eMERGEseq SNVs**

| SNV Category | Total |
|---|---|
| Ancestry | 241 |
| Fingerprinting | 184 |
| Pharmacogenomics | 125 |
| HLA (imputed) | 272 |
| Actionable clinically significant (P/LP) | 14(see above for more details) |
| Non-actionable clinically significant (P/LP) | 55 |
| Non-actionable, not clinically significant (VUS and below) | 660 |
| TOTAL | 1,551 |

#### *Copy Number Variant (CNV) Calling*

CNV calling at Partners/Broad was performed using VisCap, which infers copy number changes from targeted sequence data by comparing the fractional coverage of each exon in a gene to the median of these values across all samples in a given sequencing run.[5] BCM-HGSC CNV calls were made via Atlas-CNV, an in-house software that combines outputs from XHMM[6,7] and the GATK DepthOfCoverage tool.[6] Like VisCap, Atlas-CNV infers the presence of CNVs from normalized coverage differences to other samples in the same sequencing batch and refines these predictions with a pair of quality control metrics.[8] CNV calls were confirmed by orthogonal technology: Droplet Digital PCR (Bio-Rad) at Partners/Broad and Multiplex Ligation-dependent Probe Amplification (MRC-Holland) at the BCM-HGSC. Detected CNVs were filtered based on the clinical site's gene reporting preferences and ClinGen haplosensitivity and tri-sensitivity scores (see Gene Dosage Curations in Web Resources) and then manually reviewed. Partners/Broad required a minimum of three contiguous exons for reporting; BCM-HGSC required two.

#### *Analytical Validation*

To validate sensitivity, specificity, and reproducibility of the eMERGEseq panel, the performance of both SCs was compared using a common reference sample (NA12878). In addition, each group separately examined previously tested clinical samples, containing known pathogenic variants that were uniquely available to their laboratory. Subsequent additional validation analyses were performed to accommodate lower DNA input amounts, based on sample availability (BCM-HGSC).

#### *Ongoing Proficiency*

Ongoing proficiency testing to monitor laboratories' continuing performance for the eMERGEseq panel involved interlaboratory exchange of previously tested eMERGE samples as part of a proficiency testing program for general sequencing platforms, with all results being concordant to date (see the Supplemental Subjects and Methods for further details).

### Variant Interpretation

#### *General Approach to Interpretation*

Variant classifications from both laboratories were based on ACMG/Association of Medical Pathology (ACMG/AMP) criteria[9] with ClinGen Sequence Variant Interpretation Working Group modifications as well as additional specifications for some of the eMERGEseq genes as established by ClinGen Expert Panels (see Sequence Variant Interpretation in Web Resources). Additional local data accrued from previous case studies were combined with manual literature and public data review for final decisions. Non-ACMG 56 genes underwent an in-depth clinical curation effort using the ClinGen framework for gene-disease validity assessment,[10] followed by actionability assessment by the eMERGE Clinical Annotation Working Group (WG). The WG included more than 6 active MDs (including clinical geneticists) and more than 6 members with clinical laboratory genetics training, among approximately 50 members. The gene-disease pairs were presented at in-person and teleconference meetings attended by WG members of each site. The WG created the consensus list that all sites considered actionable.

#### *Legacy Variant Interpretation*

In order to harmonize prior interpretations and to assess likely ongoing differences, the BCM-HGSC and Partners LMM exchanged data from 1,047 previously interpreted variants in the 109 eMERGE genes and evaluated discrepancies (see Results).

#### *Ongoing Harmonization*

Monthly data exchanges identified any differences of interpretation of non-PGx variants intended for clinical reporting. These discrepancies were reviewed during a bi-weekly interpretation/harmonization teleconference call. Cases of unresolvable variants were presented to the eMERGE Clinical Annotation WG to attempt resolution and/or track their occurrence. All reported variants are submitted to ClinVar along with their interpretations.

#### *Pharmacogenomics (PGx)*

The SCs worked with the eMERGE PGx working group to select variants to be included on the clinical reports provided to participants, to interpret diplotypes, and to select drugs for therapeutic recommendations, guided by the Clinical Pharmacogenomics Implementation Consortium (CPIC) guidelines (Web Resources). Twenty PGx variants in seven genes were deemed to be clinically actionable and were therefore selected for return to participants. Table S3 includes details of the PGx genes and variants reported and the drugs associated. For two PGx genes, *CYP3A5* (MIM: 605325) and *SLCO1B1* (MIM: 604843), the gene panel included only one of three variants discussed in the CPIC guidelines. *CYP3A5* was deemed not reportable, as two SNVs important for predicting phenotype for African Americans and Latinos are not included on the gene panel. *SLCO1B1* was deemed reportable, as the one SNV included in the panel serves as a tag SNV for the remaining two SNVs.

The BCM-HGSC included PGx results on individual patient reports, while Partners LMM produced a batch report that accommodates one to hundreds of patients for bulk consumption and EHR integration by sites. Sample PGx reporting formats can be found in the supplemental data (Sample Clinical Reports and Table S8). The CPIC drugs that were included in the PGx report were largely the same with some minor differences (see Table S3).

### Data Management

#### *Sample Intake*

Each site was provided barcoded tubes by the SC for DNA shipping. Sample identifiers and metadata were uploaded using an "eMERGE requisitioning sheet" via secure portals (see eMERGE Sample Submission Portal and Clinical Research Sequencing Platform in Web Resources). The requisitioning spreadsheet contains fields for sample information (name [optional], sex, date of birth/age, US state of residence, site-specific ID), as well as eMERGE-specific metadata including patient "disease area" (from a list defined by the network, see Table S4 details), disease status

**Table 4. Assay Performance and Optimization at the Sequencing Sites**

| | BCM-HGSC | | | Broad | | |
|---|---|---|---|---|---|---|
| | Acceptance Criteria | Original | Low Input | Acceptance Criteria | Measured at ~250× MTC | Measured at ~400× MTC |
| Assay sensitivity (SNV + indel) | | 100% | 100% | ≥95% | 100% | 100% |
| Assay sensitivity (CNV) | | 97.7% | 98.3% | n/a | 100%[a] | n/a |
| Assay specificity (point variant + indel) | | 100% | 100% | ≥95% | 100% | 100% |
| Assay reproducibility | ≥95% | >98% | >97% | ≥95% | 98.5% | 99.6% |
| % of >20× coverage for targeted regions | ≥99% | >99% | >99% | ≥95% | 99% | 99% |
| Depth of mean coverage | >200× | >200× | >200× | n/a | ≥250× | ≥400× |

[a]CNV sensitivity at Broad/LMM is for events ≥3 consecutive exons.

and test indication, eMERGE project ID, and barcode number on the tube. An additional option was to add phenotype terms in a free-text field, primarily based on the MonDO ontology and occasionally additional local codes largely derived from Human Phenotype Ontology (HPO) terms (Table S4). A simple .csv file structure was used by both SCs so that sites could upload all metadata at the time of sample batch shipment. For the BCM-HGSC SC, the sample accession was directly into a cloud environment, managed by DNAnexus, while for the Partners-Broad SC, a custom portal operating in the Broad's local environment was employed for intake followed by transfer to the GeneInsight system for analysis and reporting, with all systems being HIPAA compliant. Local identifiers were then generated to track the samples as they progressed through DNA sequencing and variant calling. Orders were reviewed and approved by the SCs prior to sample shipping and accession. Upon receipt, the samples were subjected to volume and concentration quality control checks.

#### *Data Delivery and Reporting*

Each SC developed custom reporting methods (see Supplemental Subjects and Methods for examples). Partners/Broad site users have a unique, password-protected account and are able to view only orders and metadata from their own site. The Broad portal authorization procedures are customized to allow for secure transfer of sequencing output files and metadata to both Partners and DNAnexus via APIs. The BCM-HGSC sites are delivered reports from the DNAnexus environment via DNAnexus APIs. Users were provided individual logins for accessing PDF reports and structured content in a harmonized .xml format.

#### *GeneInsight*

Partners/Broad sites used the commercial tool, GeneInsight (Sunquest Information Systems), for local report management.[11] This tool was configured to create a De-identified Case Repository (DCR) which contains a de-identified record of all cases and associated variants from both Partners/Broad and the BCM-HGSC supported sites.

#### *DNAnexus Data Commons*

The BCM-HGSC clinical sites were provided with two data access points in the DNAnexus infrastructure. One provides a restricted space for accessing protected health information (PHI)-containing clinical reports, while another acts as a general space for the de-identified records of each case and associated variants. Users were provided individual logins and selectively granted access to one or both access points. Data for sites that were served by the BCM-HGSC were provided both .xml and .pdf formats, at the time of reporting. De-identified, structured versions of the Partners-Broad reports are downloaded from the DCR and also stored in the DNAnexus Data Commons projects, creating a comprehensive repository of de-identified clinical reports.

#### *Variant Updates*

Two complementary mechanisms were developed to enable delivery of variant updates from the SCs to the sites as new evidence leading to a classification change becomes available. At Partners/Broad, individual participant results are stored in an eMERGE-specific instance of the GeneInsight database that is linked to Partners LMM's GeneInsight instance enabling communication of variant updates.[12] If Partners updates a variant, sites that have signed up receive proactive notification emails if a reported variant identified in one or more of their cases is updated. Hyperlinks are provided in those emails that allow sites to directly access updated information on the variant in each case, which facilitates the choice to return an updated result to a participant. In addition, Partners is generating an .xml file for each variant interpretation change alert, which sites can consume through other electronic interfaces. At the BCM-HGSC, participant results are stored in a database that is routinely queried for variants with new actionable interpretations. If such a variant is found in a previously reported sample, an amended report is issued via DNAnexus and sites are notified. Variant updates are included in the ongoing variant interpretation harmonization process described above.

### eMERGE III Samples and Raw Data Storage

Results were analyzed from the eMERGE III eMERGEseq data, which consisted of 25,015 samples. These included 14,515 from Baylor and 10,500 from Partners-Broad. The associated BAM, xml, and vcf files are available on the eMERGE Commons, accessible to sites as well as outside investigators who apply for access (see eMERGE Network in Web Resources). Data are also being submitted to dbGaP for controlled public access (phs001616.v1.p1).

## Results

### Network Overview

The eMERGE III network established a Clinical and Discovery Platform that consists of 11 clinical study sites, 2 DNA SCs, and a coordinating center (CC) (Figure 1). Participants were enrolled at each site, where blood was collected, and DNA was extracted locally and sent to one of two SCs for targeted sequencing. Analysis and interpretation of the DNA sequence data was performed at each SC, and the data were returned to the clinical sites for return to

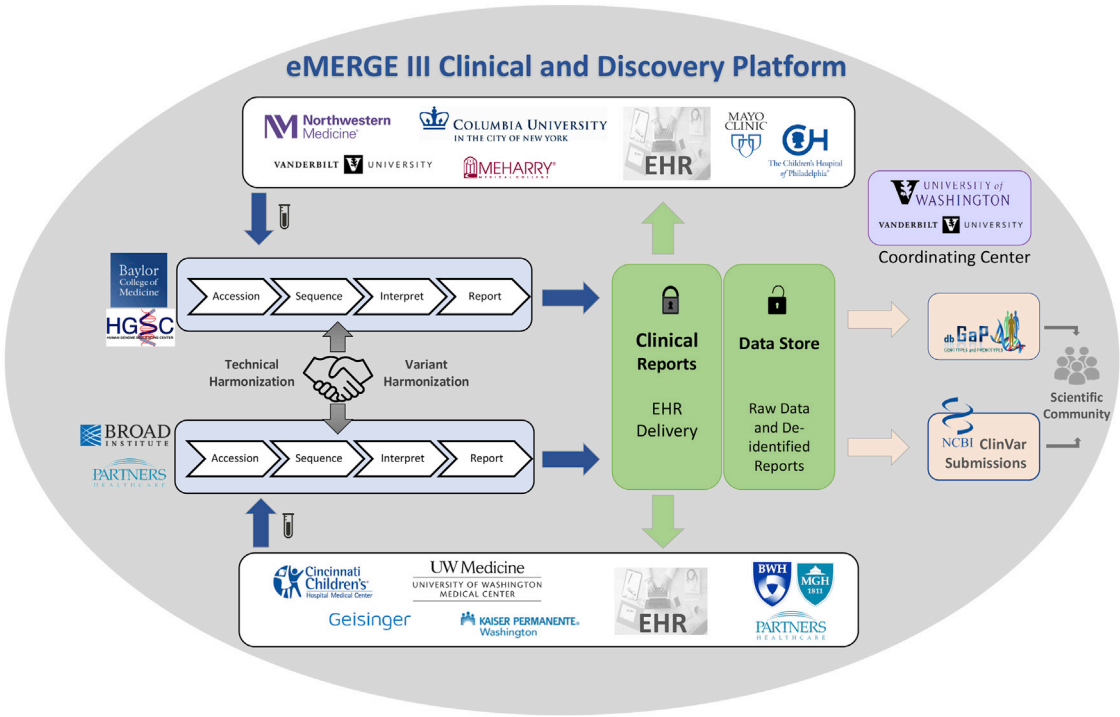



**Figure 1. eMERGE III Network Overview**
The eMERGE III network is comprised of 11 study sites, 2 sequencing centers (SCs), and a coordinating center (CC). The different components and processes involved in the data flow across both the clinical and research discovery arms of the network are highlighted in this figure and described in more detail in Network Overview.

participants. Raw data were accrued for data mining purposes by eMERGE investigators and approved affiliates. Subsequently, raw data are released to dbGaP and interpreted variants to ClinVar.

An early decision of the program was to utilize DNA capture "panels" of approximately 500 kb, in order to generate genomic data from the eMERGE participants, as an alternative to whole-exome sequencing (WES) or whole-genome sequencing (WGS). This choice reflected a balance between available fiscal resources and a reasonable selection of content to explore return of actionable results and focused discovery efforts. It should be noted that there are other efforts within eMERGE to support discovery from research platforms, including more than 100,000 GWAS arrays and more than 5,000 exome and genome sequences generated to date. However, this effort was distinct in focusing on a CLIA platform intended for clinical return of actionable results. The use of the panel enabled testing of 109 genes and 1,551 additional sites of single-nucleotide variation in each sample. Across the network, ~25,000 samples were assayed, ~2,500 from each site (Table S5). The study is therefore large enough to allow robust analysis of specific phenotypes as well as to gain experience with a sufficient number of patients at each site to develop processes to support the return of actionable genetic results.

Prior population studies suggested that the genes included on the panels would reveal thousands of newly identified single-nucleotide and structural variants. A small subset of these would be expected to be pathogenic, and the program aimed to report to participants only those variants that were pathogenic or likely pathogenic according to the ACMG/AMP guidelines[9] or those with actionable pharmacogenomic associations. Each site would have the option of a customized clinical reporting framework, as well as full access to all network data to guide decisions and harmonize interpretations.

This elaborate network reflects a real-world situation, where a full complement of testing, reporting, and research require coordination and harmonization of many components. First, the selection of gene targets and the rules for reporting must agree. Next, the technical aspects of DNA capture and sequencing required standardization and ongoing comparison. The DNA changes must be interpreted and reported with the same conclusions, regardless of where testing occurred. Finally, file structure standardizations and data management practices must be organized. A detailed list of components (Table 1) that require coordination and harmonization illustrates the magnitude of the challenge.

## Technical Validation of Capture Panels

Coordination and harmonization of the DNA capture panel process at the two CAP/CLIA-certified DNA sequencing laboratories was demanding because in addition to different DNA capture reagents, the local processes of sample preparation, library construction, hybrid capture, and sequencing represented complex workflows with many variables. As an alternative to compelling

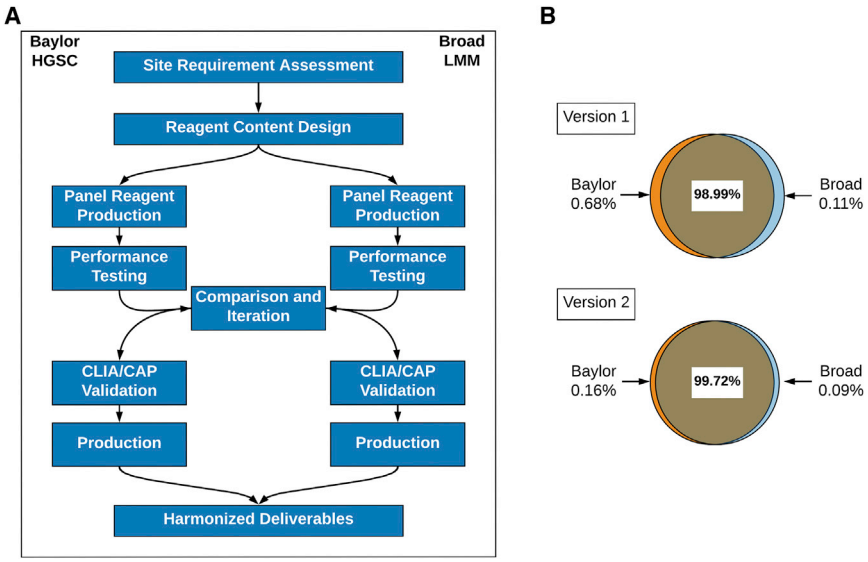



**Figure 2. eMERGEseq Panel Test Development and Validation**

(A) Technical harmonization of two DNA capture panels. Coordination and harmonization of all the components of the DNA gene capture panel process at the two sequencing centers.

(B) Base coverage. Percentage of bps covered ≥20× across sequencing centers. Percent of bases in the panel targeted region covered in each version of the panel design and the extent to which these bases overlap between the genome centers is shown. Version 2 is the final version used for data generation.

each laboratory to adopt unfamiliar methods, the harmonization was achieved through phases of coordinated design, comparing initial high-level technical performance, and via ongoing monitoring of proficiency (Figure 2A). The harmonization process aimed to reduce any impact on the overall program due to the heterogeneity of capture reagents or sequencing methods between two sites and for the end users to be able to compare data from each laboratory without batch effects.

Design was coordinated by first agreeing on the intended limits to reporting, e.g., number of bases adjacent to exons to be reported (see Subjects and Methods and Figure S1). Each laboratory employed slightly different criteria for the selection of the range of transcripts to be tested, reflecting a lack of harmony of public databases. Possible differences in design were resolved by selection of the union of all possible exons to be considered and validated by iterative sharing of the capture design files ("bed files"). The detailed design specifications can be found in Table S1.

Preliminary testing of the technical performance of the two capture reagents utilized both local test samples and a shared sample reference set (see Subjects and Methods). The technical performance was shared between the SCs by measuring the coverage of individual bases and other key technical metrics (Tables 4 and S2). Overall sequence coverage goals and the extent to which poorly covered regions could be tolerated were agreed upon *a priori,* and the technical comparison was straightforward between SCs. In general, the sequencing reagents performed well, although the presence of some uncovered bases in the first panel designs led each group to modify the initial reagents to optimize performance (Figure 2B). Throughout, the comparative performance of the two reagents informed the progress of technical development and illustrated the synergism from closely monitoring similar processes.

For final validation, both groups measured overall sensitivity and specificity on a reference sample (NA12878) as well as sensitivity to detect known pathogenic variants from previously tested clinical samples that were uniquely available to them. Groups also incorporated evaluation of variance in processing including varying coverage from ~250× to 400× (Broad) and input amounts of 250 ng and 500 ng (Baylor). Summary results of the respective validation studies are shown in Table 4. Panel optimization results and coverage analyses can be found in Table S2. The impact of the ~0.2% of targeted bases that were not effectively covered via the optimized panel designs was evaluated by the network for impact on clinical decision making. The majority of missing data was judged to be of little consequence although small regions of some genes (e.g., *RYR1* [MIM: 180901], *CACNA1B* [MIM: 601012]) could not be recovered by either platform (Table S2).

Once the data production phase of the program was initiated, the ongoing performance was monitored by sharing production metrics and via the ongoing CAP/CLIA proficiency program that included exchange of samples and comparison of DNA variation data. As of this publication, mean coverage of Broad production samples is ~420×, percent of targeted bases covered ≥20× is 99.7%, and percent of targeted bases with zero coverage is 0.17%. These metrics, collected from >7,000 production samples, closely match the performance of the validation set. Mean coverage of the BCM-HGSC production samples is ~340×, percent of targeted bases covered ≥20× is 99.8%, and percent of targeted bases with zero coverage is 0.04%. These metrics, collected from >9,600 production samples, also closely match the performance of the validation set.

## eMERGE III Cohort

The eMERGEseq cohort is comprised of 25,015 biobank or prospectively recruited participants representing 11 eMERGE sites. These were either unselected for any specific phenotype or were enriched for specific phenotypes depending on site-specific clinical and research interests. A brief summary of the nature of each site-specific sample repository, including the total number of participants per

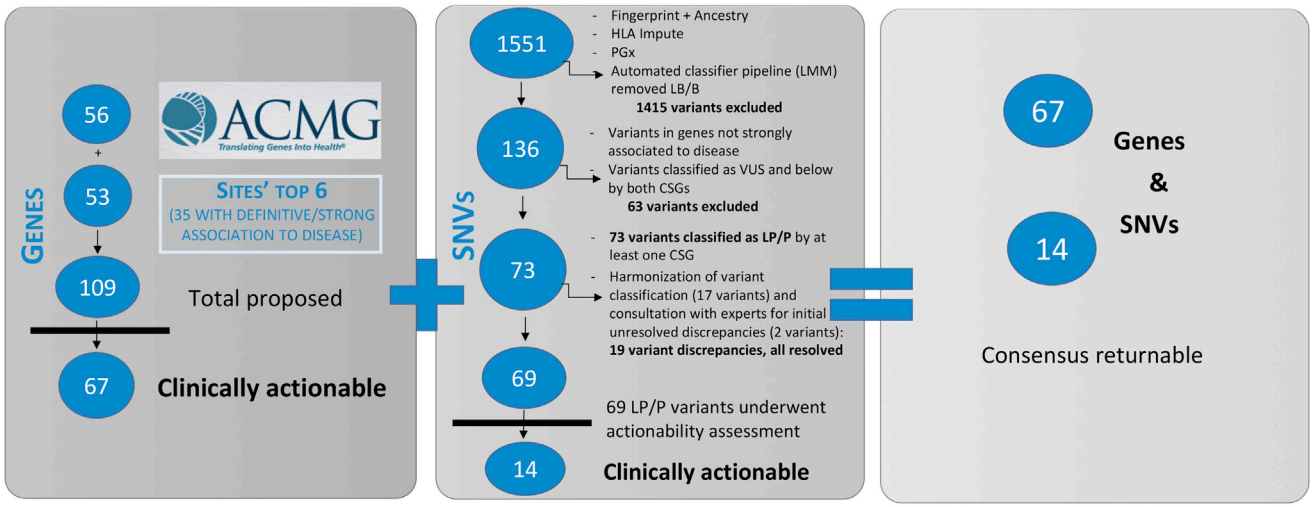



**Figure 3. Content Development for the eMERGEseq Panel**
Left: ClinGen gene-disease validity assessment for all site top six proposed genes. Those with definite and strong association to disease were considered for further actionability analyses.
Middle: Clinical assessment for a subset of single nucleotide variants (SNVs). Those deemed P/LP were considered for actionability analyses.
Right: Final consensus list of returnable content. This included all the ACMG56 genes, in addition to 11 genes and 14 variants that were deemed actionable by the eMERGE Clinical Annotation Working Group.

site can be found in Table S5. A more detailed description of the clinical cohorts involved in this study, including enrollment criteria, are reported elsewhere (A. Gordon et al., 2018, American Society of Human Genetics, abstract).

### Genetic Ancestry

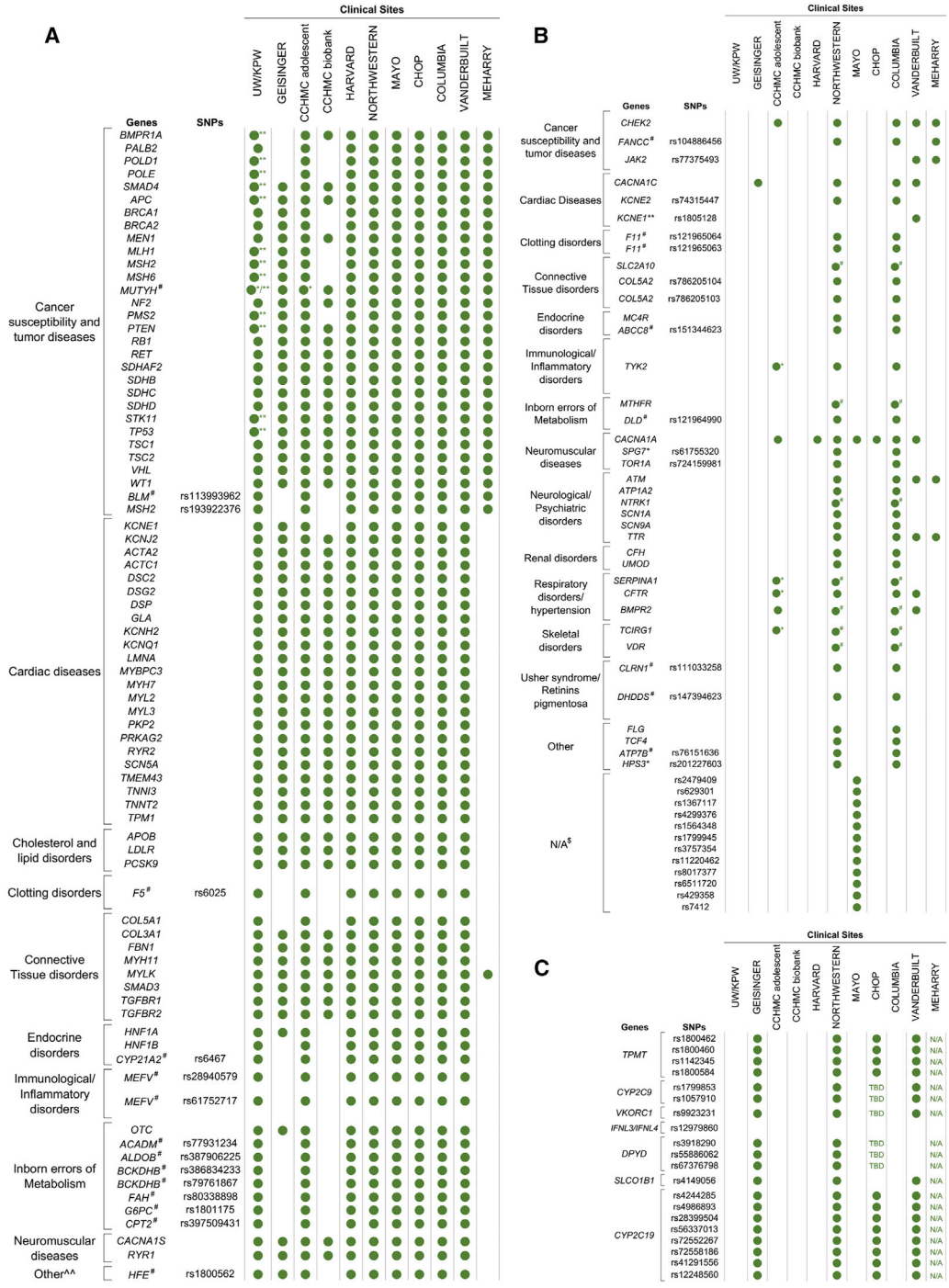
Genetic ancestry within the diverse eMERGEseq dataset was determined by using common variants throughout the eMERGEseq panel, including ancestry informative marker SNVs. Principal component analysis of genetic ancestry (Figure S2) and qualitative comparison to self-reported ancestry (Table S6) were performed as a part of various quality control analyses applied on the cohort. The self-reported race and genetically determined race appear to generally match.

### Clinical Content Validation and Site-Specific Return of Results Plans

Gene selection by sites for inclusion on the eMERGEseq panel was driven by both clinical and research needs leading to a final list for panel design of 109 genes, including the "ACMG56"[2] and 53 additional site selected genes. Evidence review using the ClinGen gene-disease validity framework identified 35 of the additional 53 genes as having definite or strong association to disease. These genes were considered for further actionability analyses (see Figure 3). Most of the 18 genes with lower levels of validity were included by sites to enable research on these genes, reflecting the diverse goals of the eMERGE network including discovery as well as return of results.

A subset of the genotyping SNVs were also evaluated for possible return. This excluded 1,415 SNVs submitted for HLA analyses, fingerprinting, and ancestry typing or already designated for PGx return. Of those remaining, some had been previously classified as likely benign or benign and were thus excluded from further analyses of potential pathogenicity. The remaining 136 variants were considered for further clinical assessment. Seventy-three variants were classified as either likely pathogenic or pathogenic by at least one of the SCs. Of these, 19 had discrepant classifications between the two SCs. These were resolved by variant re-assessment and scoring on published evidence as well as combined internal evidence from both SCs. For two variants, the eMERGE Clinical Annotation WG was consulted to assist in resolving interpretation differences. A final list of 69 pathogenic/likely pathogenic (P/LP) variants was established and further considered for actionability analyses (Figure 3).

The eMERGE Clinical Annotation WG evaluated the medical actionability of the 35 non-ACMG56 genes for which we had applied ClinGen criteria and defined as having at least one strong/definitive disease association, as well as 69 P/LP pathogenic variants, based on whether there was a substantially increased risk of serious disease that could be prevented or managed differently if the risk were known. In addition to the ACMG56, 11 genes and 14 variants were deemed actionable by the eMERGE Clinical Annotation WG and placed on a consensus list of returnable content (Tables 2 and 3, Figure 3). While sites agreed that this list represented content that would generally be medically actionable in adults, some sites did not return results from all genes on the consensus list and/or chose to return additional content based on their research interests, patient populations, and IRB-approved return of results protocols (Figure 4). For example, not all sites chose




# Only bi-allelic (homozygous, compound heterozygous) variants were returned.
* Carrier status was returned.
** VUS's were returned for 13 colorectal cancer genes in this site.
^^ Other include: dermatological (FLG) &
Pitt-Hopkins syndrome (TCF4), Iron storage disorders (HFE),
Copper metabolism (ATP7B), Hermansky-Pudlak syndrome (HPS3).
^ KCNE1 on the gene list (variant considered a risk allele)
$ No clinical interpretation was provided for these variants.
TBD - To be determined; N/A - Not Applicable

**Figure 4. Site-Specific Reportable List of Genes/SNPs for Which Pathogenic or Likely Pathogenic Variants Will Be Returned**
(A) Consensus list of returnable SNPs/genes. Inclusions are indicated with a green dot and exclusions are indicated by no dot.
(B) Site-specific list. Non-consensus genes/SNPs with site-specific inclusions indicated with a green dot
(C) Site-specific PGx list. Pre-determined SNPs for return in PGx genes. Inclusions are indicated with a green dot. UW/KPW, University of Washington/Kaiser Permanente Washington; CHOP, Children's Hospital of Philadelphia; CCHMC, Cincinnati Children's Hospital Medical Center

to return *HFE* (MIM: 613609) p.Cys282Tyr homozygotes. Additionally, of the 11 sites, one that included pediatric biobank participants opted not to report variants in genes that increase risk of adult-onset diseases but are not actionable during childhood. Another site limited its actionable gene-disease pair return list to cancer-associated genes. Four other sites requested return for additional genes and SNVs that were not on the consensus list, again due to study differences. For example, a clinical site whose research included the creation and return of a polygenic risk score requested genotypes at 12 SNP sites associated with low-density lipoprotein cholesterol (LDL-C) risk be included on their report. Another site returned variants of uncertain significance in 13 colorectal cancer (CRC [MIM: 114500])-associated genes for a subset of their samples derived from a cohort of participants with CRC or polyps. A full list of the content that was returned for each site can be found in Table S7 and summarized in Figure 4.

For PGx returnable content, 20 variants in 7 genes were deemed clinically actionable by the PGx working group, yet only 4 sites chose to return PGx results to participants, and for those that did, they did not return results from all genes (Figure 4). For example, none of the sites elected to return diplotypes associated with *IFNL3/IFNL4*. Return of PGx results was in part influenced by which sequencing center was assigned to a site, due to differences in the types of reports being issued (PGx results included on individual patient reports for BCM-HGSC versus separate batched reports with PGx results from Partners-Broad).

### Data Intake and Delivery

Data intake and delivery represented challenges for the network due to the plan to test distributed, heterogeneous EHR systems and other data sources used by sites and the need to deliver updated data interpretations. All demands were required to be met while managing issues of compliance and security for PHI protection. These challenges mimicked real-world situations as these are identical needs for any health care organization opting to interact with a research enterprise or reference laboratory. The data management required the development of three main informatic components: data intake, clinical reporting, and the de-identified case repository and data commons.

Firstly, data intake and accessioning for each site was facilitated by an agreement of the specific PHI metadata to be supplied with each sample, as well as an agreement of a set of required "indications for testing" that represented the primary phenotype data that tracked each sample through the network (see Subjects and Methods).

The second is clinical reporting. Within each pipeline, the standard validated product was a PDF report that was returned to the clinical investigators (see Supplemental Data for examples of reports). Each clinical site had custom requirements for the report content that reflected local preferences for data to be returned to patients. Each SC also had different reporting requirements; for example, some sites requested negative reports, others returned only positive reports.[1] Most sites also requested data in structured formats to enable direct integration onto their local EHRs (see Data S1 and S2 for examples).

The five clinical sites served by the Partners-Broad SCs received results delivered through the GeneInsight platform, which enabled storage and query of clinical reports. The six sites served by the BCM-HGSC utilized custom applications developed for report delivery. Possible difficulties in data sharing between different parts of the network were anticipated and obviated by development of an agreed .xml standard. This standard was based upon the GeneInsight system specifications and facilitated communication across all components (see Subjects and Methods and Aronson et al.[13]). The clinical sites therefore had two options—they could either use a stand-alone tool for report data management or alternatively the report data could be parsed into local customized systems.

For those sites using the GeneInsight platform, automated alerts were delivered immediately upon LMM variant reclassifications that affected an eMERGE report. Most alerts then led to requests for report amendments with a total of 16 amendments delivered by LMM for 7 variant reclassifications to date. In addition, ten amendments were issued by BCM after routine queries for variant updates. For PGx data, in addition to receiving results in PDF reports (either individual reports by the BCM-HGSC or batch reports by Partners-Broad), a standardized data format was also developed to deliver structured PGx data in the form of both variant level and diplotype results allowing sites to directly integrate PGx results into the EHR for clinical decision support.

Finally, the network required all deidentified data to reside together, to enable data mining for both basic research and to better inform clinical decision making with access to larger clinical datasets. There were two independent but complementary mechanisms for this. First, the GeneInsight tool maintains a record of all returned variant data from both sites in a de-identified case repository allowing an easy search interface for clinically reported variants. A second site maintained the full set of eMERGE raw data in a cloud environment, managed by DNAnexus. This "eMERGE Commons" was structured to house each DNA sequence file in the BAM format, as well as the annotations for the data in a vcf format. As clinical report delivery for the data generated in the Baylor SC also utilized the DNAnexus infrastructure, the full set of identified clinical reports and de-identified raw data were both resident in the cloud. The access permissions for the data were managed to allow only the clinical providers to access their patients' clinical reports. The full set of raw data was available to all eMERGE investigators after PHI information had been removed.

### Variant Interpretation Harmonization

To ensure consistency of results being returned across the eMERGE consortium, variant interpretation was harmonized between the SCs (Figure 5). In a pre-test

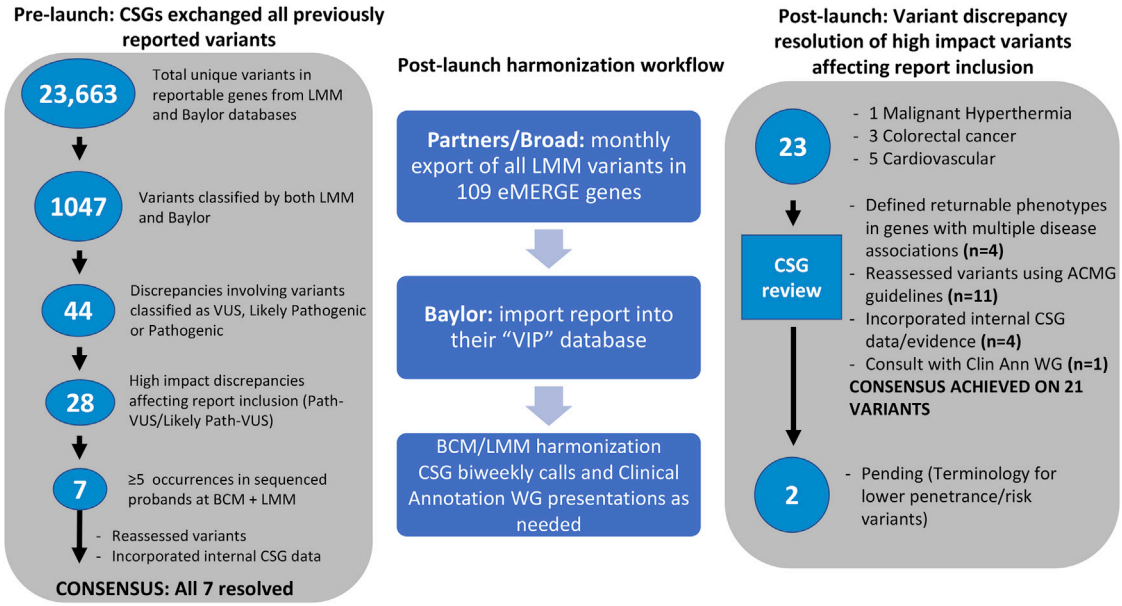



**Figure 5. Variant Harmonization Process Overview**
Pre-launch and post-launch harmonization processes involving the exchange of variants in reportable genes between the sequencing centers and the identification, prioritization, and the resolution of discrepancies affecting report inclusion.

launch, both SCs exchanged variants in reportable genes from their respective databases, totaling 23,663 unique variants. Of those, 1,047 were previously classified by both SCs. The pre-test launch data exchange showed 90% concordance in variant classification among variants classified as VUS, likely pathogenic, and pathogenic by at least one SC. When likely pathogenic and pathogenic variants were grouped together, the concordance was 93%. When all variant classifications were considered, including benign versus likely benign, the data showed a 67.5% concordance. However, only 28, or 3% of the variants were deemed to affect reporting (VUS versus pathogenic 1.9%, VUS versus likely pathogenic 1.1%). The two SCs resolved all differences that would affect inclusion on clinical reports (i.e., P/LP versus VUS).

An ongoing process was also developed to ensure continuous harmonization of variant interpretation (Figure 5). As of May 2018, 23 initial discrepancies of interpretation of variants from five disease areas were considered, based upon potential to affect report inclusion. Most discrepancies in variant interpretation (83%) were immediately resolved when re-assessed by the SCs by using ACMG guidelines, incorporating additional laboratory-specific evidence, after defining returnable phenotypes in genes with multiple disease associations (for example malignant hyperthermia [MIM: 145600] versus myopathy [MIM: 117000] for *RYR1*), or defining terminology for lower penetrance/risk variants. For one variant, resolution required input from additional eMERGE investigators through the eMERGE Clinical Annotation WG.

Three variants (p.Ile1307Lys in *APC* [MIM: 611731], p.Met54Thr in *KCNE2* [MIM: 603796], and p.Asp85Asn in *KCNE1* [MIM: 176261]) were noteworthy as the interpretations were more discrepant upon initial assessment (i.e., "two-steps:" pathogenic versus likely benign), although the evidence used by both centers was identical. These represented variants that have significantly reduced penetrance, leading to difficulties applying the ACMG/AMP classification framework, which is designed primarily for highly penetrant Mendelian disorders. Nevertheless, some sites chose to return the *APC* variant as it imparts a 2-fold risk of CRC in Ashkenazi Jewish individuals, even though its effect in other populations in unclear. Other sites elected to return the *KCNE2* variant, as it has been associated with variable presentations such as arrhythmias (MIM: 611493) and long QT syndrome (MIM: 613693).[14–16] This type of classification discordance highlights the need for guidance on classification terminology for low penetrance variants for not only the eMERGE network but for the entire medical genetics community.

## Aggregate Findings and Return of Results

A total of 8,437,788 variants were detected among the 25,015 case subjects that have been collected and analyzed via the eMERGEseq panel. A subset of these were excluded from further analyses due to a LB/B classification by the SCs or by an auto-classification pipeline based on allele frequency thresholds or for having a low-quality score. The remaining variants underwent a filtration process which returns (1) predicted loss-of-function variants with a minor allele frequency (MAF) $< 1\%$, (2) variants previously classified by the SCs as likely pathogenic (LP)/pathogenic (P) regardless of MAF, and (3) ClinVar P/LP as well as HGMD "DM" variants with a MAF $< 5\%$. This pipeline resulted in 9,653 unique variants requiring further assessment. After expert review, these were further categorized as benign

(1%), likely benign (8%), VUS (69%), LP (7%), P (12%), or deemed as low penetrance risk alleles (0.5%). In addition, 205 unique copy number variants have been detected across the reviewed samples, with 141 gains and 64 losses. Of these, 30% were deemed reportable and were returned to sites. In summary, these data led to a total of 1,497 case subjects that have a LP/P variant that would require a positive report to be issued.

Results being returned to sites currently fall into three categories: (1) indication-based returnable results that include all sequence and copy number variants related to the site-provided indication for testing, (2) non indication-based consensus returnable results that include all sequence and copy number variants in genes and SNVs comprising the consensus list of returnable content (see Clinical Content Validation and Site-Specific Return of Results Plans) that are not related to the indication for testing, and thus considered secondary findings, and (3) non indication-based site-specific returnable results which include variants in additional site-requested genes that are not on the consensus list and not related to the indication for testing. Additionally, both SCs are returning results on pre-selected PGx SNVs as either an addendum to individual patient reports or in a batch report that contains up to ~185 samples (see Subjects and Methods).

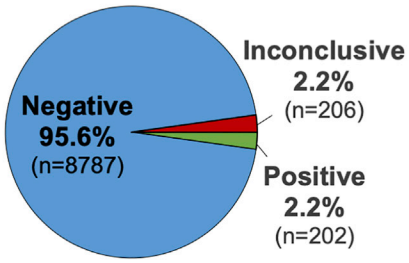
The positive rate for each category of findings is depicted in Figure 6. For all 25,015 case subjects that have been reviewed, 9,195 (37%) had an indication for testing. Of these, 202 (2.2%) had positive findings relevant to the indication for testing (Figure 6A). Moreover, of all individuals sequenced, 1,039 (4.2%) had additional/secondary findings of medical significance in genes and SNVs from the consensus list, that are being returned to participants (Figure 6B). 17,175 participants (69%) were enrolled in sites who were interested in returning pathogenic and/or likely pathogenic variants in additional genes or SNVs that were not on the consensus list. In 265 cases (1.5%), a non-indication based, site-specific returnable pathogenic or likely pathogenic variant was identified (Figure 6C). 37% of these variants were in *CHEK2* (MIM: 604373), a tumor suppressor gene, and are associated with an increased risk for a variety of cancers. A full list of all positive findings returned to participants with and without indications are listed in Table S10.

Other variants from consensus list genes and SNVs that were not related to the indication of testing were associated with cancer, cardiac disease, familial hypercholesterolemia (MIM: 143890, 144010, 603776) and hemochromatosis (MIM: 235200) (Figure 6). For indication-based assessments, detection rates were highest for breast/ovarian cancer (MIM: 114480, 604370, 600185) (39%), hyperlipidemia (28%), and CRC/polyps (19%). Some phenotypes had no disease-causing variants identified due to either the absence of genes causative for the disorders on the eMERGEseq panel or the lack of a clear monogenic disease etiology for the disorder (e.g., abnormality of pain sensation [MIM: 243000], pediatric migraine [MIM: 188840]). The rate of P/LP variants detected in participants without a clinical indication differed from site to site, ranging from 2% to 11%, depending upon the basis for participant selection, which were reflective of the underlying study designs of the individual sites. The overall positive rate for secondary findings was skewed higher for one site (Geisinger), where a subset of participants were preselected for a suspicious variant(s) previously identified in an exome study.[17] On the other hand, two sites had lower rates than expected either because their cohort had an indication related to genes in the secondary findings list that led to the removal of these genes from secondary findings reporting or because the site did not choose to return all results from the consensus list. When data from Geisinger participants preselected for suspicious variants were excluded, the frequency of secondary findings was similar across sites, ranging from 1.8% to 5.1%, suggesting that the complexity of the network did not otherwise distort these results, and reflecting the success of the data and process harmonization. A further analysis of the factors that influence the rate of secondary findings return is underway (A. Gordon et al., 2018, American Society of Human Genetics, abstract).

For PGx results, reports depicting genotype and related diplotype data, including whether the reported diplotype for each gene and resulting phenotype would result in a recommendation to modify dosage, have been issued for all participants from 11 sites. The frequency of the reported diplotypes were concordant with the CPIC published frequency tables for each major race/ethnic group (see CPIC in Web Resources). One difference for diplotype interpretation was particularly illustrative of the role of harmonization. When both rs1800460 and rs1142345 are identified in *TPMT* (MIM: 187680), it cannot be ascertained whether these variants are in *cis*, resulting in a *TPMT*1/*3A* diplotype and intermediate metabolizer phenotype, or in *trans*, resulting in a *TPMT*3B/*3C* diplotype and a poor metabolizer phenotype. One SC emphasized the more common diplotype in their report, while the other emphasized the higher risk of the rarer diplotype under some drug regimens. With input from the sites and the eMERGE PGx working group, it was decided that the more common genotype would be reported with a warning that the rarer genotype could not be ruled out.

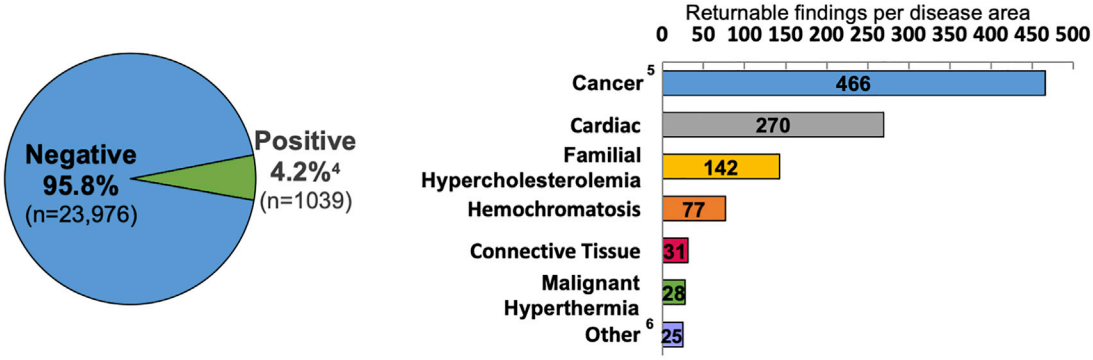
Across the 20 loci (7 genes) and 11 drug types, diplotype analysis prompted recommendation for potential non-standard drug dosing in at least one drug in 93% (23,232/25,015) of participants. Overall, the percentage of participants with actionable PGx results, resulting in a recommendation to potentially adjust standard drug dosing or use of an alternate drug based on their metabolizer phenotype, ranged from 2% (for *DPYD* [MIM: 612779] genotypes associated to response to Fluoropyrimidines) to 57% (for *IFNL3* [MIM: 607402]/*IFNL4* [MIM: 615090] genotypes associated to response to pegylated interferon-α (PEG-IFN-α) and Ribavarin). Site-specific PGx results across all tested genes leading to potential dosage

**A Indication-based returnable results (n=9,195)**

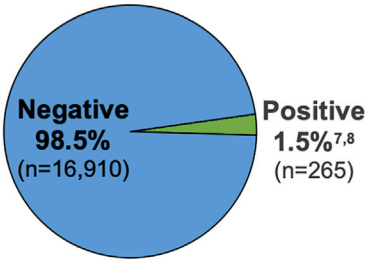



| Indication[1] | Total | Positive | Negative | Inconclusive |
|---|---|---|---|---|
| Colorectal cancer/Polyps | 2593 | **39** | 2348 | 206 |
| Arrhythmia | 188 | **5** | 183 | n/a |
| Asthma | 135 | **0** | 135 | n/a |
| Cardiomyopathy | 16 | **1** | 15 | n/a |
| Chronic Kidney Disease | 1069 | **7** | 1062 | n/a |
| Obesity | 33 | **0** | 33 | n/a |
| Pulmonary Hypertension | 28 | **0** | 28 | n/a |
| Tubular Sclerosis Complex | 12 | **11** | 1 | n/a |
| Abnormality of pain sensation | 678 | **0** | 678 | n/a |
| Autistic Behavior | 102 | **1** | 101 | n/a |
| Ehlers-Danlos Syndrome | 89 | **1** | 88 | n/a |
| Hyperlipidemia[2] | 3057 | **56** | 3001 | n/a |
| Pediatric Migraine | 473 | **0** | 473 | n/a |
| Breast/Ovarian Cancer[3] | 427 | **79** | 348 | n/a |
| at risk for colorectal cancer | 126 | **0** | 126 | n/a |
| at risk for Breast/Ovarian Cancer | 38 | **0** | 38 | n/a |
| At risk for prostate cancer | 131 | **2** | 129 | n/a |
| **TOTAL** | 9195 | **202** | 8787 | 206 |

**B Non indication-based consensus returnable results (n=25,015)**




**C Non indication-based site-specific returnable results (n=17,175)**




| Disease/Disorder | P/LPVariants in site specific genes | Total |
|---|---|---|
| Cancer | *CHEK2* | 96 |
| | *JAK2* | 3 |
| Cardiac | *KCNE1 (risk allele)* | 62 |
| | *KCNE2* | 4 |
| Clotting | *F11* | 3 |
| Endocrine | *M4CR* | 12 |
| Neuromuscular | *TOR1A* | 4 |
| | *CACNA1A* | 3 |
| Neurological | *TTR* | 1 |
| | *ATM* | 36 |
| | *APT1A2* | 1 |
| Respiratory/ Hypertension | *SERPINA1*[8] | 27 |
| | *CFTR*[8] | 7 |
| | *BMPR2* | 1 |
| Dermatological | *FLG* | 2 |
| Renal | *CFH* | 1 |
| | **TOTAL** | **265** |

**Figure 6. Aggregate Findings Returned to Sites**

The positive rate for each category of returnable findings for all 25,015 participants from the eMERGE III study is shown.

(A) Indication-based returnable results. For those with an indication for testing, the different indications are depicted . [1]Four positive and two inconclusive reports had an additional secondary finding; [2]587 patients had colorectal cancer and hyperlipidemia;[3]findings from 67 consensus genes except for 2 in *CHEK2*.

(B) Non indication-based consensus returnable results. Secondary findings from the consensus gene list across the entire eMERGE III cohort are broken down per disease area. [4]14 reports had two pathogenic variants. Skewed positive rate due to one site with sample selection based on suspicious genotype (11% positive); [5]colorectal cancer (40%), breast/ovarian cancer (37%), other cancers (22%); [6]other: includes immunological/inflammatory disorders, inborn errors of metabolism, endocrine disorders, neurological disorders, clotting disorders, Myhre syndrome, and neuromuscular diseases.

(C) Non indication-based site-specific returnable results. For a subset of participants, the number of pathogenic and likely pathogenic variants in site-specific additional genes that are not on the consensus list are shown. [7]Ten participants had a site-specific variant and an additional consensus returnable variant. Of these ten site-specific variants returned, three were relevant to the indication for testing and seven were non-indication-based findings; [8]14 *SERPINA1* and 5 *CFTR* variants were reported as carrier status.

adjustment recommendations for 11 drug types can be found in Table S9.

The majority of returned data reflected variants with relatively clear interpretations for participants, with variants that either had a large body of published evidence or were straightforward to interpret. However, understanding the actual risk to patients to develop disease in those without an indication is more challenging, with risk being dependent on what is known about the penetrance of disease for the gene and variant as well as other individual factors such as family history and environmental factors (e.g., diet, exercise, exposures, etc.). In addition, in several cases, there were more interesting and unexpected findings.

The first finding involved what appeared to be a whole chromosome gain of chromosome 12. An NGS-based CNV calling algorithm detected a gain in all exons of six eMERGEseq genes on chromosome 12 (*CACNA1C* [MIM: 114205], *PKP2* [MIM: 602861], *VDR* [MIM: 601769], *MYL2* [MIM: 160781], *HNF1A* [MIM: 142410], and *POLE* [MIM: 174762]), which was confirmed by ddPCR. *CACNA1C*, and *POLE* are located near the telomeric end of the chromosome 12 p and q arms, respectively, supporting a whole chromosome gain. Given that chromosome 12 trisomies are embryonic lethal, this CNV was assumed to be either of somatic origin or occurring as a mosaic variant. The former scenario is more likely as trisomy 12 is the most common somatic chromosomal aberration in chronic lymphocytic leukemia (CLL [MIM: 151400]) (see Atlas of Genetics in Web Resources) but has also been observed in other B cell lymphoproliferative disorders and is associated with a less favorable prognosis.[18] Rarely, trisomy 12 has been reported as a mosaic variant in individuals with a variety of clinical phenotypes ranging from reportedly normal to multiple congenital anomalies, dysmorphic features, and developmental delay.[18–22] Most of these were identified prenatally, with less than ten case subjects reported postnatally and even fewer detected in peripheral blood (for reviews see Chen et al.[21] and Hong et al.[22]). Additional clinical information provided by the site indicated that this patient has a complex medical history including diabetes, heart disease, and a diagnosis of CRC at 87. While this finding is from a blood draw in early January 2016, this individual's last complete blood count in 2010 showed no evidence of increased lymphocytes or any other abnormality suggesting a CLL diagnosis. While this type of result was not anticipated within the reporting scope for eMERGE III, upon further consultation with the site, this finding was included in the clinical report of the individual to encourage additional testing and/or management.

A second case with unexpected findings was associated with another copy number variant call. A duplication for all exons of *OTC* (MIM: 300461) and *GLA* (MIM: 300644), confirmed by ddPCR, was observed in a 40-year-old male not selected for phenotype. These genes are the only two present on the X chromosome on the eMERGEseq panel. Given that *OTC* and *GLA* are on the p and q arms, respectively, the observed duplication is most likely a single event spanning the entire X chromosome. This is most consistent with a male with Klinefelter syndrome (47,XXY). Additional clinical information provided by the site confirmed a prior diagnosis of Klinefelter syndrome that had been confirmed by chromosomal karyotyping. Although a clinical report was not issued for this individual, these findings serve to further validate the sensitivity of NGS-based copy number calling.

The third unexpected category of findings was that six individuals presented with apparently mosaic variants in genes that predispose to cancer or cardiomyopathy (*TP53* [MIM: 191170], *CHEK2*, *ATM* [MIM: 607585], *MYH7* [MIM: 160760]). The presence of mosaics was based upon the ascertainment of allelic variants that were present in <30% of the DNA sequence reads at the variant site. Initial observations were screened manually to eliminate false positives due to mis-mapping to pseudogene sites or other technical errors. The presence of the mosaic variants was subsequently confirmed by Sanger sequencing and clinical reporting offered to the referring sites.

## Discussion

The introduction of clinical sequencing into the phase III of the eMERGE network has provided a framework for large-scale clinical translation of genomic data in healthcare, as well as for the seamless integration of research studies into clinical data management. The network integrated many research groups with diverse interests and a common mission to deliver genomic health care. To stimulate and address challenges for the delivery of genomic medicine, a large number of samples were tested and state of the art methods for interpretation and data delivery were applied.

A primary driver for the study design was cost and a focus on exploring the return of actionable genetic findings and therefore a gene-panel was chosen as a primary platform for genomic analyses. Whole-exome sequencing was considered. However, while exomes would have offered increased flexibility and saved time in design and testing, the network determined that a more focused target of ~100 genes was needed to stay within the budget for testing all 25,015 participants and focus on a primary goal of developing experience around return of actionable results in biobank participants. In addition, sites individually contributed research data on subjects using high density genotyping arrays allowing for genome-wide association studies which are not discussed here.

Initially, predictions were made as to the major challenges that would be faced and the most likely obstacles to achieving a smooth flow of clinical results, while maintaining access to research data. However, most of the actual challenges were not anticipated. For example, the variety of different consents used to support the process sometimes stipulated requirements inconsistent with the

network-wide decisions being made. As each site's sequencing got started, these types of site-specific challenges were uncovered. Many sites altered their decisions around the reportable content and details of their reporting needs (e.g., which genes were reportable; whether negative reports were needed; whether reports should contain certain recommendations for genetic counseling, etc.). There was evolving work around how to structure pharmacogenomic results to flow into EHRs and work to ensure the accurate provision of phenotypes from the sites to the SCs. One site needed accommodation for lower DNA input. These "hiccups" led to significant delays in getting each site started with their sequencing and clinical reports. However, once a smooth workflow was developed for each site, the SCs were able to ramp up the rate of sequencing, interpretation, and reporting. For example, during the first half of the project, 9,245 cases were completed, versus 15,770 cases completed during the second half.

The work described here supports one of the major goals of the eMERGE III project, which is to study the return of actionable genetic variants to biobank participants and assess clinical outcomes. The outcomes being tracked include the ordering of any additional tests, starting new medication, and undergoing new procedures as well as overall healthcare utilization. The protocols for returning results in eMERGE III, including consent processes and the various components involved in the return of results process such as timing, mechanism of delivery, options to receive primary versus secondary findings, and the return of positive versus neutral results, have been previously described.[1] For those sites that are returning negative results, most are doing so via letters to the participants. Both quantitative and qualitative studies, in the form of surveys and interviews, respectively, are being conducted by two sites to better understand how participants perceive such results, in particular the dissonance that may result when such results are received in the setting of a known family history of a disease (for example breast cancer). Data are currently being collected and analyzed and results will be reported separately. Furthermore, a follow-up study to explore variants of uncertain significance (VUSs), that were not reported but were in a "VUS leaning pathogenic" subcategory, is now beginning to allow phenotypes present within the EHR data to inform pathogenicity of these variants.

## Conclusions

An important outcome of the study is the generation of real data that reflects the practicality of such a large-scale biobank study. The network has provided an accurate estimate of the frequency of returnable results within the interrogated gene set. Further, the study has established the ability for two sequencing centers to adequately harmonize both the technical and interpretive aspects of clinical sequencing tests, a critical achievement to the standardization of genomic testing. Furthermore, the eMERGE network has accomplished the integration of structured genomic results directly into multiple electronic health record systems, setting the stage for the use of clinical decision support to enable genomic medicine.

## Data and Code Availability

The datasets generated and/or analyzed during the current study will be publicly available in the dbGaP repository under phs001616.v1.p1 and pre-dbGaP submission access can also be requested on the eMERGE Network website (see Web Resources).

## Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.07.018.

## Consortia Members Contribution

Manuscript Leadership: Hana Zouk,* Eric Venner,* Donna M. Muzny, Niall J. Lennon, Heidi L. Rehm,# Richard A. Gibbs.#

Test Design, Validation, Metric Tracking, CNV detection: Niall J. Lennon,* Kimberly Walker,* Adam S. Gordon, Mark Bowser, Maegan V. Harden, Theodore Chiang, Elizabeth D. Hynes, Jianhong Hu, Matthew S. Lebo, Alyssa Macbeth, Lisa Mahanta, Eric Venner, Tsung-Jung Wu, Gail P. Jarvik, Hana Zouk, Heidi L. Rehm, Richard A. Gibbs, Birgit Funke,# Donna M. Muzny.#

Validity and Actionability Assessment for Genes and SNPs: Hana Zouk, Magalie Leduc, Emily Kudalkar, Adam S. Gordon, Clinical Annotation WG, Eric Venner, Heidi L. Rehm, Gail P. Jarvik, Birgit Funke.

Data Intake and Delivery: Lawrence J. Babb, Mullai Murugan, EHRI WG, Darren C. Ames, Maegan V. Harden, Chet Graham, Lisa Mahanta, Matthew S. Lebo, Heidi L. Rehm, Richard A. Gibbs, Samuel Aronson, Eric Venner.

Pharmacogenomics Reporting: Barbara Klanderman, Hana Zouk, Eric Venner, Elizabeth D. Hynes, Chiao-Feng Lin, Chet Graham, Lisa Mahanta, Donna M. Muzny, Steven Scherer, Matthew S. Lebo, Richard A. Gibbs, Birgit Funke, Magalie Leduc.

Variant Interpretation Harmonization and Result Interpretation: Yunyun Jiang,* Leora Witkowski,* Yaping Yang, David R. Murdock, Christine M. Eng, Matthew Varugheese, Eric Venner, Clinical Annotation WG, Birgit Funke, Richard A. Gibbs, Heidi L. Rehm, Magalie Leduc,# Hana Zouk.#

Clinical Site Representatives for Return of Results: Nancy D. Leslie (CCHMC), Melanie F. Myers (CCHMC), Cynthia A. Prows (CCHMC), Wendy Chung (Columbia), David Fasel (Columbia), Maddalena Marasa (Columbia), Hila Milo Rasouly (Columbia), Chunhua Weng (Columbia), Julia Wynn (Columbia), Melissa A. Kelly (Geisinger), Marc S. Williams (Geisinger), Gail P. Jarvik (UW), Eric B. Larson (KPW), Kathleen A. Leppig (KPW), James D. Ralston (KPW), David C. Kochan (Mayo), Iftikhar J. Kullo (Mayo), Noralane M. Lindor (Mayo), Philip Lammers (Meharry), Rajbir Singh (Meharry), Duane T. Smoot (Meharry), Christin Hoell (Northwestern), Laura J. Rasmussen-Torvik (Northwestern), Maureen E. Smith (Northwestern), Robert C. Green (Partners), Jordan W. Smoller (Partners), Josh F. Peterson (VUMC), Sara L. Van Driest (VUMC), Quinn S. Wells (VUMC), Georgia L. Wiesner (VUMC).

BCM-Human Genome Sequencing Center: Donna M. Muzny, Eric Venner, Jianhong Hu, Kimberly Walker, Sara E. Kalla, Theodore Chiang, Tsung-Jung Wu, Ritika Raj, Andrea Foster, Adithya

Balasubramanian, Jesse Muniz, Shawn Denson, Gauthami Chandanavelli, Wen Liu, Harshad Mahadeshwar, Sean M. Vargas, Lan Zhang, Xiuping Liu, Qiaoyan Wang, Joy C. Jayaseelan, Darren C. Ames, Divya Kalra, Beenish Riza, Jessica De La Cruz, Liwen Wang, Viktoriya Korchina, Christie Kovar, Yunyun Jiang, Magalie Leduc, David R. Murdock, Yaping Yang, Victoria Yi, Mullai Murugan, Christine M. Eng, Richard A. Gibbs.

Partners Laboratory for Molecular Medicine/Broad Institute Sequencing Center: Hana Zouk, Maegan V. Harden, Leora Witkowski, Samuel Aronson, Lawrence J. Babb, Samantha Baxter, Mark Bowser, Wendy Brodeur, Sheila Dodge, Phil Dunlea, Christopher Friedrich, Tim DeSmet, Michael J. Dinsmore, Mark Fleharty, Chet Graham, Elizabeth D. Hynes, Emily Kudalkar, Matthew S. Lebo, Chiao-Feng Lin, Hayley Lyon, Alyssa Macbeth, Lisa Mahanta, Jim Meldrim, Thomas E. Mullen, Robert C. Onofrio, Kara Slowik, Gina Vicente, Michael W. Wilson, Betty Woolf, Stacey Gabriel, Birgit Funke[#], Niall J. Lennon[#], Heidi L. Rehm[#].

Coordinating Center: Melissa Basford, Brittany City, David R. Crosslin, Adam S. Gordon, Kayla M. Howell, Gail P. Jarvik, Jodell E. Linder, Ian B. Stanaway, Laura A. Woods, Josh F. Peterson (PI).

eMERGE Working Group Co-Chairs: Clinical Annotation: Gail P. Jarvik and Heidi Rehm; EHR Integration: Samuel Aronson and Casey Overby-Taylor; Genomics: David R. Crosslin, Megan J. Puckelwartz and Patrick M.A. Sleiman; Outcomes: Hakon Hakonarson, Josh F. Peterson and Marc S. Williams; PGx: Cynthia A. Prows and Laura Rasmussen-Torvik; Phenotyping: George Hripcsak, Wei-Qi Wei and Chunhua Weng; Return of Results/ELSI: Ingrid Holm and Iftikhar J. Kullo.

eMERGE Principal Investigators: Rex L. Chisholm (Steering Committee Chair), Samuel E. Adunyah, David R. Crosslin, Josh C. Denny, Ali Gharavi, Richard A. Gibbs, Hakon Hakonarson, John Harley, George Hripcsak, Gail P. Jarvik, Elizabeth W. Karlson, Iftikhar J. Kullo, Eric B. Larson, Niall J. Lennon, Shawn Murphy, Josh F. Peterson, Heidi L. Rehm, Dan M. Roden, Marylyn D. Ritchie, Richard R. Sharp, Maureen E. Smith, Jordan W. Smoller, Chunhua Weng, Scott T. Weiss, Marc S. Williams.

NHGRI Staff: Rongling Li, Ken L. Wiley Jr., Jyoti Dayal, Sheethal Jose, Teri Manolio.

## Additional Consortia Members

Debra Abrams, Ladia Albertson-Junkans, Berta Almoguera, Paul Appelbaum, Samuel Aronson, Sharon Aufox, Hana Bangash, Lisa Bastarache, Meckenzie Behr, Barbara Benoit, Elizabeth Bhoj, Suzette J. Bielinski, Sarah T. Bland, Carrie Blout, Erwin P. Bottinger, Harrison Brand, Murray Brilliant, Pedro Caraballo, David S. Carrell, Andrew Carroll, Lisa Castillo, Victor Castro, Kurt D. Christensen, Christopher G. Chute, Beth L. Cobb, John J. Connolly, Paul Crane, Katherine Crew, Mariza De Andrade, Ozan Dikilitas, Todd L. Edwards, Alex Fedotov, Qiping Feng, Robert Freimuth, Stephanie M. Fullerton, Vivian Gainer, M. Geoffrey Hayes, Andrew M. Glazer, Joseph T. Glessner, Jessica Goehringer, Justin H. Gundelach, Heather S. Hain, Margaret Harr, Andrea Hartzler, Scott Hebbring, Nora Henrikson, Andrew Hershey, Yoonjung Yoonie Joo, Navya Shilpa Josyula, Anne E. Justice, Brendan J. Keating, Eimear E. Kenny, Dustin Key, Krzysztof Kiryluk, Terrie Kitchner, Eric Klee, Leah Kottyan, Ming Ta (Michael) Lee, Wayne H. Liang, Todd Lingren, James G. Linneman, Cong Liu, John Lynch, Bradley Malin, Keith Marsolo, Michelle L. McGowan, Elizabeth McNally, Frank Mentch, Jonathan Mosley, Shubhabrata Mukherjee, Bahram Namjou, Yizhao Ni, Aniwaa Owusu Obeng, Thomas N. Person, Lynn Petukhova, Cassandra J. Pisieczko, Siddharth Pratap, Alanna Kulchak Rahm, Arvind Ramaprasan, Andrea Ramirez, Luke Rasmussen, Soumya Raychaudhuri, Catherine Rives, Elisabeth A. Rosenthal, Avni Santani, Dan Schaid, Stuart Scott, Aaron Scrol, Soumitra Sengupta, Ning Shang, Himanshu Sharma, Joshua C. Smith, Sunghwan Sohn, Justin Starren, Mary Stroud, Jessica Su, Kasia Tolwinski, David Veenstra, Miguel Verbitsky, Michael Wagner, Theresa Walunas, Peter S. White, Janet L. Williams, Ge Zhang.

Equal contributions indicated with symbols ([*#]).

## Declaration of Interests

Samuel Aronson is employed at Partners HealthCare, which receives royalties on sales of GeneInsight Software. Samuel Aronson's team receives funding from Sunquest and has received funding from Novartis for development of SMART on FHIR apps. Andrew Carroll is employed at Google Inc. and is a former employee of DNAnexus. Paul Crane served as a consultant to Eisai on efforts unrelated to this manuscript. David Crosslin serves on a consulting board for UnitedHealth Group with precision medicine efforts, which is unrelated to this manuscript. Christine Eng is a full-time employee/faculty member of Baylor College of Medicine. Through a professional services agreement, she serves as Chief Medical Officer and Chief Quality Officer of Baylor Genetics. Richard A. Gibbs declares that Baylor College of Medicine receives payments from Baylor Genetics Laboratories, which provides services for genetic testing; Baylor College of Medicine is part owner of Codified Genomics. Robert C. Green receives personal compensation from AIA, Applied Therapeutics, Helix, Prudential, Verily, and Veritas for speaking or consulting and is co-founder of Genome Medical, Inc. Eimear E. Kenny has received speaker honorariums from Illumina and Regeneron Pharmaceuticals. Niall J. Lennon is an advisor to Genturi Inc. Elizabeth McNally serves or has served as a consultant to Invitae, Tenaya, Exonics, Pfizer, AstraZeneca, Cytokinetics, and 4D Molecular Therapeutics and founded Ikaika Therapeutics. Thomas E. Mullen is employed at and a shareholder at Quest Diagnostics. Heidi Rehm is employed at Massachusetts General Hospital, which receives royalties on sales of GeneInsight Software. Avni Santani receives royalties from Agilent Technologies and a founder of Opus Genomics. Jordan W. Smoller is an unpaid member of the Bipolar/Depression Research Community Advisory Panel of 23andMe. Eric Venner is a cofounder of Codified Genomics. Theresa Walunas completed (in 2018) legal consulting for by Pfizer Inc., Wyeth LLC, Genetics Institute, LLC, Merck KGaA, and EMD Serono, Inc. Georgia L. Wiesner is a member of the External Advisory Panel for the ClinGen Clinical Genome Resource Project. Hana Zouk is employed at Massachusetts General Hospital which receives royalties on sales of GeneInsight Software.

## Acknowledgments

The eMERGE Phase III Network was initiated and funded by the National Human Genome Research Institute (NHGRI) through the following grants: U01HG8657 (Kaiser Permanente Washington Health Research Institute/University of Washington), U01HG8685 (Brigham and Women's Hospital), U01HG8672 (Vanderbilt University Medical Center), U01HG8666 (Cincinnati Children's Hospital Medical Center), U01HG6379 (Mayo Clinic), U01HG8679 (Geisinger Clinic), U01HG8680 (Columbia University Health Sciences), U01HG8684 (Children's Hospital of Philadelphia), U01HG8673 (Northwestern University), MD007593 (Meharry Medical College),

U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center), U01HG8676 (Partners Healthcare/Broad Institute), and U01HG8664 (Baylor College of Medicine).

Received: October 26, 2018
Accepted: July 26, 2019
Published: August 22, 2019

## Web Resources

Atlas of Genetics and Cytogenetics in Oncology and Haematology, Michaux, L. (2000). +12 or trisomy 12, http://atlasgeneticsoncology.org/Anomalies/tri12ID2024.html
Clinical Research Sequencing Platform, https://portals.broadinstitute.org/portal/CRSP
CPIC Publications, https://cpicpgx.org/publications
dbGaP, https://www.ncbi.nlm.nih.gov/gap
eMERGE, https://www.genome.gov/27540473/electronic-medical-records-and-genomics-emerge-network/
eMERGE Network, https://emerge.mc.vanderbilt.edu/collaborate/
eMERGE Sample Submission Portal, https://emerge.hgsc.bcm.edu/workflow/sample-submission
Gene Dosage Curations, https://search.clinicalgenome.org/kb/gene-dosage
OMIM, https://www.omim.org/
Sequence Variant Interpretation, https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation

## References

1. Fossey, R., Kochan, D., Winkler, E., Pacyna, J.E., Olson, J., Thibodeau, S., Connolly, J.J., Harr, M., Behr, M.A., Prows, C.A., et al. (2018). Ethical Considerations Related to Return of Results from Genomic Medicine Projects: The eMERGE Network (Phase III) Experience. J. Pers. Med. *8*, 2.
2. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al.; American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.
3. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
4. Reid, J.G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C., et al. (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics *15*, 30.
5. Pugh, T.J., Amr, S.S., Bowser, M.J., Gowrisankar, S., Hynes, E., Mahanta, L.M., Rehm, H.L., Funke, B., and Lebo, M.S. (2016). VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. Genet. Med. *18*, 712–719.
6. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
7. Fromer, M., and Purcell, S.M. (2014). Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr. Protoc. Hum. Genet. *81*, 1–21.
8. Chiang, T., Liu, X., Wu, T.-J., Hu, J., Sedlazeck, F.J., White, S., Schaid, D., de Andrade, M., Jarvik, G.P., Crosslin, D., et al. (2018). Atlas-CNV: a validated approach to call Single-Exon CNVs in the eMERGESeq gene panel. Genet. Med. Published online March 20, 2019. https://doi.org/10.1038/s41436-019-0475-4.
9. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
10. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am. J. Hum. Genet. *100*, 895–906.
11. Aronson, S.J., Clark, E.H., Babb, L.J., Baxter, S., Farwell, L.M., Funke, B.H., Hernandez, A.L., Joshi, V.A., Lyon, E., Parthum, A.R., et al. (2011). The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. Hum. Mutat. *32*, 532–536.
12. Aronson, S.J., Clark, E.H., Varugheese, M., Baxter, S., Babb, L.J., and Rehm, H.L. (2012). Communicating new knowledge on previously reported genetic variants. Genet. Med. *14*, 713–719.
13. Aronson, S., Babb, L., Ames, D., Gibbs, R.A., Venner, E., Connelly, J.J., Marsolo, K., Weng, C., Williams, M.S., Hartzler, A.L., et al.; eMERGE Network EHRI Working Group (2018). Empowering genomic medicine by establishing critical sequencing result data flows: the eMERGE example. J. Am. Med. Inform. Assoc. *25*, 1375–1381.
14. Abbott, G.W., Sesti, F., Splawski, I., Buck, M.E., Lehmann, M.H., Timothy, K.W., Keating, M.T., and Goldstein, S.A. (1999). MiRP1 forms IKr potassium channels with HERG and is associated with cardiac arrhythmia. Cell *97*, 175–187.
15. Kapplinger, J.D., Tester, D.J., Salisbury, B.A., Carr, J.L., Harris-Kerr, C., Pollevick, G.D., Wilde, A.A.M., and Ackerman, M.J. (2009). Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. Heart Rhythm *6*, 1297–1303.
16. Nawathe, P.A., Kryukova, Y., Oren, R.V., Milanesi, R., Clancy, C.E., Lu, J.T., Moss, A.J., Difrancesco, D., and Robinson, R.B. (2013). An LQTS6 MiRP1 mutation suppresses pacemaker current and is associated with sinus bradycardia. J. Cardiovasc. Electrophysiol. *24*, 1021–1027.
17. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray, M.F., Smelser, D.T., Gerhard, G.S., and Ledbetter, D.H. (2016). The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. *18*, 906–913.
18. English, C.J., Goodship, J.A., Jackson, A., Lowry, M., and Wolstenholme, J. (1994). Trisomy 12 mosaicism in a 7 year old girl

with dysmorphic features and normal mental development. J. Med. Genet. *31*, 253–254.

19. Hsu, L.Y., Yu, M.T., Neu, R.L., Van Dyke, D.L., Benn, P.A., Bradshaw, C.L., Shaffer, L.G., Higgins, R.R., Khodr, G.S., Morton, C.C., et al. (1997). Rare trisomy mosaicism diagnosed in amniocytes, involving an autosome other than chromosomes 13, 18, 20, and 21: karyotype/phenotype correlations. Prenat. Diagn. *17*, 201–242.
20. DeLozier-Blanchet, C.D., Roeder, E., Denis-Arrue, R., Blouin, J.L., Low, J., Fisher, J., Scharnhorst, D., and Curry, C.J. (2000). Trisomy 12 mosaicism confirmed in multiple organs from a liveborn child. Am. J. Med. Genet. *95*, 444–449.
21. Chen, C.-P., Chang, S.-D., Su, J.-W., Chen, Y.-T., and Wang, W. (2013). Prenatal diagnosis of mosaic trisomy 12 associated with congenital overgrowth. Taiwan. J. Obstet. Gynecol. *52*, 454–456.
22. Hong, B., Zunich, J., Openshaw, A., and Toydemir, R.M. (2017). Clinical features of trisomy 12 mosaicism-Report and review. Am. J. Med. Genet. A. *173*, 1681–1686.

**The American Journal of Human Genetics, Volume *105***

## Supplemental Data

# Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network

**The eMERGE Consortium**

Figure S1.

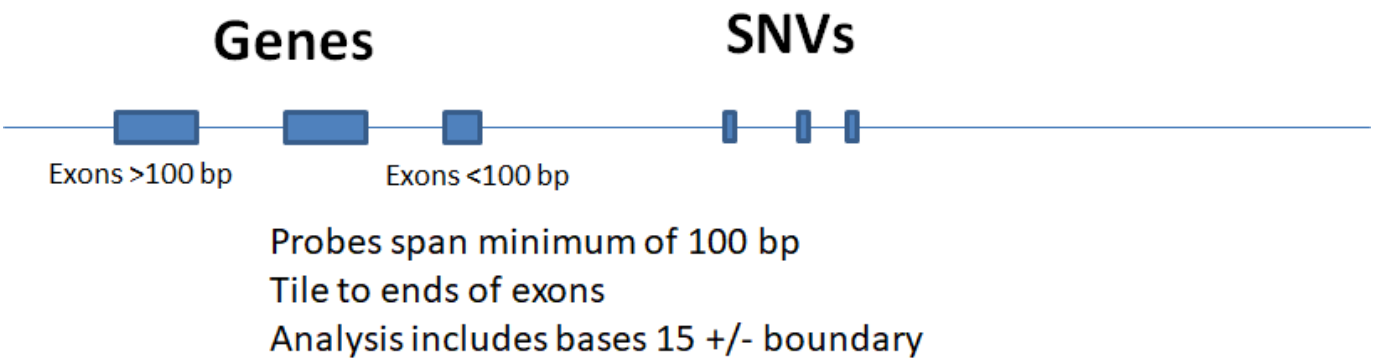



Figure S1. Capture design strategies. Probes were designed to capture targeted regions of the eMERGESeq panel according to the certain criteria that include but are not limited to those shown in the figure.

Figure S2.

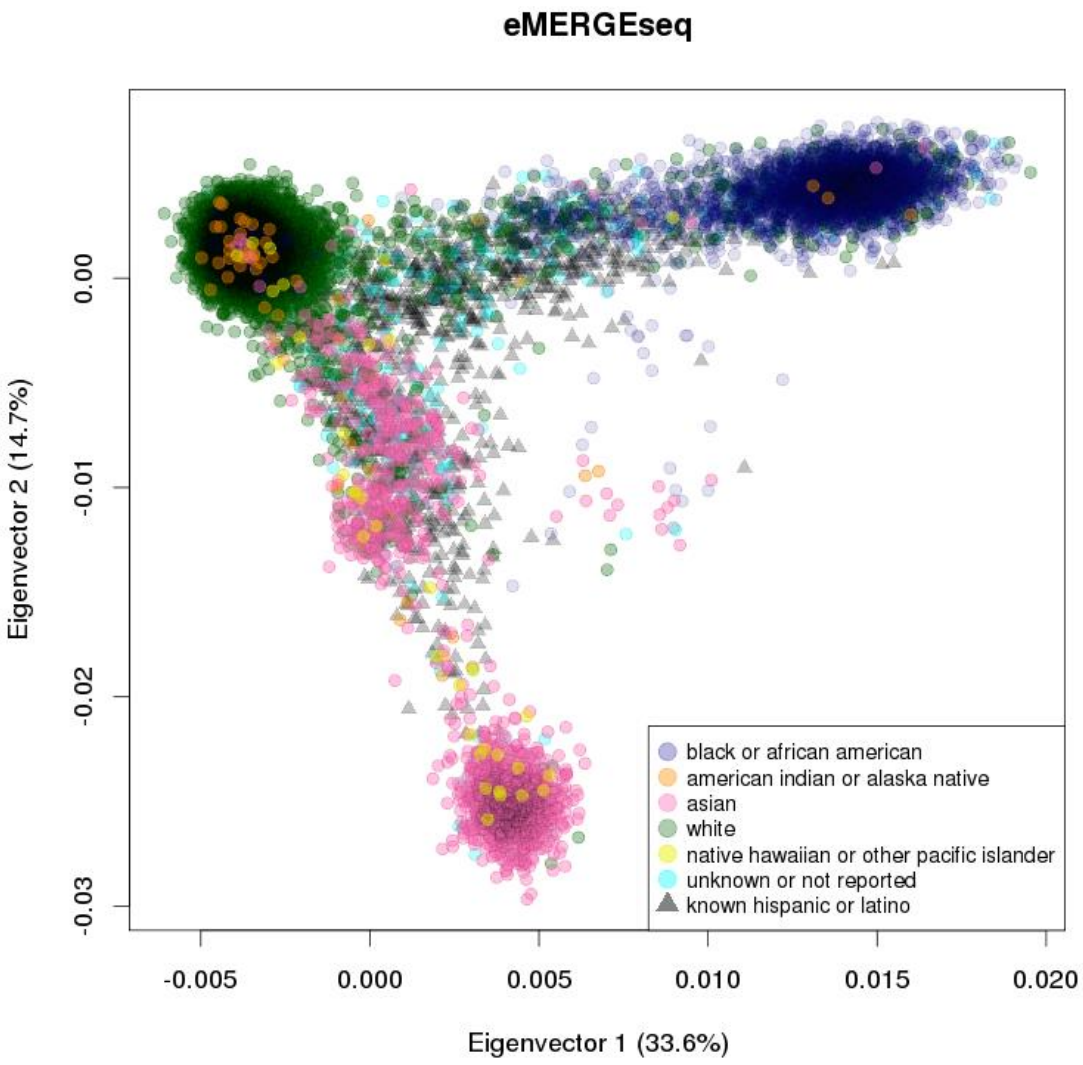



Figure S2. PCA plot of genetic ancestry for the eMERGEseq cohort.

Table S2. Capture design performance comparison between both sequencing centers before and after design optimization (Version 1 vs. Validated Panel).

| | **Broad/LMM** | | **Baylor** | |
|---|---|---|---|---|
| | Version 1 Panel | Validated Panel (v2) | Version 1 Panel | Validated Panel (v2) |
| % of Bases ≥20X | 99.30% | 99.80% | 99.80% | 99.90% |
| Number of bases <20X in ≥10% of samples | 3,364 | 732 | 1,332 | 475 |
| Gene_Exon with bases <20X in validated panel (% of bases in that exon <20X) | n/a | *CACNA1A*_Exon42 (100%)<br>*TGFBR1*_Exon01 (100%)<br>*KCNQ1*_Exon01 (44.5%)<br>*RYR1*_Exon91 (22.4%)<br>*CACNA1B*_Exon01 (19.4%)<br>*CACNA1B*_Exon19 (7.6%)<br>*RB1*_Exon09 (2.6%) | n/a | *SDHD*_Exon04 (100%)<br>*CHEK2*_Exon14 (100%)<br>*TGFBR1*_Exon01 (39.8%)<br>*PKP2*_Exon09 (14.3%)<br>*COL5A1*_Exon01 (11.8%)<br>*RYR1*_Exon91 (10.6%)<br>*APOB*_Exon29 (8.4%)<br>*CACNA1B*_Exon19 (6.4%) |
| SNV with <20X | n/a | rs25531<br>rs25532<br>rs1135840<br>rs1702294 | n/a | rs657452<br>rs6002655<br>rs8066602<br>rs8066731<br>rs35445101 |
| SNVs determined by eMerge Annotation group to be clinically actionable (14) that have low or no coverage | n/a | None | n/a | |
| ClinVar Pathogenic/Likely pathogenic variants with low/missing coverage | n/a | rs199473441[a]<br>rs397508096[c]<br>rs794728563[a]<br>rs794728544[a]<br>rs199472884[a,b] | n/a | None |

[a]single submitter for 'pathogenic' classification; [b] multiple submitters for 'likely pathogenic'; [c] multiple submitters for 'pathogenic'

Table S2. Comparison before and after design optimization (Version 1 vs. Validated Panel). While each design has unique strengths and deficiencies, the percent bases ≥ 20X coverage is greater than 99% for both sequencing sites.

Table S3. Summary of PGx report contents from sequencing centers to sites

| Genes | # Variants Genotyped | Haplotypes or *Alleles Reported | Phenotypes Reported | Associated Drugs | Reporting differences between SCs | CPIC Publications |
|---|---|---|---|---|---|---|
| *TPMT* | 4 | *1, *2, *3A, *3B, *3C, *4 | TPMT Normal, Intermediate, or Poor Metabolizer | Azathioprine, Mercaptopurine, Thioguanine (thiopurines, immunosuppressant) | None | Relling et al. 2011 PMID 21270794; Relling et al. 2013 PMID 23422873 |
| *CYP2C9* | 2 | *1, *2, *3 | CYP2C9 Normal, Intermediate, or Poor Metabolizer | Warfarin (anticoagulant), Phenytoin/fosphenytoin (anticonvulsant) | BCM-HGS reported CYP2C9 and VKORC1 together for warfarin only given that the phenytoin dosage guidelines rely on the HLA-B allele, which is not assessed in this assay. | Johnson et al. 2011 PMID: 21900891, Johnson et al. 2017 PMID: 28198005; Caudle et al. 2014 PMID: 25099164 |
| *VKORC1* | 1 | c.-1639G>A (rs9923231) | VKORC1 Normal, Intermediate or Low Expression | Warfarin (anticoagulant) | None | Johnson et al. 2011 PMID: 21900891, Johnson et al. 2017 PMID: 28198005 |
| *IFNL3* | 1 | c.-3180G>A (rs12979860) | IFNL3 Favorable or Unfavorable | PEG-IFN-α, Ribavirin (antiviral) | None | Muir et al. 2014 PMID: 24096968 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Response | | | |
| *DPYD* | 3 | *1, *2A, *13, c.2846A>T | DPYD Normal, Intermediate, or Poor Metabolizer | 5-Fluorouracil, Capecitabine, Tegafur[a] (fluoropyrimidines, antineoplastic) | None | Caudle et al. 2013 PMID: 23988873 |
| *SLCO1B1* | 1 | *1A, *5 (c.521T>C; rs4149056) | SLCO1B1 Normal, Decreased, or Poor Function | Simvastatin (cholesterol reduction) | None | Wilke et al. 2012 PMID: 22617227, Ramsey et al. 2014 PMID: 24918167 |
| *CYP2C19* | 8 | *1, *2, *3, *4A, *4B, *5, *6, *7, *8, *17 | CYP2C19 Ultrarapid, Rapid, Normal, Intermediate, or Poor Metabolizer | Clopidogrel (antiplatelet); Voriconazole (antifungal); Amitriptyline, Clomipramine, Doxepin, Imipramine, Trimipramine (tricyclic antidepressants or TCAs); Citalopram, Escitalopram, Sertraline (selective serotonin reuptake inhibitors or SSRIs) | Clomipramine, Doxepin, Imipramine, Trimipramine, Sertraline are CPIC level B; BCM-HGSC only reported CPIC level A gene/drug combinations. | Scott et al. 2011 PMID: 21716271, Scott et al. 2013 PMID: 23698643; Moriyama et al. 2016 PMID: 27981572; Hicks et al. 2013 PMID: 23486447, Hicks et al. 2016 PMID: 27997040; Hicks et al. 2015 PMID: 25974703 |

[a]Tegafur/DPYD had a CPIC level A when PGx report was implemented. It was since lowered to a level C

Table S4. The 39 preferred indication terms and codes for the eMERGE III network.

| **Preferred indication term** | **Code** |
|---|---|
| Abnormal sex determination | EMERGE-GIS-LOCAL 10102-6 (abnormal sex determination) |
| Abnormality of pain sensation | EMERGE-GIS-LOCAL 10103-5 (abnormality of pain sensation) |
| Abnormality of the heart valves | EMERGE-GIS-LOCAL 10104-4 (abnormality of the heart valves) |
| Adult Migraine | EMERGE-GIS-LOCAL 10099-1 (adult migraine) |
| Amyloidosis, Hereditary, Transthyretin-Related | MIM 105210 (transthyretin amyloidosis) |
| Arrhythmia | MESH D001145 (arrhythmias, cardiac) |
| Ascending aortic dilation / Aneurysm | EMERGE-GIS-LOCAL 10106-2 (ascending aortic dilation / aneurysm) |
| Asthma | DOID 2841 (asthma) |
| Atopic dermatitis | DOID 3310 (atopic dermatitis) |
| Autistic behavior | EMERGE-GIS-LOCAL 10109-9 (autistic behavior) |
| Autoimmunity | EMERGE-GIS-LOCAL 10110-7 (autoimmunity) |
| Bipolar affective disorder | DOID 3312 (bipolar disorder) |
| Breast carcinoma | DOID 3459 (breast carcinoma) |
| Cardiomyopathy | ORPHA 167848 (cardiomyopathy) |
| Chronic kidney disease | MESH D051436 (renal insufficiency, chronic) |
| Chronic sinusitis | EMERGE-GIS-LOCAL 10111-6 (chronic sinusitis) |
| Cirrhosis | DOID 5082 (liver cirrhosis) |
| Colorectal Cancer / Polyps | EMERGE-GIS-LOCAL 10126-0 (colorectal cancer / polyps) |
| Congestive heart failure | DOID 6000 (congestive heart failure) |
| Coronary artery disease | DOID 3393 (coronary artery disease) |
| Depression | EMERGE-GIS-LOCAL 10128-8 (depression) |
| Ehlers-Danlos Syndrome | ORPHA 98249 (ehlers-danlos syndrome) |
| Familial hypercholesterolemia | EMERGE-GIS-LOCAL 10134-1 (familial hypercholesterolemia)<br>MIM 143890 (familial hypercholesterolemia) |
| Healthy | EMERGE-GIS-LOCAL 10094-6 (healthy) |
| Hyperammonemia due to ornithine transcarbamylase deficiency | MIM 311250 (ornithine carbamoyl transferase deficiency) |
| Hyperlipidemia | MESH D006949 (hyperlipidemias) |
| Hypertriglyceridemia | MESH D015228 (hypertriglyceridemia) |
| Intellectual disability | DOID 1059 (intellectual disability) |
| Not selected for trait | EMERGE-GIS-LOCAL 10093-7 (not selected for trait) |
| Obesity | DOID 9970 (obesity) |
| Opioid dependence, Neonatal abstinence | EMERGE-GIS-LOCAL 10100-8 (opioid dependence, neonatal abstinence) |
| Ovarian Cancer, epithelial included | ORPHA 213500 (ovarian cancer) |
| Pediatric Migraine | EMERGE-GIS-LOCAL 10101-7 (pediatric migraine) |
| Pulmonary Hypertension | ORPHA 422 (primary pulmonary hypertension) |
| Rheumatoid arthritis | DOID 7148 (rheumatoid arthritis) |
| Schizophrenia | DOID 5419 (schizophrenia) |
| Seizures | EMERGE-GIS-LOCAL 10118-9 (seizures) |
| Stroke | MESH D020521 (stroke) |
| Tuberous sclerosis type 1 | EMERGE-GIS-LOCAL 10135-0 (tuberous sclerosis type 1)<br>MIM 191100 (tuberous sclerosis type 1) |

Table S5. eMERGE III Clinical cohort description

| Site | Total participants | eMERGESeq Cohort summary | Any phenotype enrichment? |
|---|---|---|---|
| Vanderbilt | 2,452 | biobank - prior PGx testing or interest in research | N |
| UW/KPW | 2,500 | biobank – Colorectal cancer/Polyps diagnosis or Asian ancestry | Y |
| Columbia | 2,582 | biobank & prospective, some specific clinics and studies | Y |
| Mayo | 3,025 | CRC/P & Hyperlipidemia cohorts | Y |
| Northwestern | 2,985 | prospective recruitment across clinics, some specialty | Y |
| Geisinger | 2,500 | biobank - suspicious genotype | Y |
| Harvard | 2,500 | biobank - unselected | N |
| CCHMC | 3,000 | biobank & adolescent prospective | N |
| CHOP | 2,976 | biobank - enriched for neuro phenotypes | Y |
| Meharry | 495 | Breast, prostate, colorectal, lung cancer or high risk for developing these cancers | Y |

Table S6. Demographic information for participants for each site in the eMERGEIII network

| Site | Sex | | Total | Caucasian | African American | Asian | Hispanic | Native American | Other | Unknown |
|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | | | | | | | | |
| UW/KPW | 984 | 1516 | 2500 | 1296 | 55 | 1008 | 40 | 50 | 27 | 24 |
| Geisinger | 809 | 1691 | 2500 | 2350 | 68 | 9 | 59 | 4 | 7 | 3 |
| CCHMC[a] | 1504 | 1496 | 3000 | 1773 | 1133 | 26 | 42 | 3 | 3 | 20 |
| Harvard | 1083 | 1417 | 2497 | 2043 | 147 | 75 | 137 | 0 | 0 | 98 |
| Northwestern | 1148 | 1837 | 2985 | 2257 | 405 | 132 | 167 | 4 | 3 | 17 |
| Mayo[b] | 1066 | 1497 | 3025 | 2391 | 14 | 19 | 114 | 2 | 0 | 485 |
| CHOP | 2045 | 931 | 2976 | 1533 | 1164 | 38 | 122 | 3 | 3 | 113 |
| Columbia | 1163 | 1419 | 2582 | 829 | 205 | 109 | 358 | 4 | 22 | 1055 |
| Vanderbilt[c] | 1257 | 1105 | 2452 | 2156 | 102 | 24 | 26 | 4 | 0 | 140 |
| Meharry | 275 | 220 | 495 | 0 | 495 | 0 | 0 | 0 | 0 | 0 |
| Total | 10352 | 11890 | 22242 | 15845 | 2573 | 1377 | 967 | 53 | 20 | 1407 |

[a]1 individual had no sex information; [b]457 individuals had no sex information; [c]90 individuals had no sex information

Table S9. Participants with actionable PGx results from the eMERGE III cohort

| Drug | Gene | Recommended adjustment to standard dosing or alternate drug use based on genotype | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UW/KPW (n=2500) | CCHMC (n=3000) | Geisinger (n=2500) | Harvard (n=2500) | Vanderbilt (n=2452) | Columbia (n=2582) | Mayo (n=3025) | North-western (n=2001) | CHOP (n=2976) | Meharry (n=495) |
| **Thiopurines** | *TPMT* | 177 **(7%)** | 296 **(10%)** | 223 **(9%)** | 186 **(7%)** | 241 **(10%)** | 223 **(9%)** | 308 **(10%)** | 254 **(9%)** | 248 **(8%)** | 52 **(11%)** |
| **Warfarin[a]** | *CYP2C9/ VKORC1* | 1189 **(48%)** | 740 **(25%)** | 855 **(34%)** | 944 **(38%)** | 747 **(30%)** | 856 **(33%)** | 982 **(32%)** | 962 **(32%)** | 673 **(23%)** | 14 **(3%)** |
| **Phenytoin/ fosphenytoin** | | 579 **(23%)** | 730 **(24%)** | 858 **(34%)** | 844 **(34%)** | 849 **(35%)** | 716 **(28%)** | 993 **(33%)** | 912 **(31%)** | 724 **(24%)** | 40 **(8%)** |
| **PEG-IFN-α, Ribavirin** | *IFNL3* | 991 **(40%)** | 1940 **(65%)** | 1412 **(56%)** | 1348 **(54%)** | 0[b] (n/a) | 1512 **(59%)** | 1651 **(55%)** | 1692 **(57%)** | 1952 **(66%)** | 0[b] (n/a) |
| **Fluoropyrimidines** | *DPYD* | 35 **(1%)** | 55 **(2%)** | 57 **(2%)** | 52 **(2%)** | 0[b] (n/a) | 31 **(1%)** | 51 **(2%)** | 46 **(2%)** | 52 **(2%)** | 0[b] (n/a) |
| **Simvastatin** | *SLCO1B1* | 589 **(24%)** | 630 **(21%)** | 706 **(28%)** | 735 **(29%)** | 653 **(27%)** | 667 **(26%)** | 876 **(29%)** | 793 **(27%)** | 597 **(20%)** | 30 **(6%)** |
| **Clopidogrel** | *CYP2C19* | 1012 **(41%)** | 911 **(30%)** | 702 **(28%)** | 741 **(30%)** | 699 **(29%)** | 792 **(31%)** | 890 **(29%)** | 898 **(30%)** | 881 **(30%)** | 172 **(35%)** |
| **Voriconazole** | | 661 **(26%)** | 1018 **(34%)** | 873 **(35%)** | 850 **(34%)** | 856 **(35%)** | 781 **(30%)** | 970 **(32%)** | 1002 **(34%)** | 1030 **(35%)** | 154 **(31%)** |
| **Tricyclic antidepressants** | | 661 **(26%)** | 1018 **(34%)** | 873 **(35%)** | 850 **(34%)** | 856 **(35%)** | 781 **(30%)** | 970 **(32%)** | 1002 **(34%)** | 1030 **(35%)** | 154 **(31%)** |
| **Citalopram, Escitalopram** | | 661 **(26%)** | 1018 **(34%)** | 873 **(35%)** | 850 **(34%)** | 856 **(35%)** | 781 **(30%)** | 970 **(32%)** | 1002 **(34%)** | 1030 **(35%)** | 154 **(31%)** |
| **Sertraline** | | 661 **(26%)** | 1018 **(34%)** | 873 **(35%)** | 850 **(34%)** | 856 **(35%)** | 781 **(30%)** | 970 **(32%)** | 1002 **(34%)** | 1030 **(35%)** | 154 **(31%)** |

[a]Warfarin dosing algorithms use both genetic and nongenetic factors such as age, sex, smoking status etc. to predict appropriate dose. Follow up with physician is recommended for all individuals

[b]this site elected to not receive this PGx result

**Supplemental methods:**

**Proficiency Testing across clinical sequencing sites.**

Both BCM and Broad clinical labs are accredited by the College of American Pathologists (CAP) and are therefore required to perform biannual proficiency testing (PT) on every clinical test offered. There are several acceptable means to perform PT on a sequencing-based assay including enrollment in CAP's PT program. As part of the PT program CAP sends out reference samples with known events for the clinical lab to prepare and sequence and results are submitted to CAP for scoring. The eMERGEseq panel presented a unique opportunity for an alternate PT program - lab exchange. Since both laboratories are running the same clinical test, both could perform PT by sending previously tested clinical samples to the other lab. Results from end-to-end processing were compared for concordance. This approach has the added benefit of ensuring both laboratories remain technically harmonized throughout the duration of the project. For this program, BCM performs both the CAP PT program in conjunction with the alternate PT program described above.

Results of mid-2017 PT sample from BCM run at Broad: passing variant calls 100% concordant with BCM variants.

Results of mid-2017 PT sample from Broad run at BCM: passing variant calls 100% concordant with Broad variants.

**HGSC Clinical Laboratory**
One Baylor Plaza, Houston TX, 77030
Phone: 713.798.6539 Fax: 713.798.5741
www.hgsc.bcm.edu questions@hgsc.bcm.edu

**Preliminary Report**

goo.gl/cLdKB8

| | | | |
|---|---|---|---|
| Patient Name: | | Sample Collected Date: | |
| Patient ID: | | Sample Received Date: | 03/16/2016 |
| Age: | | Report Date: | |
| DOB: | | Sample Type: | |
| Sex: | | Indication for Testing: | |
| Patient Sample ID: | | Ordering Physician Name: | |
| Accession #: | | | |

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

# Preliminary eMERGE-Seq Panel Sequencing Report

## HGSC-CL

This test interrogates the protein-coding and exon-splicing regions of 109 genes as well as 1551 single-nucleotide polymorphisms that may impact human health and disease. Clinical interpretation and reporting are provided for pathogenic and likely pathogenic variants for genes and single nucleotide polymorphisms as described in the methodology section.

### PATHOGENIC AND/OR LIKELY PATHOGENIC VARIANTS DETECTED

A homozygous c.350G>A (p.R117H) pathogenic variant in the CFTR (NM_000492.3) gene was detected in this individual. Defects in CFTR are the cause of cystic fibrosis (CF) [MIM 219700], an autosomal recessive common generalized disorder of the exocrine glands which impairs clearance of secretions in a variety of organs. It is characterized by the triad of chronic bronchopulmonary disease (with recurrent respiratory infections), pancreatic insufficiency (which leads to malabsorption and growth retardation), and elevated sweat electrolytes. Defects in CFTR are also the cause of congenital bilateral absence of the vas deferens (CBAVD) [MIM 277180], an important cause of sterility in men and could represent an incomplete form of cystic fibrosis, as the majority of men suffering from cystic fibrosis lack the vas deferens.

**Table 1: Details of Pathogenic and Likely Pathogenic Variants**

| Disease | Inh. | Gene | Position (NCBI 37) | Variant | Zyg. | Notes | Interpretation |
|---|---|---|---|---|---|---|---|
| Cystic fibrosis [MIM 219700]; Congenital bilateral absence of the vas deferens [MIM 277180] | AR | CFTR | chr7 g.117171029G>A | c.350G>A p.R117H | Homozygous | PMID 2344617, 23420618, 21228398, 21594800, 10103316, 24440181, 21507732, 22366207, 7506096, 22975760, 23951356, 12767731, 22658665, 20619026, 26324139, 23974870, 19880712, 23891399, 15246977, 26846474, 22332135, 20021716, 21520337, 23378603, 20797923, 20923678, 18778819, 19885835, 22572128, 23751316; rs78655421; [5T ] | Pathogenic |

**Table 2: Details of Copy Number Variants:**

No CNVs found for this sample.

**Table 3: Details of Pharmacogenomic Variants**

Pharmacogenomics variants are returned for the following genes: CYP2C19, DPYD, INFL3, SLCO1B1, TPMT, CYP2C9/VKORC1. Star alleles are determined based on the variants detected by this assay. Star alleles may not be accurately defined due to the limitations of this assay which include: 1) The presence of additional variants defining functional and non functional alleles in a patient, not detected by this assay, and 2) the lack of ability to determine the phase of the variants when a star allele is defined by multiple variants. Additionally, undetected genetic and/or non genetic factors such as drug-drug interactions, may also impact the phenotype. This pharmacogenomic report is limited to CPIC level A alleles and drug recommendations. Additional (level B and lower) drugs may be metabolized by these

reported enzymes; and additional enzymes, not reported here, may affect the metabolism of a reported drug. Refer to the current recommendation for dosage guidelines. See Methodology for details.

| Gene | Drug | Diplotype | Phenotype | Recommendation |
|---|---|---|---|---|
| CYP2C19 | clopidogrel | *1/*1 | Normal metabolizer | https://cpicpgx.org/guidelines/guideline-for-clopidogrel-and-cyp2c19/ |
| | voriconazole | | | https://cpicpgx.org/guidelines/guideline-for-voriconazole-and-cyp2c19/ |
| | citalopram, escitalopram | | | https://cpicpgx.org/guidelines/guideline-for-selective-serotonin-reuptake-inhibitors-and-cyp2d6-and-cyp2c19/ |
| | amitriptyline | | | https://cpicpgx.org/guidelines/guideline-for-tricyclic-antidepressants-and-cyp2d6-and-cyp2c19/ |
| DPYD | capecitabine | *1/*1 | Normal DPD activity and "normal" risk for fluoropyrimidine toxicity | https://cpicpgx.org/guidelines/guideline-for-fluoropyrimidines-and-dpyd/ |
| | fluorouracil | | | |
| | tegafur | | | |
| IFNL3 | peginterferon alfa-2a | rs12979860 C/C | Favorable response genotype | https://cpicpgx.org/guidelines/guideline-for-peg-interferon-alpha-based-regimens-and-ifnl3/ |
| | peginterferon alfa-2b | | | |
| | ribavirin | | | |
| SLCO1B1 | simvastatin | rs4149056 T/C | Intermediate function, Intermediate simvastatin induced myopathy risk | https://cpicpgx.org/guidelines/guideline-for-simvastatin-and-slco1b1/ |
| TPMT | azathioprine | *1/*1 | High activity | https://cpicpgx.org/guidelines/guideline-for-thiopurines-and-tpmt/ |
| | mercaptopurine | | | |
| | thioguanine | | | |
| CYP2C9 | warfarin | *1/*3 | Intermediate metabolizer | https://cpicpgx.org/guidelines/guideline-for-warfarin-and-cyp2c9-and-vkorc1/ |
| VKORC1 | | T/T | | |

**Interpretation of Pharmacogenomic Variants:**

This individual is homozygous for the wild type allele of the CYP2C19 gene. Based on the genotype result, this patient is predicted to have a CYP2C19 normal metabolizer phenotype. This genotype information can be used by patients and clinicians as part of the shared decision-making process for several drugs metabolized by CYP2C19 including clopidogrel, voriconazole, amitriptyline, citalopram and escitalopram. For clopidogrel, individuals with this diplotype are expected to have normal platelet inhibition and normal residual platelet aggregation in response to clopidogrel. Label recommended dosage and administration are recommended. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-clopidogrel-and-cyp2c19/. For voriconazole, normal voriconazole metabolism is expected in individuals with this genotype. Initiate therapy with recommended standard of care dosing. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-voriconazole-and-cyp2c19/. For citalopram and escitalopram, initiate therapy with recommended starting dose. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-selective-serotonin-reuptake-inhibitors-and-cyp2d6-and-cyp2c19/. For amitriptyline, initiate therapy with recommended starting dose. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-tricyclic-antidepressants-and-cyp2d6-and-cyp2c19/. For citalopram, escitalopram and amitriptyline, if CYP2D6 genotyping is available, refer to the current guidelines for dosing recommendations.

This individual is homozygous for the functional allele of the DPYD gene. This genotype information can be used by patients and clinicians as part of the shared decision-making process for fluoropyrimidines (capecitabine, fluorouracil, tegafur). Based on the genotype result, this patient is predicted to have a normal DPD activity phenotype. Individuals with this diplotype are expected to have "normal" risk for fluoropyrimidine toxicity. Recommendations include the use of label recommended dosage and administration. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-fluoropyrimidines-and-dpyd/.

This individual is homozygous for the rs12979860 C/C allele in the IFNL3 gene. This variant is the strongest baseline predictor of response to peginterferon alfa and ribavirin therapy in previously untreated patients and can be used by patients and clinicians as part of the shared decision-making process for initiating treatment for hepatitis C virus infection. Based on the genotype result, this patient is predicted to have an increased likelihood of response (higher sustained virologic response rate) to peginterferon alfa and ribavirin therapy as compared with patients with unfavorable response genotype. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-peg-interferon-alpha-based-regimens-and-ifnl3/.

This individual is heterozygous for the rs4149056 T/C allele in the SLCO1B1 gene. This genotype information can be used by patients and clinicians as part of the shared decision-making process for simvastatin and other drugs affected by SLCO1B1. Based on the genotype result, this patient is predicted to have intermediate SLCO1B1 function. This patient may be at risk for an adverse response to medications that are affected by SLCO1B1. To avoid an untoward drug response, dose adjustments may be necessary for medications affected by SLCO1B1. If simvastatin is prescribed to a patient with intermediate SLCO1B1 function, there is an increased risk for developing simvastatin-associated myopathy; such patients may need a lower starting dose of simvastatin or an alternate statin agent. Refer to current guidelines for dosage and recommendations at https://cpicpgx.org/guidelines/guideline-for-simvastatin-and-slco1b1/.

This individual is homozygous for the normal high activity allele of the TPMT gene. Decreased TPMT gene activity is associated with toxicity and myelosuppression in response to thiopurines, and this genotype information can be used by patients and clinicians as part of the shared decision-making process for initiating treatment. Based on the genotype result, this patient is predicted to have normal TPMT function. Individuals with this diplotype are expected to have a normal response to mercaptopurine, azathioprine and thioguanine. A normal dose of thiopurine and adjustment following the disease-specific guidelines is recommended. Refer to current guidelines for dosage and recommendations for each specific thiopurine drug at https://cpicpgx.org/guidelines/guideline-for-thiopurines-and-tpmt/.

This individual is heterozygous for the low function allele in the CYP2C9 gene. Based on the genotype result, this patient is predicted to have intermediate CYP2C9 function.This individual is also homozygous for the variant allele for the VKORC1 gene. Expression level of the VKORC1 gene is associated with warfarin sensitivity. Based on the genotype result, this patient is predicted to have high sensitivity to warfarin.

**Comments & Recommendations:**

It is recommended that correlation of these findings with the clinical phenotype be performed. Genetic counseling for the patient and at-risk family members is recommended.

This is a preliminary report because the variant has not yet been confirmed by Sanger sequencing.

**Gene Coverage:**

All genes have 100% of targeted bases sequenced to redundant coverage of 20x or greater with the following exceptions: APOB (99.39%), CACNA1B (96.45%), COL5A1 (98.03%), GRM5 (99.92%), KCNQ1 (94.28%), PKP2 (98.76%), PRKAG2 (99.83%), RYR1 (98.69%), TGFBR1 (93.56%). Further information, including specific coverage for this patient's sample, is available in the ExCID report.

**Methodology:**

1. eMERGE-Seq Version 2 NGS Panel: for the paired-end pre-capture library procedure, genome DNA is fragmented by sonicating genome DNA and ligating to the Illumina multiplexing PE adapters (reference 1). The adapter-ligated DNA is then PCR amplified using primers with sequencing barcodes (indexes). For target enrichment capture procedure, the pre-capture library is enriched by hybridizing to biotin labeled in-solution probes (reference 2) at 56°C for 16 - 19 hours. For massively parallel sequencing, the post-capture library DNA is subjected to sequence analysis on Illumina HiSeq platform for 100 bp paired-end reads. The following quality control metrics of the sequencing data are generally achieved: >70% of reads aligned to target, >99% target base covered at >20X, >98% target base covered at >40X, average coverage of target bases >200X. SNP concordance to SNPTrace genotype array: >99%. This test may not provide detection of certain genes or portions of certain genes due to local sequence characteristics or the presence of closely related pseudogenes. Gross deletions or duplications, changes from repetitive sequences may not be accurately identified by this methodology. Genomic rearrangements cannot be detected by this assay.

2. As a quality control measure, the individual's DNA is also analyzed by a SNP-array (Fluidigm SNPTrace panel (reference 3) ). The SNP data are compared with the NGS panel data to ensure correct sample identification and to assess sequencing quality.

3. Data are analyzed by the Mercury 3.4 (reference 4) pipeline. The output data from Illumina HiSeq are converted from bcl file to FastQ file by Illumina bcl2fastq 1.8.3 software, and mapped to the hg19 human genome reference by the BWA program (reference 5). The variant calls are performed using Atlas-SNP and Atlas-indel developed in-house by BCM HGSC. Copy number variants were detected using Atlas-pcnv v0, developed in-house by the BCM HGSC. Variant annotations are performed using the Cassandra tool, developed in-house. Neptune version $VERSION was used to match variants against curated variants in the VIP database version [/hgsccl/next-gen/neptune/vip/vip.2016-11-07] and generate this report.**

4. The variants were interpreted according to ACMG guidelines (reference 6) and patient phenotypes. Synonymous variants, intronic variants not affecting splicing site, and common benign variants are excluded from interpretation unless they were previously reported as pathogenic variants. Reviewed variants are added to the VIP database for inclusion on future reports. It should be noted that the interpretation of the data is based on our current understanding of genes and variants at the time of reporting.

Clinical interpretation and reporting are provided for pathogenic and likely pathogenic variants as requested by BMGL for the following 68 medically actionable genes: ACTA2, ACTC1, APC, APOB, BMPR1A, BRCA1, BRCA2, CACNA1A, CACNA1S, COL3A1, COL5A1, DSC2, DSG2, DSP, FBN1, GLA, HNF1A, HNF1B, KCNE1, KCNH2, KCNJ2, KCNQ1, LDLR, LMNA, MEN1, MLH1, MSH2, MSH6, MUTYH, MYBPC3, MYH11, MYH7, MYL2, MYL3, MYLK, NF2, OTC, PALB2, PCSK9, PKP2, PMS2, POLD1, POLE, PRKAG2, PTEN, RB1, RET, RYR1, RYR2, SCN5A, SDHAF2, SDHB, SDHC, SDHD, SMAD3, SMAD4, STK11, TGFBR1, TGFBR2, TMEM43, TNNI3, TNNT2, TP53, TPM1, TSC1, TSC2, VHL, WT1, the following non medically actionable genes: ANK2, ATM, ATP1A2, BMPR2, CACNA1C, CFH, CFTR, CHEK2, FLG, MC4R, MTHFR, NTRK1, SCN1A, SCN9A, SERPINA1, SLC2A10, TCF4, TCIRG1, TTR, TYK2, UMOD, VDR, the following medically actionable SNPs: rs77931234, rs387906225, rs79761867, rs386834233, rs113993962, rs397509431, rs6467, rs6025, rs80338898, rs1801175, rs1800562, rs28940579, rs61752717, rs193922376; and non medically actionable SNPs: rs151344623, rs76151636, rs111033258, rs786205104, rs786205103, rs147394623, rs121964990, rs121965064, rs121965063, rs104886456, rs201227603, rs74315447, rs61755320, rs724159981. For autosomal recessive disorders, only homozygous or biallelic variants will be returned. Variants in exon 3 of the FLG gene are not reported.

5. Variants related to patient phenotypes are confirmed by Sanger sequencing if the variant has been observed and confirmed fewer than 5 times by our laboratory or the Baylor Genetics Laboratory. Sanger confirmation is noted in the 'Notes' section of the tables if performed.

6. For the pharmogenomic variants, the star alleles are determined based on the variants detected by this assay. Alleles reported for TPMT are limited to *1, *2, *3A, *3B, *3C and *4. Alleles reported for CYP2C19 are limited to *1, *2, *4A, *4B, *5, *6, *7, *8, *17. If reported, alleles for DPD are limited to *1, *2A , *13 and rs67376798. Alleles reported for CYP2C9 are limited to *1, *2 and *3; and rs9923231 for VKORC1. Additional rare star alleles have been reported with reduced or no function for TPMT, CYP2C19 and DPD; however, the variants defining these additional star alleles are not detected with this assay. For SLCO1B1, this assay only detects rs4149056. The minor C allele at rs4149056 defines the SLCO1B1*5 (rs4149056 alone) but also tags the *15 and *17 alleles. Thus a *5 allele may represent a *15 or *17 allele. However, the magnitude of the phenotypic effect is similar for *5, *15, and *17 alleles.

** The VIP variant database was developed in conjunction with Baylor Genetics and the Partners Healthcare Laboratory for Molecular Medicine.

# Sample clinical report (Partners/Broad)

PARTNERS HEALTHCARE | PERSONALIZED MEDICINE

MASSACHUSETTS GENERAL HOSPITAL | BRIGHAM AND WOMEN'S HOSPITAL

**Laboratory for Molecular Medicine**

65 Landsdowne Street, Cambridge MA 02139

Phone: (617) 768-8500 Fax: (617) 768-8513

www.partners.org/personalizedmedicine/lmm

Unit Number(s):

Lab Accession: **PM-16-A07001**

Patient Name: **68282000, 10038000**

Birth Date: **1/1/1800**

Age Sex: **215 Year old Female**



## MOLECULAR DIAGNOSTICS REPORT

| | | | |
|---|---|---|---|
| **Specimen Type:** | DNA, Isolated - Blood, Peripheral (edit) | **Received Date:** | 9/1/2016 |
| **Related Accession(s):** | | **Referring Facility:** | HARVARD |
| **Referring Physician:** | EMERGE-CLINIC-TEST | **Referring Fac. MRN:** | |
| **Copies To:** | GENEINSIGHT<br>EMERGE-HUB GENEINSIGHT | **Lab Control Number:**<br>**Family Number:** | 10038000_68282000-0_SM-B3ZZZ<br>FAB123 |

**TEST DESCRIPTION -** Copy Number Variation Analysis
Sequence Confirmation Test eMERGE III Sequencing Panel

**TEST PERFORMED -** CNV-a; SeqConV2; EMERGE-pnlC

**INDICATION FOR TEST -** Not selected for trait

**RESULTS**

**DNA VARIANTS:**
Heterozygous c.338C>A (p.Ser113X), Exon 4, PMS2, Pathogenic

**INTERPRETATION:**
**Positive.** DNA sequencing of the coding regions and splice sites of 97 genes (see methodology section below) identified the variants listed above. Copy number analysis using NGS could not be completed because data did not meet quality standards for CNV detection. For a list of exons that are incompletely covered please see "Additional notes and disclaimers" section below.

SUMMARY:
This individual carries a Pathogenic variant in the PMS2 gene. The available information on this variant is described below. Disease-causing variants in the PMS2 gene are strongly associated with Lynch syndrome and this individual may be at risk for developing colorectal cancer / polyps.

ADDITIONAL NOTES AND DISCLAIMERS:
Disease penetrance and severity can vary due to modifier genes and/or environmental factors. The significance of a variant should therefore be interpreted in the context of the individual's clinical manifestations

DETAILED VARIANT INTERPRETATIONS:
p.Ser113X, c.338C>A (PMS2; NM_000535.5; Chr7g.6043336G>T; GRCh37):
The p.Ser113X variant in PMS2 has not been previously reported in individuals with Lynch syndrome and was absent from large population studies. This nonsense variant

**MOLECULAR DIAGNOSTICS REPORT**

leads to a premature termination codon at position 113, which is predicted to lead to a truncated or absent protein. Heterozygous loss of function of the PMS2 gene is an established disease mechanism in Lynch syndrome (http://www.ncbi.nlm.nih.gov/books/NBK1211/). In summary, this variant meets our criteria to be classified as pathogenic for Lynch syndrome (http://www.partners.org/personalizedmedicine/LMM) based upon predicted impact to the protein and absence in controls.

**RECOMMENDATION:**
Genetic counseling is recommended for this individual and their relatives. Familial variant testing is available for other relatives if desired. For assistance in locating genetic counseling services or disease specialists, please call the laboratory at 617-768-8500 or email at LMM@partners.org.

Please note that variant classifications may change over time if more information becomes available. Please contact us at 617-768-8500 or LMM@partners.org.

**TEST INFORMATION**

**BACKGROUND:**
The eMERGE (electronic MEdical Records and GEnomics) network combines DNA biorepositories with electronic health record (EHR) systems for large-scale discovery and clinical implementation research in genomic medicine. A main goal is the return of genomic testing results to patients in a clinical care setting. In phase III, participating sites are sequencing 109 clinically relevant genes in ~25,000 participants using a custom next generation sequencing panel.

**METHODOLOGY:**
Test content (target region): This test includes 109 genes (including the ACMG56 genes; PMID: 23788249, and additional genomic positions for known variants. For reference sequences exons/positions covered see http://personalizedmedicine.partners.org/Laboratory-For-Molecular-Medicine/).

Note that this test may not detect variants in regions with difficult sequence contexts (e.g. high or low GC content) and is generally not designed to detect deep intronic variants as well as variants in the 5' and 3'UTR. Regions with high sequence homology are only included in this test if a unique Sanger sequencing assay can be designed to rule out false positive calls.

Sample preparation, sequencing, variant calling and confirmation: This test is performed by next generation sequencing using sonicated genomic DNA (Covaris) followed by target enrichment (Illumina Rapid Capture Custom Kit), Illumina HiSeq sequencing (76 bp paired-end reads) and alignment/variant calling (BWA/GATK). A custom script is used to generate calls for the individual genomic positions. Sample identity is confirmed by comparing NGS derived genotypes of a custom set of SNPs to results generated for the same specimen using a fingerprinting genotype array. Samples with ≤95% of the target region covered at ≥20X are failed and repeated. Copy number variants (CNVs) of ≥3 exons are detected by an in-house developed tool (VisCap, PMID: 26681316). This assay is 99.33% sensitive to detect single nucleotide variants (95% CI = 96.30-99.88%), 100% sensitive to detect indels (95% CI = 79.61-100.00%) and 100% sensitive to detect CNVs (n=4). All variants included on this report are confirmed (SNVs and indels: Sanger sequencing, CNVs: ddPCR).

# Sample clinical report (Partners/Broad)

**Laboratory for Molecular Medicine**
Partners HealthCare Personalized Medicine

Accession: **PM-16-A07001**
Patient Name: **68282000, 10038000**

## MOLECULAR DIAGNOSTICS REPORT

Variant annotation and filtration: All variants within the coding sequence of the included genes (default: exons +/- 5 bp) are subjected to the following process: Variant annotations are derived from ExAC (vs 0.3), ClinVar (April 2016 release), HGMD (2016.1), 1000 Genomes (Phase 3), Alamut Batch (vs 1.4.4), (dbnsfp vs 3.1), and LMM's GeneInsight knowledge base (vs 5.3.2). The following variant types are further analyzed: a) Loss of function variants with a minor allele frequency (MAF)<1%, b) Variants previously classified as pathogenic or likely pathogenic regardless of MAF, c) ClinVar pathogenic or likely pathogenic and HGMD DM variants with a MAF<5%.

Variant interpretation and clinical reporting: Variants assessment is based on in-house developed expert criteria and the most recent ACMG classification framework (PMID: 25741868) with disease and gene-specific modifications when applicable. Please note that variant classifications can change over time. Reporting is restricted to pathogenic and likely pathogenic variants in a subset of eMERGE network genes and variants consisting of 62 genes and 1 variant in an additional gene: ACTA2, ACTC1, APC, APOB, BRCA1, BRCA2, CACNA1C, CACNA1S, COL3A1, DSC2, DSG2, DSP, FBN1, GLA, HNF1A, KCNE1, KCNH2, KCNJ2, KCNQ1, LDLR, LMNA, MEN1, MLH1, MSH2, MSH6, MUTYH, MYBPC3, MYH11, MYH7, MYL2, MYL3, MYLK, NF2, OTC, PCSK9, PKP2, PMS2, PRKAG2, PTEN, RB1, RET, RYR1, RYR2, SCN5A, SDHAF2, SDHB, SDHC, SDHD, SMAD3, SMAD4, STK11, TGFBR1, TGFBR2, TMEM43,
TNNI3, TNNT2, TP53, TPM1, TSC1, TSC2, VHL, WT1 and HFE (rs1800562). Carrier status for autosomal recessive conditions will not be reported.

This test was developed and its performance characteristics determined by the Laboratory for Molecular Medicine at Partners HealthCare Personalized Medicine (LMM, 65 Landsdowne St, Cambridge, MA 02139; 617-768-8500; CLIA#22D1005307). It has not been cleared or approved by the U.S. Food and Drug Administration (FDA). The FDA has determined that such clearance or approval is not necessary.

**REFERENCES:**

Hendriks YM, Jagmohan-Changur S, van der Klift HM, Morreau H, van Puijenbroek M, Tops C, van Os T, Wagner A, Ausems MG, Gomez E, Breuning MH, Bröcker-Vriends AH, Vasen HF, Wijnen JT. 2006. Heterozygous mutations in PMS2 cause hereditary nonpolyposis colorectal carcinoma (Lynch syndrome). Gastroenterology. 130(2):312-22.

REPORT by Matthew Lebo Ph.D., on Friday September 09, 2016 at 04:22:23PM
Final Diagnosis by **Matthew Lebo Ph.D.**, Electronically signed on Monday September 12, 2016 at 10:56:03AM