

1 **HARMONIZING CLINICAL SEQUENCING AND INTERPRETATION**

2 **FOR THE EMERGE III NETWORK**

3
4 The eMERGE Consortium*

5 *detailed contributions can be found below

6 Correspondence:
7 Richard A. Gibbs, PhD
8 Baylor College of Medicine
9 1 Baylor Plaza #226, Houston, TX 77030
10 agibbs@bcm.edu

11
12 Heidi L. Rehm, PhD
13 Center for Genomic Medicine
14 Massachusetts General Hospital
15 Simches Research Building, CPZN-5-812
16 185 Cambridge St, Boston, MA 02114
17 617-643-3217
18 hrehm@mgh.harvard.edu
19

20
21 **Keywords:** eMERGE; electronic health record; clinical sequencing; harmonization; next generation
22 sequencing
23
24
25
26
27
28
29
30
31
32
33
34

35 **ABSTRACT**

36
37 **Background:** The eMERGE III Network was tasked with harmonizing genetic testing protocols linking multiple
38 sites and investigators.

39 **Methods:** DNA capture panels targeting 109 genes and 1551 variants were constructed by two clinical
40 sequencing centers for analysis of 25,000 participant DNA samples collected at 11 sites where samples were
41 linked to patients with electronic health records. Each step from sample collection, data generation,
42 interpretation, reporting, delivery and storage, were developed and validated in CAP/CLIA settings and
43 harmonized across sequencing centers.

44 **Results:** A compliant and secure network was built and enabled ongoing review and reconciliation of clinical
45 interpretations while maintaining communication and data sharing between investigators. Mechanisms for
46 sustained propagation and growth of the network were established. An interim data freeze representing 15,574
47 sequenced subjects, informed the assay performance for a range of variant types, the rate of return of results
48 for different phenotypes and the frequency of secondary findings. Practical obstacles for implementation and
49 scaling of clinical and research findings were identified and addressed. The eMERGE protocols and tools
50 established are now available for widespread dissemination.

51 **Conclusions:** This study established processes for different sequencing sites to harmonize the technical and
52 interpretive aspects of sequencing tests, a critical achievement towards global standardization of genomic
53 testing. The network established experience in the return of results and the rate of secondary findings across
54 diverse biobank populations. Furthermore, the eMERGE network has accomplished integration of structured
55 genomic results into multiple electronic health record systems, setting the stage for clinical decision support to
56 enable genomic medicine.

62 INTRODUCTION

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

70

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

77

78 A second group of practical obstacles arises from the heterogeneity of sites and tools used to collect patients'
79 and participants' data. On many occasions, even straightforward measurements cannot be meaningfully
80 combined when they are derived from different sites, if they are obtained with different instruments, or from
81 different clinical genetic testing laboratories using different molecular reagents. Thus, some information
82 regarding the method of measurement must accompany the measurement data for harmonization, integration
83 and standardization across populations.

84

85 Despite these challenges, the desire to improve medical care by advancing genetic discovery provides incentive
86 for data harmonization. The underlying processes, including participant interaction as well as methods for
87 phenotyping, sequencing, and genetic variant interpretation therefore need to be studied and standardized.
88 Further, the demands of harmonized data flow, storage, and management must be met. Each process must also

89 attend to the tension between respect for patient privacy (e.g. HIPAA laws) and the ability to access data to
90 facilitate research. A list of practical obstacles is presented in Table 1.

91
92 The current phase (III) of the United States National Institute of Health's Electronic Medical Records and
93 Genomics (eMERGE) program¹ aims to study and improve these processes for delivery of clinical and research
94 data, in a multi-center network, while providing actionable genetic results derived from a next-generation
95 sequencing platform to eMERGE research participants. The network builds upon experience with participant
96 consent, obtaining clinical data from the EHR, genotyping and return of results, expanding processes to inform
97 care and catalyze research.

98
99 **SUBJECTS AND METHODS** (More details of certain methods are included in Supplementary Material)

100
101 **(i) eMERGEseq Panel Overview:**

102 **Panel Design and Content:** A gene panel comprising a total of 109 genes and approximately 1,400 SNV sites
103 was informed by network input. The design process considered potential actionability of findings and local
104 research interests, as well as gene size. The 109 genes included 56 based upon the American College of Medical
105 Genetics and Genomics (ACMG) actionable finding list². Additionally, each site nominated 6 genes relevant to
106 their Specific Aims, including discovery-focused genes associated with clinical phenotypes in need of further
107 study. All nominated genes apart from titin (*TTN*), which was excluded due to its size, were included in the final
108 panel design for a total of 109 genes. Further, eMERGEseq content included several categories of single
109 nucleotide variants (SNVs): 1) ancestry informative markers and QC/fingerprinting loci (N=425), 2) a suite of
110 SNVs selected to inform HLA type (N=272), 3) pathogenic SNVs in genes not included on the panel for which
111 return of results was planned (N=14), 4) pathogenic or putatively pathogenic SNVs in genes not included on the
112 panel for which return of results was not planned (N=55; for some, penetrance is poorly understood), 5) SNVs
113 related to site-specific discovery efforts (N=718), and 6) pharmacogenomic variants (N=125), selected based on
114 potential actionability, allele frequency and space available on the platform. A summary of all eMERGEseq
115 content can be found in Table 2, with additional details provided in Table S1. All sequence and SNV data are

116 shared across the network for research, and a subset of the content, namely the clinically actionable variants
117 associated with disease or drug response, are included in clinical reports for return to the participants.

118
119 **(ii) Panel Sequencing:**

120
121 **Reagent:** The gene and SNV list was used to direct construction of targeted capture platforms at two sequencing
122 centers (SCs): The Baylor College of Medicine Human Genome Sequencing Center [BCM-HGSC], Houston TX;
123 Broad Institute and Partners Laboratory for Molecular Medicine, Cambridge, MA. Broad used Illumina Rapid
124 Capture probes for this panel and the BCM-HGSC used Roche-Nimblegen methods. Each group created in-
125 solution capture probes spanning the entire targeted regions of the eMERGEseq panel. Probes were designed
126 to be complementary to specified exons or SNV sites with a minimum span of 100 nucleotides. Tiling was limited
127 to exonic sequence and analyses included +/- 15 intronic flanking bases (Figure S1).

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

134 **Sequencing and Primary Analysis:** Samples from DNA capture using the custom capture reagents were
135 sequenced using standard Illumina technologies. Post-sequence processing at each site utilized preferred
136 alignment and variant calling algorithms. The variant calling pipeline at Broad incorporates Picard deduplication,
137 BWA alignment, and GATK variant calling for SNVs and short InDels³. At the BCM-HGSC the alignment using
138 BWA-MEM and variant calling using Atlas were instantiated within the Mercury Pipeline⁴.

139 **Panel Fill-in:** A common set of reference samples were initially sequenced at each SC. The chosen parameters
140 to monitor performance were coverage of targeted sequence and percentage of the targeted bases at or above
141 20X coverage. Both groups sequenced cohorts of control samples and identified systematically poorly-covered
142 bases as those with less than 20X coverage in >10% of tested samples. Based on this conservative threshold,

143 both groups went through a process of enriching with more targeting probes ('fill-in'), to boost underperforming
144 regions, prior to final validation. The reagent performance is described in Table 3, with additional details in
145 supplementary Table S2.

146 **Copy Number Variant (CNV) Calling:** CNV calling at Partners/Broad was performed using VisCap, which infers
147 copy number changes from targeted sequence data by comparing the fractional coverage of each exon in a
148 gene to the median of these values across all samples in a given sequencing run⁵. BCM-HGSC CNV calls were
149 made via Atlas-CNV, in-house software that combines outputs fromXHMM^{6,7} and the GATK DepthOfCoverage
150 tool⁶. Like VisCap, Atlas-CNV infers the presence of CNVs from normalized coverage differences to other
151 samples in the same sequencing batch, and refines these predictions with a pair of quality control metrics⁸. CNV
152 calls were confirmed by orthogonal technology - Droplet Digital PCR (Bio-Rad, Hercules, CA) at Partners/Broad
153 and Multiplex Ligation-dependent Probe Amplification (MRC-Holland, Amsterdam, Netherlands) at the BCM-
154 HGSC. Detected CNVs were filtered based on clinical site's gene reporting preferences and ClinGen
155 haplosensitivity and tri-sensitivity scores (<https://search.clinicalgenome.org/kb/gene-dosage>), and then manually
156 reviewed. Partners/Broad required a minimum of three contiguous exons for reporting, BCM-HGSC required
157 two.

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

163 **Ongoing Proficiency:** Ongoing proficiency testing involved interlaboratory exchange of previously tested
164 eMERGE samples and CAP proficiency testing for general sequencing platforms with all results concordant to
165 date (see Supplementary Materials: Supplemental Methods).

166 (iii) Variant Interpretation:

167 **General Approach to Interpretation:** Variant classifications from both laboratories were based on
168 ACMG/Association of Medical Pathology (ACMG/AMP) criteria with ClinGen Sequence Variant Interpretation
169

Working Group modifications as well as additional specifications for some of the eMERGEseq genes as established by ClinGen Expert Panels⁹. Additional local data accrued from previous case studies was 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 assessment¹⁰.

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 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; interpret diplotypes; and select drugs for therapeutic recommendations, guided by CPIC guidelines. Twenty PGx variants in seven genes were deemed to be clinically actionable and selected for return to participants. Table S4 includes details of the PGx genes and variants reported and the drugs associated. For two PGx genes, *CYP3A5* and *SLCO1B1*, 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 Material (Sample eMERGE report HGSC-CL, LMM Sample PGx Batch Report). The CPIC drugs that were included in the PGx report were largely the same with some minor differences (see Table S4).

197 **(iv) Data Management:**

198 **Sample Intake:** Each site was provided barcoded tubes by the SC for DNA shipping. Sample identifiers and
199 metadata were uploaded using an 'eMERGE requisitioning sheet' via secure portals^{11,12}. The requisitioning
200 spreadsheet contains fields for sample information (name [optional], sex, date of birth/age, US state of residence,
201 site-specific ID), as well as eMERGE-specific metadata including; patient 'disease area' (from a list defined by
202 the network - see supplementary material for details), disease status and test indication, eMERGE project ID
203 and barcode number on the tube. An additional option was to add phenotype terms in a free-text field, primarily
204 based on the MonDO ontology and occasionally additional local codes largely derived from Human Phenotype
205 Ontology (HPO) terms (See Supplemental Material: preferred indication terms). A simple .csv file structure was
206 used by both SCs so that sites could upload all metadata at the time of sample batch shipment. For the BCM-
207 HGSC SC, the sample accession was directly into a cloud environment, managed by DNAnexus, while for the
208 Partners-Broad site a custom portal operating in the Broad's local environment was employed for intake followed
209 by transfer to the GeneInsight system for analysis and reporting; all systems were HIPAA compliant. Local
210 identifiers were then generated to track the samples as they progressed through DNA sequencing and variant
211 calling. Orders were reviewed and approved by the SCs prior to sample shipping and accession. Upon receipt,
212 the samples were subjected to volume and concentration quality control checks.

213 **Data Delivery and Reporting:** Each SC developed custom reporting methods (see Supplementary Material for
214 examples). Partners/Broad site users have a unique, password-protected account and are only able to view
215 orders and metadata from their own site. The Broad portal authorization procedures are customized to allow for
216 secure transfer of sequencing output files and metadata to both Partners and DNAnexus via APIs. The BCM-
217 HGSC sites are delivered reports from the DNAnexus environment via DNAnexus APIs. Users were provided
218 individual logins for accessing pdf reports and structured content in a harmonized .xml format.

219 **GeneInsight:** Partners/Broad sites used the commercial tool, GeneInsight (Sunquest Information Systems,
220 Tucson, AZ), for local report management¹³. This tool was configured to create a De-identified Case Repository
221 (DCR) which contains a de-identified record of all cases and associated variants from both Partners/Broad and
222 the BCM-HGSC supported sites.

223 **DNAnexus Data Commons:** The BCM-HGSC clinical sites were provided with two data access points in the

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

230 **Variant Updates:** Two complementary mechanisms were developed to enable delivery of variant updates from
231 the SCs to the sites as new evidence leading to a classification change becomes available. At Partners/Broad,
232 individual participant results are stored in an eMERGE-specific instance of the GeneInsight database that is
233 linked to Partners LMM's GeneInsight instance enabling communication of variant updates¹⁴. If Partners updates
234 a variant, sites that have signed up receive proactive notification emails if a reported variant identified in one or
235 more of their cases is updated. Hyperlinks are provided in those emails that allow sites to directly access updated
236 information on the variant in each case, which facilitates the choice to return an updated result to a participant.
237 In addition, Partners is generating an .xml file for each variant interpretation change alert, which sites can
238 consume through other electronic interfaces. At the BCM-HGSC, participant results are stored in a database that
239 is routinely queried for variants with new actionable interpretations. If such a variant is found in a previously-
240 reported sample, an amended report is issued via DNAnexus and sites are notified. Variant updates are included
241 in the ongoing variant interpretation harmonization process described above.

242 243 **(V) Data Freeze and Raw Data Storage:**

244 In order to analyze preliminary results from the eMERGE III eMERGEseq data, an interim freeze of samples
245 sequenced by November 2017 was generated. These 15,754 samples (9633 from Baylor and 6121 Broad-
246 Partners) and the available associated data are described in detail in the Supplementary Material. The
247 associated BAM, xml and vcf files are available on the eMERGE Commons, accessible to sites as well as outside
248 investigators who apply for access (<https://emerge.mc.vanderbilt.edu/collaborate/>). Data are also submitted to
249 dbGaP for controlled public access.

251 RESULTS

252 (i) Network Overview:

253 The eMERGE III network established a Clinical and Discovery Platform that consists of 11 clinical study sites,
254 two DNA SCs and a coordinating center (CC) (Figure 1). Participants were enrolled at each site, blood collected,
255 DNA extracted locally and sent to one of two SCs for targeted sequencing. Analysis and interpretation of the
256 DNA sequence data was performed at each SC, and the data returned to the clinical sites for return to
257 participants. Raw data were accrued for data mining purposes by eMERGE investigators and approved affiliates.
258 Subsequently raw data are released to dbGaP and interpreted variants to ClinVar.

259 An early decision of the program was to utilize DNA capture ‘panels’ of approximately 500 kb, in order to generate
260 genomic data from the eMERGE participants, as an alternative to genotyping, whole exome sequencing (WES)
261 or whole genome sequencing (WGS). This choice reflected a balance between available fiscal resources and a
262 reasonable selection of content to explore return of actionable results and focused discovery efforts. The use of
263 the panel enabled testing of 109 genes and approximately 1400 additional sites of single nucleotide variation in
264 each sample. Across the network, ~25,000 samples are being assayed, ~2500 from each site. The study is
265 therefore large enough to allow robust analysis of specific phenotypes, as well as to gain experience with a
266 sufficient number of patients at each site to develop processes to support the return of actionable genetic results.

267 Prior population studies suggested that the genes included on the panels would reveal thousands of newly
268 identified single nucleotide and structural variants. A small subset of these would be expected to be pathogenic,
269 and the program aimed to report to participants only those variants that were pathogenic or likely pathogenic
270 according to the ACMG/AMP guidelines¹⁵, or those with actionable pharmacogenomic associations. In addition,
271 it was aimed to provide data from the panel that informed possible pharmacological responses. Each site would
272 have the option of a customized clinical reporting framework, as well as full access to all network data to guide
273 decisions and harmonize interpretations.

274 This elaborate network reflects a real-world situation, where a full complement of testing, reporting, and research
275 require coordination and harmonization of many components. First, the selection of gene targets and the rules
276 for reporting must agree. Next, the technical aspects of DNA capture and sequencing required standardization
277 and ongoing comparison. The DNA changes must be interpreted and reported with the same conclusions,

278 regardless of where testing occurred. Finally, file structure standardizations and data management practices
279 must be organized. A detailed list of components (Table 1) that require coordination and harmonization illustrates
280 the magnitude of the challenge.

281 **(ii) Technical Validation of Capture Panels:**

282 Coordination and harmonization of the DNA capture panel process at the two CAP/CLIA certified DNA
283 sequencing laboratories was demanding because in addition to different DNA capture reagents, the local
284 processes of sample preparation, library construction, hybrid capture, and sequencing represented complex
285 workflows with many variables. As an alternative to compelling each laboratory to adopt unfamiliar methods, the
286 harmonization was achieved through phases of coordinated design, comparing initial high level technical
287 performance and via ongoing monitoring of proficiency (Figure 2a). The harmonization process aimed to reduce
288 any impact on the overall program of the heterogeneity of capture reagents or sequencing methods between two
289 sites and for the end users to be able to compare data from each laboratory without batch effects.

290
291 Design was coordinated by first agreeing on the intended limits to reporting, e.g. number of bases adjacent to
292 exons to be reported (see methods, Figure S1). Each laboratory employed slightly different criteria for the
293 selection of the range of transcripts to be tested, reflecting a lack of harmony of public databases. Possible
294 differences in design were resolved by selection of the union of all possible exons to be considered, and validated
295 by iterative sharing of the capture design files ('bed files'). The detailed design specifications can be found in
296 Table S1.

297
298 Preliminary testing of the technical performance of the two capture reagents utilized both local test samples and
299 a shared sample reference set (see methods). The technical performance was shared between the SCs by
300 measuring the coverage of individual bases and other key technical metrics (Table 3, Table S2). Overall
301 sequence coverage goals and the extent to which poorly covered regions could be tolerated were agreed upon
302 *a priori*, and the technical comparison was straightforward between SCs. In general, the sequencing reagents
303 performed well, although the presence of some uncovered bases in the first panel designs led each group to
304 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 ~250X to 400X (Broad) and input amounts of 250 ng and 500 ng (Baylor). Summary results of the respective validation studies are shown in Table 3. 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*, *CACNA1B*) could not be recovered by either platform.

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 ~420X, % of targeted bases covered $\geq 20X$ is 99.7%, and % of targeted bases with zero coverage is 0.17%. These metrics, collected from >7000 production samples, closely match the performance of the validation set. Mean coverage of the BCM-HGSC production samples is ~340X, % of targeted bases covered $\geq 20X$ is 99.8%, and % of targeted bases with zero coverage is 0.04%. These metrics, collected from >9600 production samples, also closely match the performance of the validation set.

(iii) 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

331 research on these genes, reflecting the diverse goals of the eMERGE network including discovery as well as
332 return of results.

333
334 A subset of the 1415 site submitted SNVs were for fingerprinting and ancestry, HLA, or PGx categories, or have
335 been previously classified as likely benign or benign and were thus excluded for further analyses of potential
336 pathogenicity. The remaining 136 variants were considered for further clinical assessment. Seventy three
337 variants were classified as either likely pathogenic or pathogenic by at least one of the SCs. Of these, 19 had
338 discrepant classifications between the two SCs. These were resolved by variant re-assessment and scoring on
339 published evidence as well as combined internal evidence from both SCs. For two variants, the eMERGE Clinical
340 Annotation WG was consulted to assist in resolving interpretation differences. A final list of 69 pathogenic/likely
341 pathogenic variants was established and further considered for actionability analyses (Figure 3).

342
343 The eMERGE Clinical Annotation WG evaluated the 35 non-ACMG56 strong/definitive genes and 69 associated
344 pathogenic/likely pathogenic variants, based on whether there was a substantially increased risk of serious
345 disease that could be prevented or managed differently if the risk were known. In addition to the ACMG56, 12
346 genes and 14 variants were deemed actionable by the eMERGE Clinical Annotation WG and placed on a
347 consensus list of returnable content (Table 2, Figure 3). While sites agreed that this list represented content that
348 would generally be returnable, some sites requested modifications be made to the consensus list based to their
349 return of results plans (Figure 4). For example, of the 11 sites, one that included pediatric biobank participants
350 opted not to report variants in genes that increase risk of adult onset diseases but are not actionable during
351 childhood. Also, not all sites chose to return *HFE* p.Cys282Tyr homozygotes. Four other sites requested
352 additional genes and SNVs that were not on the consensus list. A full list of the content that was returned for
353 each site can be found in Table S3. Additionally, one site is returning variants of uncertain significance in 13
354 colorectal cancer genes for a subset of their samples derived from a colorectal cancer cohort. Another clinical
355 site requested genotypes at twelve SNP sites associated with low-density lipoprotein cholesterol (LDL-C) risk be
356 included on their report.

358 **(iv) Data Intake and Delivery:**

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

365
366 (a) **Data intake:** Data intake and accessioning for each site was facilitated by an agreement of the specific PHI
367 metadata to be supplied with each sample, as well as an agreement of a set of required 'indications for testing'
368 that represented the primary phenotype data that tracked each sample through the network (see Methods).

369
370 (b) **Clinical reporting:** Within each pipeline, the standard validated product was a pdf report that was returned
371 to the clinical investigators (see supplementary material for examples of reports: (Sample eMERGE report
372 HGSC-CL, Sample eMERGE report HGSC-CL XML, Sample eMERGE report Partners Broad, Sample eMERGE
373 report Partners Broad XML). Each clinical site had custom requirements for the report content, that reflected
374 local preferences for data to be returned to patients. Each SC also had different reporting requirements – for
375 example, some sites requested negative reports, others only returned positive reports¹⁶. Most sites also
376 requested data in structured formats to enable direct integration onto their local EHRs.

377
378 The five clinical sites served by the Partners-Broad CSG received results delivered through the GeneInsight
379 platform, which enabled storage and query of clinical reports. The six sites served by the BCM-HGSC utilized
380 custom applications developed for report delivery. Possible difficulties in data sharing between different parts of
381 the network were anticipated and obviated by development of an agreed .xml standard. This standard was based
382 upon the GeneInsight specifications and facilitated communication across all components (See Methods and¹⁷).

383
384 The clinical sites therefore had two options – they could either use a stand-alone tool for report data management

385 or alternatively the report data could be parsed into local customized systems.

386
387 For PGx data, in addition to receiving results in pdf reports (either individual reports by the BCM-HGSC, or batch
388 reports by Partners-Broad), a standardized data format was also developed to deliver structured PGx data in the
389 form of both variant level and diplotype results allowing sites to directly integrate PGx results into the EHR for
390 clinical decision support.

391
392 (c) **Research and Discovery via Data Commons and the De-identified Case Repository:** Finally, the network
393 required all deidentified data to reside together, to enable data mining for both basic research and to better inform
394 clinical decision making with access to larger clinical datasets. There were two independent but complementary
395 mechanisms for this. First, the GeneInsight tool maintains a record of all returned variant data from both sites in
396 a de-identified case repository allowing an easy search interface for clinically reported variants.

397
398 A second site maintained the full set of eMERGE raw data in a cloud environment, managed by a middle-ware
399 vendor, DNAnexus. This 'eMERGE Commons' was structured to house each DNA sequence file in the BAM
400 format, as well as the annotations for the data in a vcf format. As clinical report delivery for the data generated
401 in the Baylor SC also utilized the DNAnexus infrastructure, the full set of identified clinical reports and de-
402 identified raw data were both resident in the cloud. The access permissions for the data were managed to allow
403 only the clinical providers to access their patients' clinical reports. The full set of raw data was available to all
404 eMERGE investigators as sensitive PHI information had been removed.

405
406 **(v) Variant Interpretation Harmonization:**

407 To ensure consistency of results being returned across the eMERGE consortium, variant interpretation was
408 harmonized between the SCs (Figure 5). In a pre-test launch, both SCs exchanged variants in reportable genes
409 from their respective databases, totalling 23,663 unique variants. Of those, 1047 were previously classified by
410 both SCs. The pre-test launch data exchange showed 90% concordance in variant classification among variants
411 classified as VUS, likely pathogenic and pathogenic by at least one SC. When likely pathogenic and pathogenic

412 variants were grouped together, the concordance was 93%. When all variant classifications were considered,
413 including benign vs likely benign, the data showed a 67.5% concordance. However, only 28, or 3% of the variants
414 were deemed to affect reporting (VUS vs pathogenic 1.9%, VUS vs likely pathogenic 1.1%). The two SCs
415 resolved all differences that would affect inclusion on clinical reports (i.e. pathogenic/likely pathogenic versus
416 VUS).

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

424 Three variants (p.Ile1307Lys in *APC*, p.Met54Thr in *KCNE2*, and p.Asp85Asn in *KCNE1*) were noteworthy as
425 the interpretations were more discrepant upon initial assessment (i.e. 'two-steps': pathogenic vs likely benign),
426 although the evidence used by both centers was identical. These represented variants that have significantly
427 reduced penetrance, leading to difficulties applying the ACMG/AMP classification framework, which is designed
428 primarily for highly penetrant Mendelian disorders. Nevertheless, some sites chose to return the *APC* variant as
429 it imparts a two-fold risk of colorectal cancer in Ashkenazi Jewish individuals, even though its effect in other
430 populations is unclear. Other sites elected to return the *KCNE2* variant, as it has been associated with variable
431 presentations such as arrhythmias, sinus bradycardia and long QT syndrome^{18–201}. This type of classification
432 discordance highlights the need for guidance on classification terminology for low penetrance variants for not
433 only the eMERGE network, but for the entire medical genetics community.

434 **(vi) Return of Results and Aggregate Findings:**

435 As of December 2017, 15,754 cases had been collected and analyzed via the eMERGEseq panel. To coordinate
436 analyses, these samples were included in a 'data freeze' termed 'eMERGEseq Data Freeze 1.0' (see
437 supplementary methods). All of the 15,754 data freeze samples have passed through at least the primary variant

438 assessment and review stage. For these assessed cases, a total of 1,913,377 variants were detected. A subset
439 of these were excluded from further analyses due to a LB/B classification by the SCs or by an auto-classification
440 pipeline based on allele frequency thresholds, or for having a low quality score. The remaining variants
441 underwent a filtration process which returns a) predicted loss of function variants with a minor allele frequency
442 (MAF) <1%, b) variants previously classified by the SCs as Likely Pathogenic(LP)/Pathogenic(P) regardless of
443 MAF, and c) ClinVar P/LP as well as HGMD “DM” variants with a MAF<5%. This pipeline resulted in 4786 unique
444 variants requiring further assessment. After expert review, these were further categorized as Benign (1%), Likely
445 Benign (8%), VUS (69%), LP (7%), P (12%) or deemed as low penetrance risk alleles (0.5%). In addition, 95
446 unique copy number variants have been detected across the reviewed samples, with 74 gains and 34 losses. Of
447 these, 35% were deemed reportable and were returned to sites. In summary, these data lead to a total of 679
448 cases projected to have a LP/P variant that would require a positive report to be issued.

449
450 Results being returned to sites currently fall into three categories: 1) Indication-based returnable results that
451 include all sequence and copy number variants related to the site-provided indication for testing, 2) non-
452 indication-based consensus returnable results that include all sequence and copy number variants in genes and
453 SNVs comprising the consensus list of returnable content (see clinical content validation section) that are not
454 related to indication for testing, and thus considered secondary findings and 3) non indication-based site-specific
455 returnable results which include variants in additional site-requested genes that are not on the consensus list
456 and not related to the indication for testing. Additionally, both SCs are returning results on pre-selected PGx
457 SNVs as either addendums to individual patient reports or in a batch report that contains up to ~185 samples
458 (See Methods).

459
460 The positive rate for each category of findings is depicted in Figure 6. For the 15,754 cases that have been
461 reviewed, 5,909 (38%) had an indication for testing. Of these, 115 (1.95%) had positive findings relevant to the
462 indication for testing. Moreover, of the 15,754 individuals sequenced, 681 (4.5%) had additional/secondary
463 findings of medical significance in genes and SNVs from the consensus list, that are being returned to
464 participants. About 4,073 participants (26%) were enrolled in sites who were interested in returning Pathogenic

465 and/or Likely Pathogenic variants in additional genes or SNVs that were not on the consensus list. In 153 cases
466 (3.5%), a non-indication based, site specific returnable Pathogenic or Likely Pathogenic variant was identified.
467 About half of these variants were in the *CHEK2* tumor suppressor gene, and are associated with an increased
468 risk for a variety of cancers. Other variants were found in genes associated with cardiac disease, familial
469 hypercholesterolemia and hemochromatosis (Figure 6). For indication-based assessments, detection rates were
470 highest for hyperlipidemia (44%), colorectal cancer/polyps (34%) and breast/ovarian cancer (19%). Some
471 phenotypes had no disease-causing variants identified due to either the absence of genes causative for the
472 disorders on the eMERGEseq panel or the lack of a clear monogenic disease etiology for the disorder (e.g.
473 abnormality of pain sensation, pediatric migraine). The rate of Pathogenic/Likely Pathogenic variants detected
474 in participants without a clinical indication differed from site to site, ranging from 1.8% to 17%, depending upon
475 the basis for participant selection, which were reflective of the underlying study designs of the individual sites.
476 The overall positive rate for secondary findings was skewed higher for one site given that 1251 participants of
477 the Geisinger cohort were preselected for a suspicious variant in a parallel exome study²¹. On the other hand, 2
478 sites had lower rates than expected either because their cohort had an indication related to genes in the
479 secondary findings list that led to the removal of these genes from secondary findings reporting or because the
480 site did not choose to return all results from the consensus list. When data from Geisinger participants
481 preselected for a suspicious variants were excluded, the frequency of secondary findings was similar across
482 sites, ranging from 2.7% to 4.9%, suggesting that the complexity of the network did not distort these results, and
483 reflecting the success of the data and process harmonization. A further analysis of the factors that influence the
484 rate of secondary findings return is underway (A. Gordon et al., 2018, American Society of Human Genetics,
485 abstract).

486
487 For PGx results, reports depicting genotype and related diplotype data, including whether the reported diplotype
488 for each gene and resulting phenotype would result in a recommendation to modify dosage, have been used for
489 approximately 9,000 participants from seven sites. Overall, the frequency of the reported diplotypes were
490 concordant with the CPIC published frequency tables for each major race/ethnic group²².

492 One difference for diplotype interpretation was particularly informative. When both rs1800460 and rs1142345
493 are identified in the thiopurine methyltransferase (*TPMT*) gene, it cannot be ascertained whether these variants
494 are in *cis*, resulting in a *TPMT*1/*3A* diplotype and intermediate metabolizer phenotype, or in *trans*, resulting in
495 a *TPMT*3B/*3C* diplotype and a poor metabolizer phenotype. One SC emphasized the more common diplotype
496 in their report, while the other emphasized the higher risk of the rarer diplotype under some drug regimens. With
497 input from the sites and the eMERGE PGx working group, it was decided that the more common genotype would
498 be reported with a warning that the rarer genotype could not be ruled out.

499
500 The majority of returned data reflected variants with relatively clear interpretations for participants, with variants
501 that either had a large body of published evidence or were straightforward to interpret. Several cases, however,
502 reflect interesting and unexpected findings.

503
504 The first finding involved what appeared to be a whole chromosome gain of chromosome 12. An NGS-based
505 CNV calling algorithm detected a gain in all exons of six eMERGEseq genes on chromosome 12 (*CACNA1C*,
506 *PKP2*, *VDR*, *MYL2*, *HNF1A* and *POLE*), which was confirmed by ddPCR. The *CACNA1C* and *POLE* genes are
507 located near the telomeric end of the chromosome 12 p- and q- arms respectively, supporting a whole
508 chromosome gain. Given that chromosome 12 trisomies are embryonic lethal, this CNV was assumed to be
509 either of somatic origin or occurring as a mosaic variant. The former scenario is more likely as trisomy 12 is the
510 most common somatic chromosomal aberration in chronic lymphocytic leukemia (CLL) but has also been
511 observed in other B-cell lymphoproliferative disorders and is associated with a less favorable prognosis²³. Rarely,
512 trisomy 12 has been reported as a mosaic variant in individuals with a variety of clinical phenotypes ranging from
513 reportedly normal to multiple congenital anomalies, dysmorphic features and developmental delay^{24–28}. Most of
514 these were identified prenatally, with less than 10 cases reported postnatally and even fewer detected in
515 peripheral blood (for reviews see^{27,28}). Additional clinical information provided by the site indicated that this
516 patient has a complex medical history including diabetes, heart disease and a diagnosis of colorectal cancer at
517 87. While this finding is from a blood draw in early January 2016, this individual's last complete blood count in
518 2010 showed no evidence of increased lymphocytes or any other abnormality suggesting a CLL diagnosis. While

519 this type of result was not anticipated within the reporting scope for eMERGE III, upon further consultation with
520 the site, this finding was included in the clinical report of the individual to encourage additional testing and/or
521 management.

522
523 A second case with unexpected findings was associated with another copy number variant call. A duplication for
524 all exons of the *OTC* and *GLA* genes, confirmed by ddPCR, was observed in a 40-year-old male not selected
525 for phenotype. These genes are the only two present on the X chromosome on the eMERGEseq panel. Given
526 that *OTC* and *GLA* are on the p and q arms respectively, the observed duplication is most likely a single event
527 spanning the entire X chromosome. This is most consistent with a male with Klinefelter syndrome (47,XXY).
528 Additional clinical information provided by the site confirmed a prior diagnosis of Klinefelter syndrome that had
529 been confirmed by chromosomal karyotyping. Although a clinical report was not issued for this individual, these
530 findings serve to further validate the sensitivity of NGS-based copy number calling.

531
532 The third unexpected category of findings was that six individuals presented with apparently mosaic variants in
533 genes that predispose to cancer or cardiomyopathy (*TP53*, *CHEK2*, *ATM*, *MYH7*). The presence of mosaics was
534 based upon the ascertainment of allelic variants that were present in <30% of the DNA sequence reads at the
535 variant site. Initial observations were screened manually to eliminate false positives due to mis-mapping to
536 pseudogene sites or other technical errors. The presence of the mosaic variants was subsequently confirmed
537 by Sanger sequencing and clinical reporting offered to the referring sites.

538 539 540 **DISCUSSION:**

541 The introduction of clinical sequencing into the phase III of the eMERGE network has provided a framework for
542 large-scale clinical translation of genomic data in healthcare, as well as for the seamless integration of research
543 studies into clinical data management. The network integrated a large number of research groups with diverse
544 interests, and a common mission to deliver genomic health care. To stimulate and address challenges for the

545 delivery of genomic medicine, a large number of samples were tested and state of the art methods for
546 interpretation and data delivery were applied.

547
548 A primary driver for the study design was cost and therefore a gene-panel was chosen as a primary platform for
549 genomic analyses. Whole exome sequencing was considered. However, while exomes would have offered
550 increased flexibility and saved time in design and testing, the network determined that a more focused target of
551 ~100 genes was needed to stay within the budget for testing all 25,000 participants. In addition, sites individually
552 contributed research data on subjects using high density genotyping arrays allowing for genome-wide
553 association studies which are not discussed here.

554
555 At the outset, the predictions were made as to the major challenges that would be faced and the most likely
556 obstacles to achieving a smooth flow of clinical results, while maintaining access to research data. However,
557 most of the challenges were not anticipated. For example, one challenge was the variety of different consents
558 used to support the process as each site had a unique consent form and approach for their biobank and these
559 consents sometimes stipulated requirements inconsistent with the network-wide decisions being made. As each
560 site's sequencing got started, these types of site-specific challenges were uncovered. Many sites altered their
561 decisions around the reportable content and details of their reporting needs (e.g. which genes were reportable;
562 whether negative reports were needed; whether reports should contain certain recommendations for genetic
563 counseling, etc). There was evolving work around how to structure pharmacogenomic results to flow into EHRs
564 and work to ensure the accurate provision of phenotypes from the sites to the SCs. One site needed
565 accommodation for lower DNA input. These startup 'hiccups' led to significant delays in getting each site started
566 with their sequencing and clinical reports. However, once a smooth workflow was developed for each site, the
567 SCs were able to ramp up the rate of sequencing, interpretation and reporting. For example, during the first half
568 of the project 5713 cases were completed, versus 9525 cases completed during the second half.

572 CONCLUSIONS

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

580 581 **Consortia:**

582 The full list of members of the eMERGE consortium along with their affiliations and declarations of interests can
583 be found in Table S5 (to be submitted in an upcoming revision).

584
585 Debbie Abrams, Samuel Adunyah, Majid Afshar, David Albers, Ladia Albertson-Junkans, Jen Albrecht, Darren
586 Ames, Armand Antommaria, Paul Appelbaum, Krishna Aragam, Sandy Aronson, Sharon Aufox, Larry Babb,
587 Adithya Balasubramanian, Shawn Banta, Melissa Basford, Joan Bathon, Christopher Bauer, Samantha Baxter,
588 Meckenzie Behr, Barbara Benoit, Ashwini Bhat, Elizabeth Bhoj, Sue Bielinski, Sarah Bland, Paula Blasi, Carrie
589 Blout, Eric Boerwinkle, Scott Bolesta, Kenneth Borthwick, Erwin Bottinger, Deb Bowen, Mark Bowser, Harrison
590 Brand, Carmen Radecki Breitkopf, Murray Brilliant, Wendy Brodeur, Kevin Bruce, Adam Buchanan, Andrew
591 Cagan, Pedro Caraballo, David Carey, David Carrell, Andrew Carroll, Robert Carroll, Peter Castaldi, Berta
592 Almoguera Castillo, Lisa Castillo, Victor Castro, Bridget Chak, Gauthami Chandanavelli, Chia Yen Chen,
593 Theodore Chiang, Rex Chisholm, Ken Christensen, Wendy Chung, Chris Chute, Brittany City, Ellen Clayton,
594 Beth Cobb, Francis Cocjin, John Connolly, Nancy Cox, Paul Crane, Katherine Crew, David Crosslin, Damien
595 Croteau-Chonka, Renata Pellegrino da Silva, Mariza De Andrade, Jessica De La Cruz, Emma Davenport,
596 Colleen Davis, Dan Davis, Lea Davis, Matt Deardorff, Josh Denny, Shawn Denson, Tim Desmet, Parimala Devi,
597 Keyue Ding, Michael Dinsmore, Sheila Dodge, Qunfeng Dong, Elizabeth Duffy, Phil Dunlea, Todd Edwards,
598 Digna Velez Edwards, Mitchell Elkind, Christine Eng, Angelica Espinoza, Xiao Fan, Anna Farrell, David Fasel,

599 Alex Fedotov, Qiping Feng, Joseph Finkelstein, Mark Fleharty, Chamith Fonseka, Robyn Fossey, Andrea Foster,
600 Robert Freimuth, Christopher Friedrich, Tanya Froehlich, Malia Fullerton, Lucinda Fulton, Birgit Funke, Stacey
601 Gabriel, Vivian Gainer, Carlos Gallego, Nanibaa' Garrison, Tian Ge, Ali Gharavi, Richard Gibbs, Andrew Glazer,
602 Joe Glessner, Jessica Goehringer, Adam Gordon, Chet Graham, Robert Green, Justin Gundelach, Jyoti Gupta,
603 Hakon Hakonarson, Chris Hale, Taryn Hall, Maegan Harden, John Harley, Margaret Harr, Andrea Hartzler, Ali
604 Hasnie, Geoff Hayes, Scott Hebring, Nora Henrikson, Tim Herr, Andrew Hershey, Christin Hoell, Ingrid Holm,
605 Kayla Howell, George Hripcsak, Alexander Hsieh, Jianhong Hu, John Hutton, Jodell Linder, Gail P. Jarvik, Joy
606 Jayaseelan, Yunyun Jiang, Darren Johnson, Laney Jones, Sarah Jones, Yoonie Joo, Sheethal Jose, Navya
607 Shilpa Josyula, Hayan Jouni, Ann Justice, Sara Kalla, Divya Kalra, Elizabeth Karlson, Sekar Kathiresan, Dave
608 Kaufman, Kenneth Kaufman, Melissa Kelly, Eimear Kenny, Dustin Key, Abel Kho, Les Kirchner, Krzysztof
609 Kirylyuk, Terrie Kitchner, Barbara Klanderma, Derek Klarin, Eric Klee, Rachel Knevel, Dennis Ko, David Kochan,
610 Barbara Koenig, Viktoriya Korchina, Leah Kottyan, Christie Kovar, Joel Krier, Emily Kudalkar, Rita Kukafka, Erin
611 Kulick, Iftikhar Kullo, Philip Lammers, Eric Larson, Jennifer Layden, Joe Leader, Matthew Lebo, Magalie Leduc,
612 Mike Lee, Yvonne Lee, Niall Lennon, Kathy Leppig, Nancy D. Leslie, Bruce Levy, Matthew Lewis, Dingcheng
613 Li, Rongling Li, Wayne Liang, Chiao-Feng Lin, Noralane Lindor, Todd Lingren, James Linneman, Hongfang Liu,
614 Wen Liu, Xiuping Liu, Yuan Luo, John Lynch, Hayley Lyon, Daniel Macarthur, Alyssa Macbeth, Harshad
615 Mahadeshwar, Lisa Mahanta, Brad Malin, Vishnu Mallipeddi, Teri Manolio, Brandy Mapes, Maddalena Marasa,
616 Talar Markossian, Keith Marsolo, Jen McCormick, Michelle McGowan, Elizabeth McNally, Ana Mejia, Jim
617 Meldrim, Kelly Melissa, Frank Mentch, Jonathan Mosley, Shubhabrata Mukherjee, Tom Mullen, Jesus Muniz,
618 David Raul Murdock, Shawn Murphy, John Murray, Mullai Murugan, Donna Muzny, Melanie F. Myers, Bahram
619 Namjou, Pradeep Natarajan, Yizhao Ni, William Nichols, Nahal Nikroo, Laurie Novack, Aniwaa Owusu Obeng,
620 Adelaide Arruda Olson, Janet Olson, Robb Onofrio, Casey Overby-Taylor, Jen Pacheco, Brian Palazzo, Melody
621 Palmer, Jyoti Pathak, Peggy Peissig, Sarah Pendergrass, Kelly Perry, Nate Person, Josh F. Peterson, Lynn
622 Petukhova, Sandie Pisieczko, Kelly Pittman, Siddharth Pratap, Cynthia A. Prows, Rebecca Pulk, Alanna Kulchak
623 Rahm, Ratika Raj, James Ralston, Arvind Ramaprasan, Andrea Ramirez, Luke Rasmussen, Laura Rasmussen-
624 Torvik, Hila Milo Rasouly, Soumya Raychaudhuri, Heidi Rehm, Marylyn Ritchie, Catherine Rives, Beenish Riza,
625 Jamie R. Robinson, Dan Roden, Elisabeth Rosenthal, Jason Ross, Megan Roy-Puckelwartz, Janey Russell,

626 Mert Sabuncu, Senthikumar Sadhasivam, Mayya Safarova, William Salerno, Saskia Sanderson, Simone Sanna-
627 Cherchi, Avni Santani, Dan Schaid, Steven Scherer, Cloann Schultz, Rachel Schwiter, Stuart Scott, Aaron Scrol,
628 Soumitra Sengupta, Nilay Shah, Catherine Shain, Ning 'Sunny' Shang, Himanshu Sharma, Richard Sharp,
629 Yufeng Shen, Ben Moore Shoemaker, Patrick Sleiman, Kara Slowik, Josh Smit, Maureen Smith, Jordan Smoller,
630 Sara Snipes, Susan Snyder, Jung Son, Peter Speltz, Justin Starren, Paul Steele, Kimberly Strauss, Mary Stroud,
631 Amy Sturm, Jessica Su, Agnes Sundaresan, Michael Talkowski, Peter Tarczy-Hornoch, Stephen Thibodeau,
632 Will Thompson, Lifeng Tian, Kasia Tolwinski, Maurine Tong, Sue Trinidad, Meghna Trivedi, Ellen Tsai, Sara L.
633 Van Driest, Sean Vargas, Matthew Varugheese, David Vawdrey, Lyam Vazquez, David Veenstra, Eric Venner,
634 Miguel Verbitsky, Gina Vicente, Sander Vinks, Carolyn Vitek, Michael Wagner, Kimberly Walker, Nephi Walton,
635 Theresa Walunas, Bridget Wang, Kai Wang, Liuyang Wang, Llwen Wang, Qiaoyan Wang, Julia Wattacheril,
636 Firas Wehbe, Wei-Qi Wei, Scott Weiss, Georgia L. Wiesner, Quinn Wells, Chunhua Weng, Peter White, Cathy
637 Wicklund, Ken Wiley, Janet Williams, Marc S. Williams, Mike Wilson, Robert Winchester, Erin Winkler, Leora
638 Witkowski, Betty Woolf, Eric Wright, Tsung-Jung Wu, Julia Wynn, Yaping Yang, Zi (Carol) Ye, Meliha Yetisgen,
639 Zachary Yoneda, Ge Zhang, Kejian Zhang, Lan Zhang, Hana Zouk.

640

641 **Acknowledgements**

642 **Authors' contributions**

643

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

645

646 **Test Design, Validation, Metric Tracking, CNV detection:** Niall Lennon*, Kimberly Walker*, Donna Muzny,

647 Adam Gordon, Mark Bowser, Maegan Harden, Theodore Chiang, Elizabeth Duffy, Jianhong Hu, Matthew

648 Lebo, Alyssa Macbeth, Lisa Mahanta, Eric Venner, Tsung-Jung Wu, Gail P. Jarvik, Hana Zouk, Heidi Rehm,

649 Richard A. Gibbs, Birgit Funke#, Donna Muzny#

650

651 **Validity and Actionability Assessment for Genes and SNPs:** Hana Zouk, Magalie Leduc, Emily Kudalkar,

652 Adam Gordon, Clinical Annotation WG, Eric Venner, Heidi Rehm, Gail P. Jarvik, Birgit Funke

653

654 **Data Intake and Delivery:** Larry Babb, Mullai Murugan, EHRI WG, Darren Ames, Will Salerno, Maegan

655 Harden, Chet Graham, Lisa Mahanta, Matthew Lebo, Heidi Rehm, Richard A. Gibbs, Sandy Aronson, Eric

656 Venner

657

658 **Pharmacogenomics Reporting:** Barbara Klanderma, Hana Zouk, Eric Venner, Elizabeth Duffy, Chiao-Feng

659 Lin, Chet Graham, Lisa Mahanta, Donna Muzny, Rebecca Pulk, Steven Scherer, Matthew Lebo, Richard A.

660 Gibbs, Birgit Funke, Magalie Leduc

661

662 **Variant Interpretation Harmonization and Result Interpretation:** Yunyun Jiang*, Leora Witkowski*, Yaping

663 Yang, Christine Eng, Matthew Varugheese, Eric Venner, Clinical Annotation WG, Birgit Funke, Richard A.

664 Gibbs, Heidi Rehm, Magalie Leduc#, Hana Zouk#

665

666 **Clinical Site Representatives for Return of Results:** Armand Antommara (CCHMC), Nancy D. Leslie

667 (CCHMC), Melanie F. Myers (CCHMC), Cynthia A. Prows (CCHMC), Wendy Chung (Columbia), David Fasel

668 (Columbia), Hila Rasouly Milo (Columbia), Chunhua Weng (Columbia), Scott Bolesta (Geisinger), Melissa Kelly

669 (Geisinger), Rebecca Pulk (Geisinger), Marc S. Williams (Geisinger), Gail P. Jarvik (UW), Eric Larson (KPW),

670 Kathleen Leppig (KPW), James Ralston (KPW), David Kochan (Mayo), Iftikhar Kullo (Mayo), Noralene Lindor

671 (Mayo), Erin Winkler (Mayo), Christin Hoell (Northwestern), Laura J. Rasmussen-Torvik (Northwestern),

672 Maureen E. Smith (Northwestern), Robert C. Green (Partners), Jordan W. Smoller (Partners), Josh F.

673 Peterson (VUMC), Jamie R. Robinson (VUMC), Ben Shoemaker (VUMC), Sara L. Van Driest (VUMC), Quinn

674 Wells (VUMC), Georgia L. Wiesner (VUMC).

675

676 **BCM-Human Genome Sequencing Center:** Donna Muzny, Eric Venner, Jianhong Hu, Kimberly Walker, Sara

677 Kalla, Theodore Chiang, Tsung-Jung Wu, Ritika Raj, Andrea Foster, Adithya Balasubramanian, Jesus Muniz,

678 Shawn Denson, Gauthami Chandanavelli, Wen Liu, Harshad Mahadeshwar, Kimberly Strauss, Sean Vargas,

679 Lan Zhang, Xiuping Liu, Qiaoyan Wang, Joy Jayaseelan, Darren Ames, Divya Kalra, Beenish Riza, Jessica De

680 La Cruz, Brian Palazzo, Liwen Wang, William Salerno, Viktoriya Korchina, Christie Kovar, Yunyun Jiang,
681 Magalie Leduc, David Raul Murdock, Eric Boerwinkle, Victoria Yi, Yaping Yang, Mullai Murugan, Christine Eng,
682 Richard A. Gibbs

683

684 **Partners Laboratory for Molecular Medicine/Broad Institute Sequencing Center:** Hana Zouk, Maegan
685 Harden, Leora Witkowski, Samuel Aronson, Larry Babb, Samantha Baxter, Mark Bowser, Wendy Brodeur,
686 Sheila Dodge, Phil Dunlea, Christopher Friedrich, Tim Desmet, Michael Dinsmore, Mark Fleharty, Chet
687 Graham, Elizabeth D. Hynes, Emily Kudalkar, Matthew Lebo, Chiao-Feng Lin, Hayley Lyon, Alyssa Macbeth,
688 Lisa Mahanta, Jim Meldrim, Tom Mullen, Robb Onofrio, Kara Slowik, Gina Vicente, Mike Wilson, Betty Woolf,
689 Stacey Gabriel, Birgit Funke[#], Niall Lennon[#], Heidi L. Rehm[#]

690

691

692 **Coordinating Center:** Melissa Basford, David Crosslin, Adam Gordon, Kayla Howell, Gail P. Jarvik, Jodell
693 Linder (Jackson), Ian Stanaway, Josh Peterson (PI)

694

695 **eMERGE Working Group Co-Chairs:** Clinical Annotation: Gail P. Jarvik and Heidi Rehm; EHR Integration:
696 Sandy Aronson and Casey Overby-Taylor; Genomics: David Crosslin, Megan Roy-Puckelwartz and Patrick
697 Sleiman; Outcomes: Hakon Hakonarson, Josh Peterson and Marc S. Williams; PGx: Cynthia A. Prows and
698 Laura Rasmussen-Torvik; Phenotyping: George Hripcsak and Peggy Peissig; Return of Results/ELSI: Ingrid
699 Holm and Iftikhar Kullo.

700

701 **eMERGE Principal Investigators:** Rex Chisholm (Steering Committee Chair), Samuel Adunyah, David
702 Crosslin, Josh Denny, Ali Gharavi, Richard Gibbs, Hakon Hakonarson, John Harley, George Hripcsak, Gail P.
703 Jarvik, Elizabeth Karlson, Iftikhar Kullo, Philip Lammers, Eric Larson, Niall Lennon, Shawn Murphy, Josh
704 Peterson, Heidi Rehm, Dan Roden, Marylyn Ritchie, Richard Sharp, Maureen Smith, Jordan Smoller, Stephen
705 Thibodeau, Chunhua Weng, Scott Weiss, Marc S. Williams

706

707 **NHGRI Staff:** Rongling Li (Program Director), Ken Wiley (Program Director), Jyoti Dayal, Sheethal Jose, Teri
708 Manolio (Division Director)

709

710 *##*contributed equally

711

712 **DECLARATION OF INTERESTS**

713 See Table S5 (to be submitted in an upcoming revision)

714

715 **Ethics approval and consent to participate:** All 10 sample collection sites consented participants under
716 Institutional Review Board-approved protocols and the two sequencing centers had IRB-approved protocols that
717 deferred consent to the participating sites. Protocol numbers are as follows: Partners Healthcare (2015P000929),
718 Baylor College of Medicine (#H-40455).

719

720 **Consent for publication:** Not applicable

721

722 **Availability of data and material:** The datasets generated and/or analysed during the current study will be
723 publicly available in the dbGaP repository under [phs001616.v1.p1](#) and pre-dbGaP submission access can also
724 be requested on the eMERGE website <https://emerge.mc.vanderbilt.edu/collaborate/>

725

726 **Funding:** The eMERGE Phase III Network was initiated and funded by NHGRI through the following grants:
727 U01HG8657 (Kaiser Permanente Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672
728 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center);
729 U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences);
730 U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); MD007593 (Meharry
731 Medical College); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center);
732 U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine)

733

734 REFERENCES

- 735 1. National Human Genome Research Institute (2017). Electronic Medical Records and Genomics (eMERGE)
736 Network. <https://www.genome.gov/27540473/electronic-medical-records-and-genomics-emerge-network/>.
- 737 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.,
738 O'Daniel, J.M., Ormond, K.E., et al. (2013). ACMG recommendations for reporting of incidental findings in
739 clinical exome and genome sequencing. *Genet. Med.* *15*, 565–574.
- 740 3. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel,
741 G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-
742 generation DNA sequencing data. *Nat. Genet.* *43*, 491–498.
- 743 4. Reid, J.G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White,
744 S., Salerno, W., Buhay, C., et al. (2014). Launching genomics into the cloud: deployment of Mercury, a next
745 generation sequence analysis pipeline. *BMC Bioinformatics* *15*, 30.
- 746 5. Pugh, T.J., Amr, S.S., Bowser, M.J., Gowrisankar, S., Hynes, E., Mahanta, L.M., Rehm, H.L., Funke, B., and
747 Lebo, M.S. (2016). VisCap: inference and visualization of germ-line copy-number variants from targeted clinical
748 sequencing data. *Genet. Med.* *18*, 712–719.
- 749 6. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler,
750 D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing
751 next-generation DNA sequencing data. *Genome Res.* *20*, 1297–1303.
- 752 7. Fromer, M., and Purcell, S.M. (2014). Using XHMM Software to Detect Copy Number Variation in Whole-
753 Exome Sequencing Data. *Curr. Protoc. Hum. Genet.* *81*, 7.23.1–21.
- 754 8. Chiang, T., Liu, X., Wu, T.-J., Hu, J., Sedlazeck, F.J., White, S., Schaid, D., de Andrade, M., Jarvik, G.P.,
755 Crosslin, D., et al. (2018). Atlas-CNV: a validated approach to call Single-Exon CNVs in the eMERGESeq gene
756 panel.
- 757 9. Clinical Genome Resource. (2018). Sequence Variant Interpretation Working Group.

- 758 <https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation/>.
- 759 10. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J.,
760 Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the Clinical Validity of Gene-Disease
761 Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum.*
762 *Genet.* *100*, 895–906.
- 763 11. Human Genome Sequencing Center, Baylor College of Medicine (2016). eMERGE Sample Submission
764 Portal. <https://emerge.hgsc.bcm.edu/workflow/sample-submission>.
- 765 12. Broad Institute (2016). Clinical Research Sequencing Platform Home.
766 <https://portals.broadinstitute.org/portal/CRSP>.
- 767 13. Aronson, S.J., Clark, E.H., Babb, L.J., Baxter, S., Farwell, L.M., Funke, B.H., Hernandez, A.L., Joshi, V.A.,
768 Lyon, E., Parthum, A.R., et al. (2011). The GenInsight Suite: a platform to support laboratory and provider use
769 of DNA-based genetic testing. *Hum. Mutat.* *32*, 532–536.
- 770 14. Aronson, S.J., Clark, E.H., Varugheese, M., Baxter, S., Babb, L.J., and Rehm, H.L. (2012). Communicating
771 new knowledge on previously reported genetic variants. *Genet. Med.* *14*, 713–719.
- 772 15. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E.,
773 Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint
774 consensus recommendation of the American College of Medical Genetics and Genomics and the Association
775 for Molecular Pathology. *Genet. Med.* *17*, 405–424.
- 776 16. Fossey, R., Kochan, D., Winkler, E., Pacyna, J.E., Olson, J., Thibodeau, S., Connolly, J.J., Harr, M., Behr,
777 M.A., Prows, C.A., et al. (2018). Ethical Considerations Related to Return of Results from Genomic Medicine
778 Projects: The eMERGE Network (Phase III) Experience. *J Pers Med* *8*, 2.
- 779 17. Aronson, S., Babb, L., Ames, D., Gibbs, R.A., Venner, E., Connelly, J.J., Marsolo, K., Weng, C., Williams,
780 M.S., Hartzler, A.L., et al. (2018). Empowering genomic medicine by establishing critical sequencing result
781 data flows: the eMERGE example. *J. Am. Med. Inform. Assoc.* ocy051–ocy051.

- 782 18. Abbott, G.W., Sesti, F., Splawski, I., Buck, M.E., Lehmann, M.H., Timothy, K.W., Keating, M.T., and
783 Goldstein, S.A. (1999). MiRP1 forms IKr potassium channels with HERG and is associated with cardiac
784 arrhythmia. *Cell* 97, 175–187.
- 785 19. Kapplinger, J.D., Tester, D.J., Salisbury, B.A., Carr, J.L., Harris-Kerr, C., Pollevick, G.D., Wilde, A.A.M.,
786 and Ackerman, M.J. (2009). Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated
787 patients referred for the FAMILION long QT syndrome genetic test. *Heart Rhythm* 6, 1297–1303.
- 788 20. Nawathe, P.A., Kryukova, Y., Oren, R.V., Milanese, R., Clancy, C.E., Lu, J.T., Moss, A.J., Difrancesco, D.,
789 and Robinson, R.B. (2013). An LQTS6 MiRP1 mutation suppresses pacemaker current and is associated with
790 sinus bradycardia. *J. Cardiovasc. Electrophysiol.* 24, 1021–1027.
- 791 21. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray, M.F., Smelser,
792 D.T., Gerhard, G.S., and Ledbetter, D.H. (2016). The Geisinger MyCode community health initiative: an
793 electronic health record-linked biobank for precision medicine research. *Genet. Med.* 18, 906–913.
- 794 22. Clinical Pharmacogenomics Implementation Consortium (2018). CPIC Publications.
795 <https://cpicpgx.org/publications/>.
- 796 23. Michaux, L. (2000). +12 or trisomy 12. *Atlas Genet. Cytogenet. Oncol. Haematol.* 4, 81–82.
- 797 24. English, C.J., Goodship, J.A., Jackson, A., Lowry, M., and Wolstenholme, J. (1994). Trisomy 12 mosaicism
798 in a 7 year old girl with dysmorphic features and normal mental development. *J. Med. Genet.* 31, 253–254.
- 799 25. 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.,
800 Khodr, G.S., Morton, C.C., et al. (1997). Rare trisomy mosaicism diagnosed in amniocytes, involving an
801 autosome other than chromosomes 13, 18, 20, and 21: karyotype/phenotype correlations. *Prenat. Diagn.* 17,
802 201–242.
- 803 26. DeLozier-Blanchet, C.D., Roeder, E., Denis-Arrue, R., Blouin, J.L., Low, J., Fisher, J., Scharnhorst, D., and
804 Curry, C.J. (2000). Trisomy 12 mosaicism confirmed in multiple organs from a liveborn child. *Am. J. Med.*
805 *Genet.* 95, 444–449.

806 27. Chen, C.-P., Chang, S.-D., Su, J.-W., Chen, Y.-T., and Wang, W. (2013). Prenatal diagnosis of mosaic
807 trisomy 12 associated with congenital overgrowth. *Taiwan. J. Obstet. Gynecol.* *52*, 454–456.

808 28. Hong, B., Zunich, J., Openshaw, A., and Toydemir, R.M. (2017). Clinical features of trisomy 12 mosaicism-
809 Report and review. *Am. J. Med. Genet. A* *173*, 1681–1686.

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833 **FIGURE TITLES AND LEGENDS**

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

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

839 a. Technical Harmonization of Two DNA Capture Panels. Coordination and harmonization of all the
840 components of the DNA gene capture panel process at the two sequencing centers.

841
842
843 b. Base Coverage. Percentage of bps covered $\geq 20X$ Across Sequencing Centers. % of bases in the panel
844 targeted region covered in each version of the panel design and the extent to which these bases overlap between
845 the genome centers is shown. Version 2 is the final version used for data generation.

846 **Figure 3: Content Development for the eMERGEseq Panel**

847
848 Left panel: ClinGen gene-disease validity assessment for all site top 6 proposed genes. Those with definite
849 and strong association to disease were considered for further actionability analyses.

850 Middle panel: Clinical assessment for a subset of single nucleotide variants (SNVs). Those deemed
851 Pathogenic/Likely Pathogenic were considered for actionability analyses.

852 Right panel: Final consensus list of returnable content. This included all the ACMG56 genes, in addition to
853 12 genes and 14 variants were deemed actionable by the eMERGE Clinical Annotation Working Group.

854 **Figure 4: Site specific reportable list of genes/SNPs for which Pathogenic or Likely Pathogenic 855 variants will be returned**

856 a. Consensus List. Consensus list of returnable SNPs/Genes with site-specific exclusions indicated with a
857 blue dot

858 b. Site specific List. Non-consensus Genes/SNPs with site-specific inclusions indicated with a green dot

860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886

Figure 5: Variant harmonization process overview.

Left panel: 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.

Figure 6: Aggregate findings returned to sites

The positive rate for each category of returnable findings for 15,754 participants from the eMERGESeq Data Freeze 1.0 is shown. For those with an indication for testing, the different indications are depicted (left). Secondary findings from the consensus gene list across the entire Data Freeze 1.0 cohort are broken down per disease area (middle). 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 (right).

887 **TABLES:**

888 **Table 1: Items to be harmonized**

Item	Challenge	Comments
Collection Sites	Sample Type	Agreed to blood*^
	Sample Quality	Minimal quantity specified*
	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 (200x; 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*	Consensus content	68 genes and 14 SNVs
	Site specific genes and SNVs	See Figure 4 and Table S3
	Updates	Variant reclassifications provided
Data Delivery	Physicians Clinical reports	Pdfs, Consumable xml structure; GenInsight;
	Network access to interpreted variants and de-identified reports	GenInsight deidentified 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
--	--------------------------------------	---

889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913

*Exceptions contributed to extended TAT; ^ BCM-HGSC accepted saliva from some sites for a predetermined number of samples.

914 **Table 2: Genes and SNVs on the eMERGEseq Panel**

915 A. List of 109 eMERGE genes, PGx and actionable SNVs

Disease category	Gene‡
Cancer susceptibility and tumor diseases	APC, BLM (rs113993962), <u>BMP1A</u> , BRCA1, BRCA2, <i>CHEK2</i> , MEN1, MLH1, MSH2 (including rs193922376), MSH6, MUTYH, NF2, <u>PALB2</u> , PMS2, <u>POLD1</u> , <u>POLE</u> , PTEN, RB1, RET, SDHAF2, SDHB, SDHC, SDHD, <u>SMAD4</u> , STK11, TP53, TSC1, TSC2, VHL, WT1
Cardiac diseases	ACTA2, ACTC1, <u>ANK2</u> , <i>CACNA1C</i> , DSC2, DSG2, DSP, GLA, <u>KCNE1</u> , KCNH2, <u>KCNJ2</u> , KCNQ1, LMNA, MYBPC3, MYH7, MYL2, MYL3, PKP2, PRKAG2, RYR2, SCN5A, TMEM43, TNNI3, TNNT2, TPM1
Cholesterol and lipid disorders	<i>ANGPTL3</i> , <i>ANGPTL4</i> , <i>APOA5</i> , APOB, <i>APOC3</i> , LDLR, PCSK9, <i>PLTP</i> , <i>SLC25A40</i>
Endocrine disorders	CYP21A2 (rs6467), <u>HNF1A</u> , <u>HNF1B</u> , <i>MC4R</i> , <i>PON1</i>
Connective Tissue disorders	COL3A1, <u>COL5A1</u> , FBN1, MYH11, MYLK, SMAD3, <i>SLC2A10</i> , TGFBR1, TGFBR2
Neuromuscular diseases	<u>CACNA1A</u> , <i>CACNA1B</i> , <i>CACNA1S</i> , RYR1
Inborn errors of Metabolism	ACADM (rs77931234), ALDOB (rs77931234), BCKDHB (rs386834233, rs386834233), FAH (rs80338898), G6PC (rs1801175), CPT2 (rs397509431), <u>OTC</u> , <i>MTHFR</i>
Immunological/Inflammatory disorders	<i>IL-33</i> , <i>IL-4</i> , MEFV (rs28940579, rs61752717), <i>TNF</i> , <i>TYK2</i>
Neurological/Psychiatric disorders	<i>APOE</i> , <i>ATM</i> , <i>ATP1A2</i> , <i>GRM1</i> , <i>GRM2</i> , <i>GRM5</i> , <i>GRM7</i> , <i>GRM8</i> , <i>NTRK1</i> , <i>SC1NA</i> , <i>SCN9A</i> , <i>TTR</i>
Respiratory disorders/hypertension	<i>BPMR2</i> , <i>CFTR</i> , <i>CORIN</i> , <i>SERPINA1</i>
Renal disorders	<i>CFH</i> , <i>UMOD</i>
Skeletal disorders	<i>TCIRG1</i> , <i>VDR</i>
Other	F5 (clotting disorder; rs6025), <i>FLG</i> (<i>dermatological</i>), HFE (iron storage disorder; rs1800562), <i>TCF4</i> (<i>Pitt-Hopkins syndrome</i>), <i>TSLP</i> (<i>association with many complex disorders</i>)
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)

916 ‡ ACMG56 genes are indicated in regular font, Actionable site TOP-6 genes are underlined, non-actionable
 917 TOP-6 genes are in italics, actionable SNVs are indicated by their rs number.
 918

919

920

921

922

923

924
925

B. 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	1551

926
927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

Table 3: Assay performance and optimization at the sequencing sites

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

946

*CNV sensitivity at Broad/LMM is for events ≥3 consecutive exons.

Figure 1.

eMERGE III Clinical and Discovery Platform

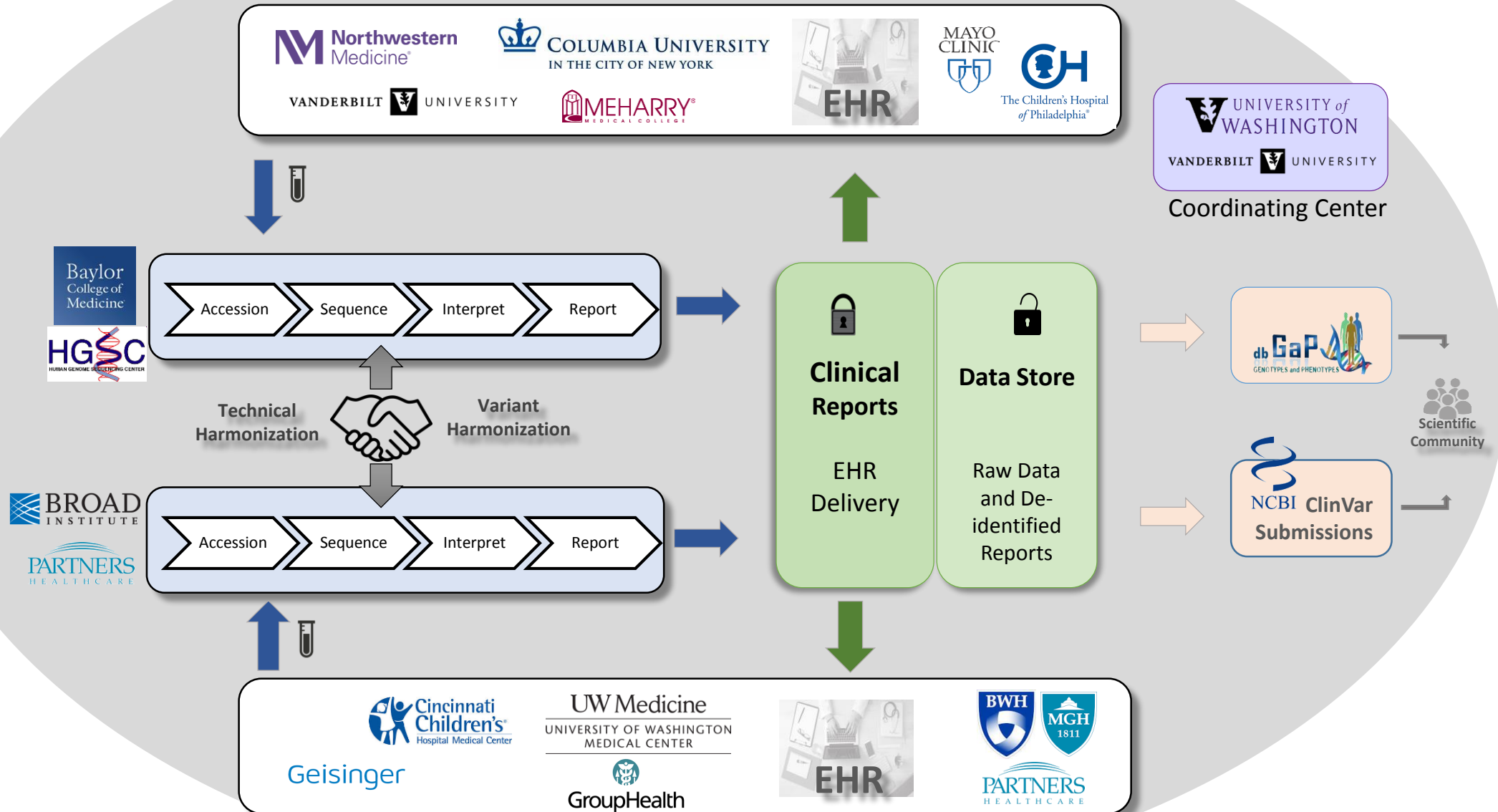
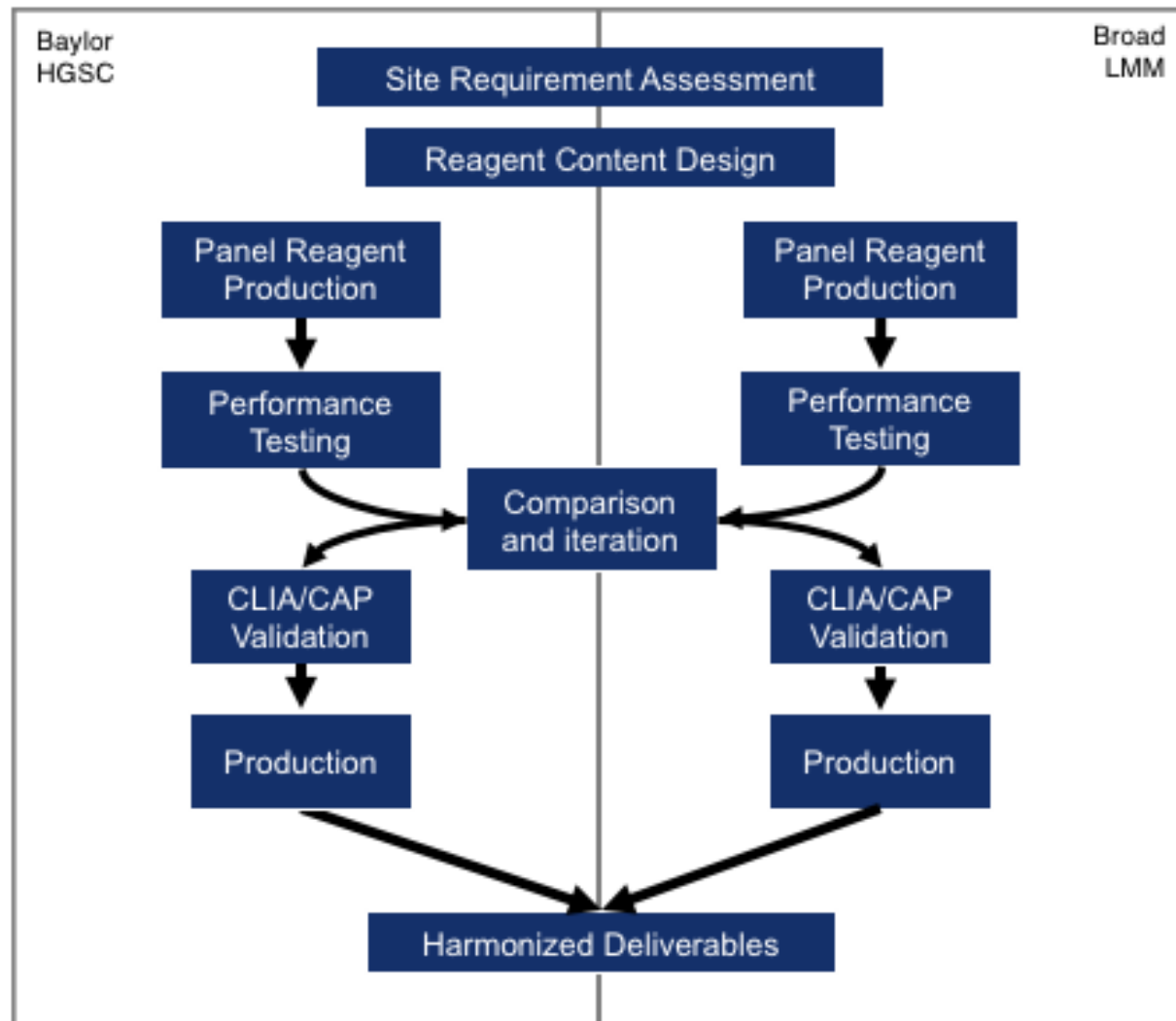


Figure 2.

a.



b.

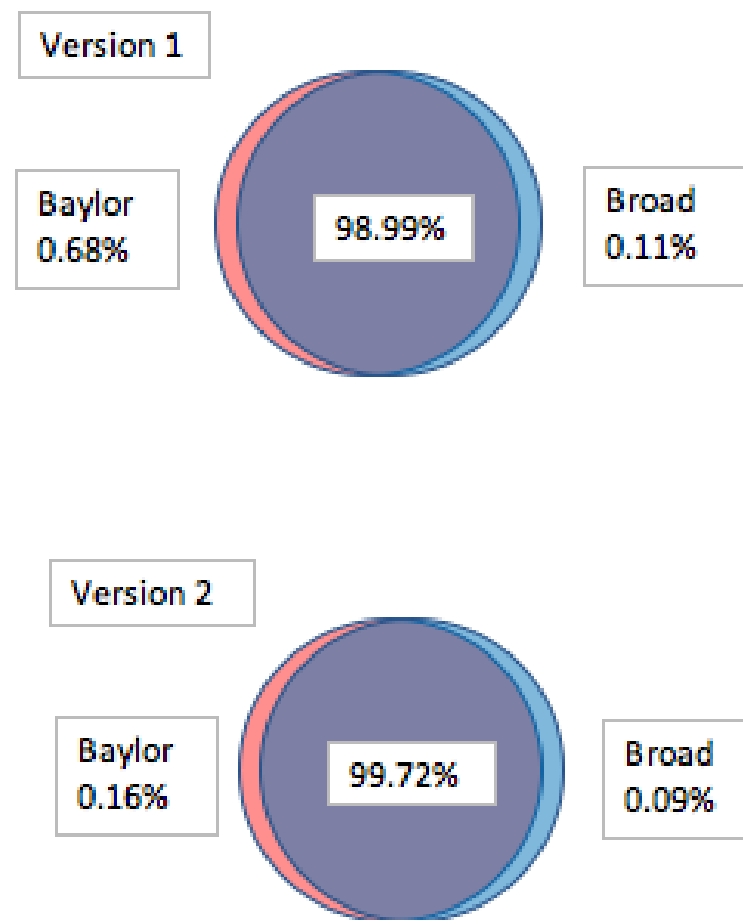


Figure 3.

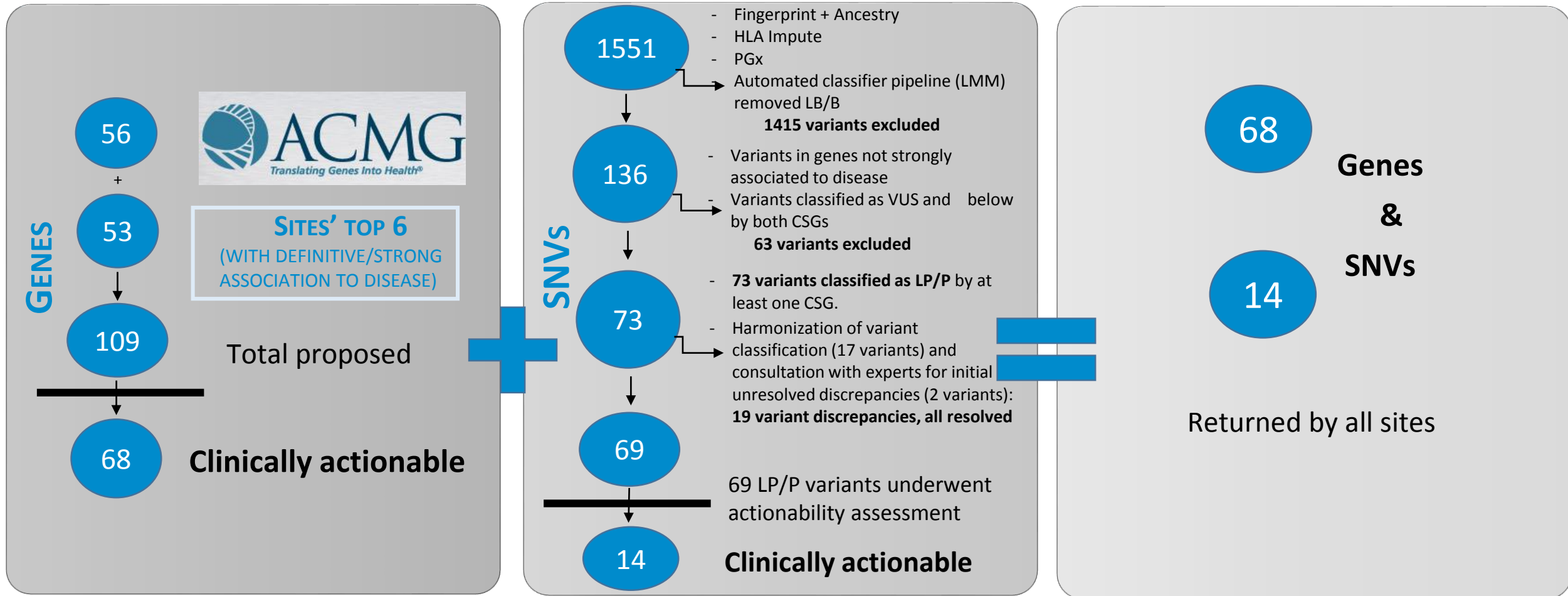
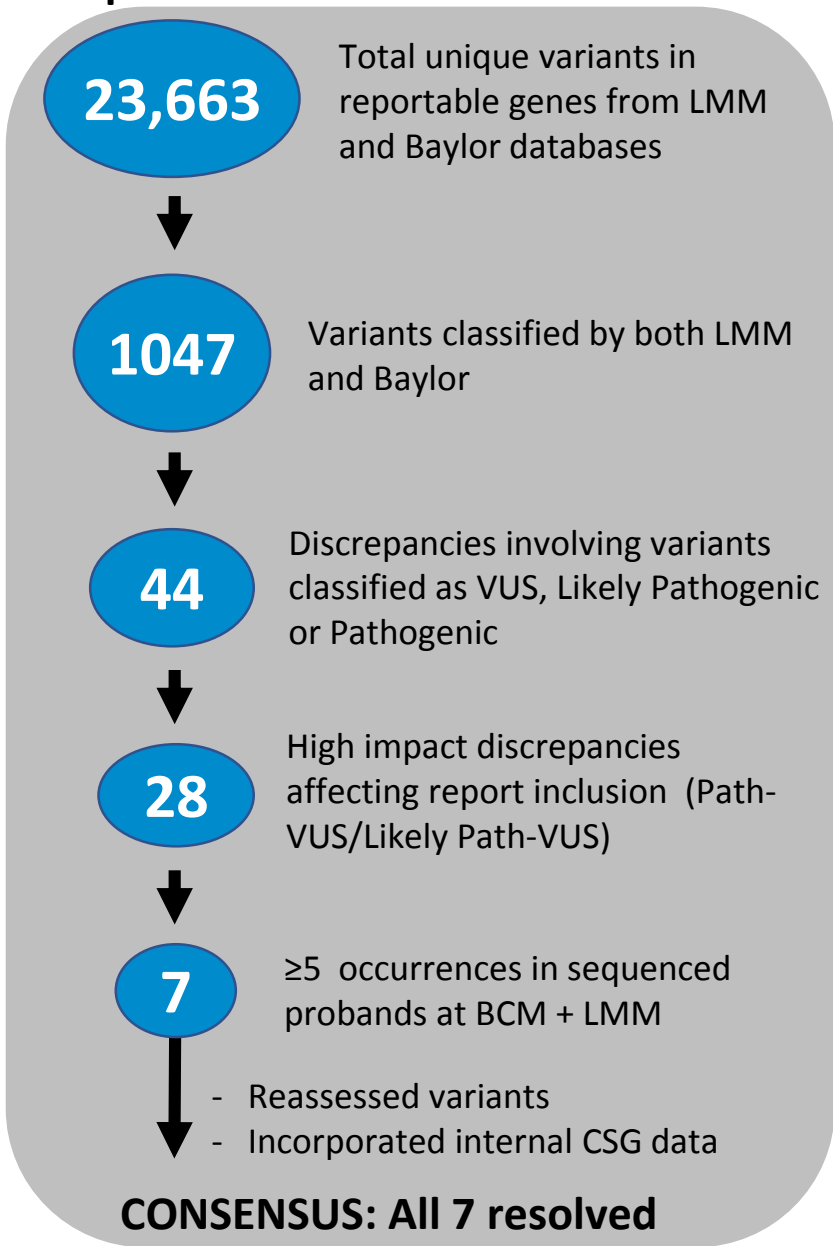
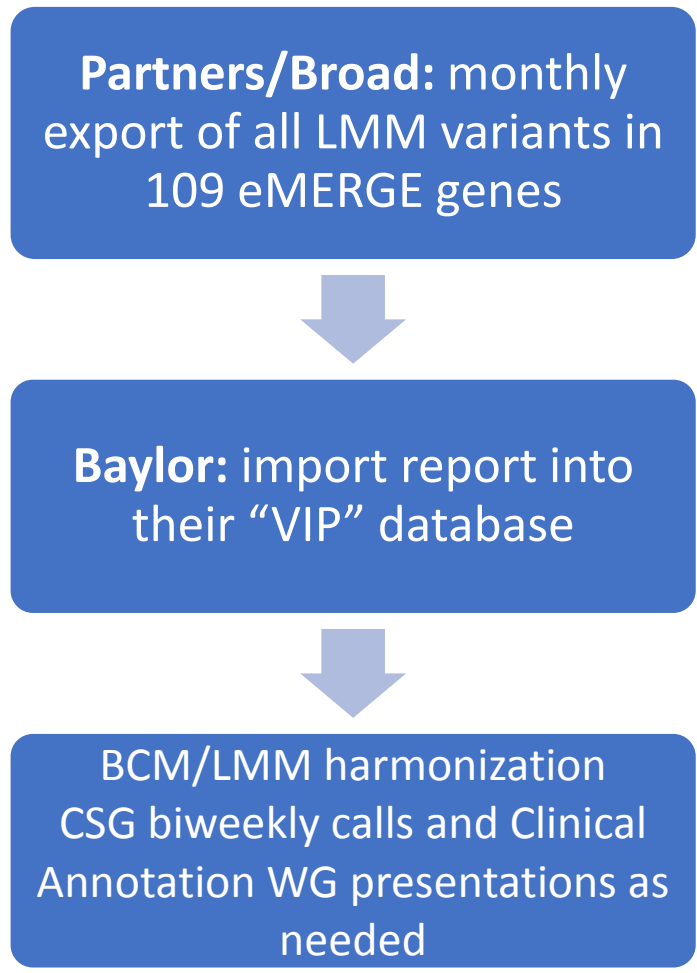


Figure 5.

Pre-launch: CSGs exchanged all previously reported variants



Post-launch harmonization workflow



Post-launch variant discrepancy resolution of high impact variants affecting report inclusion

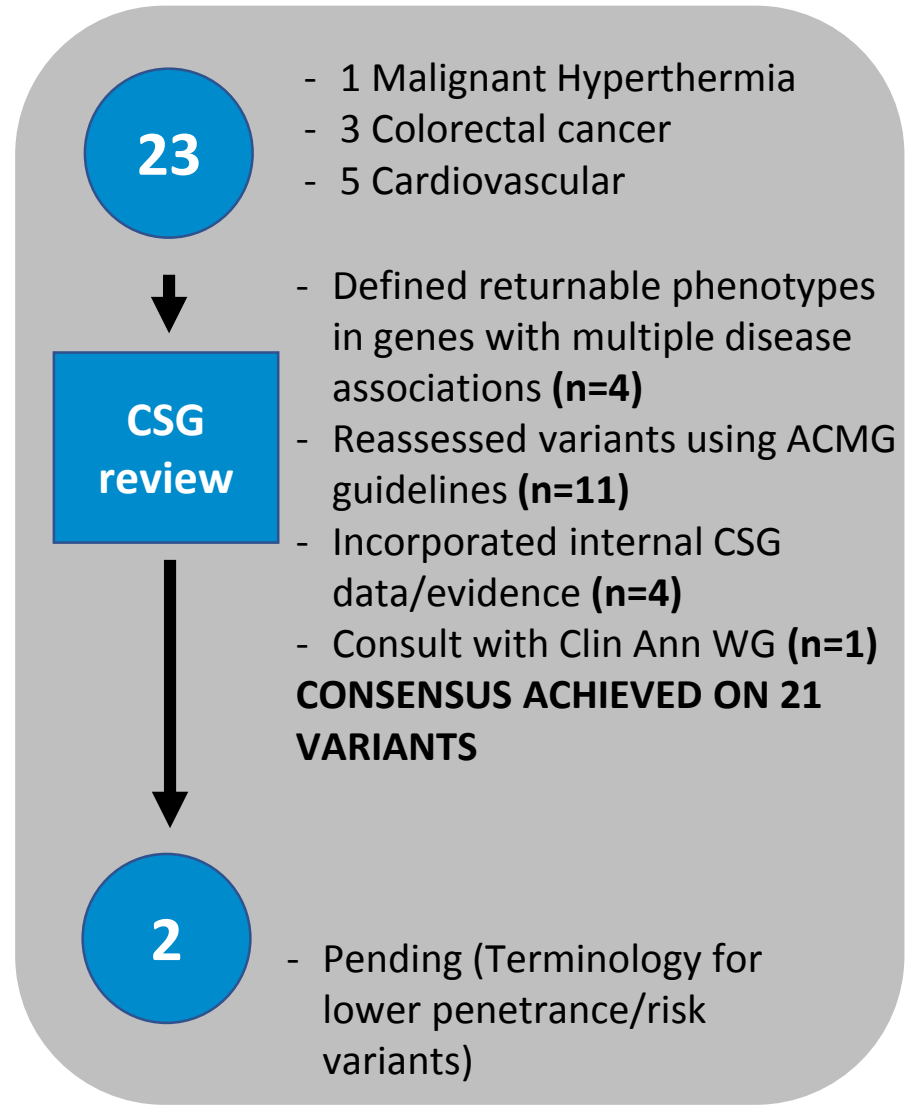
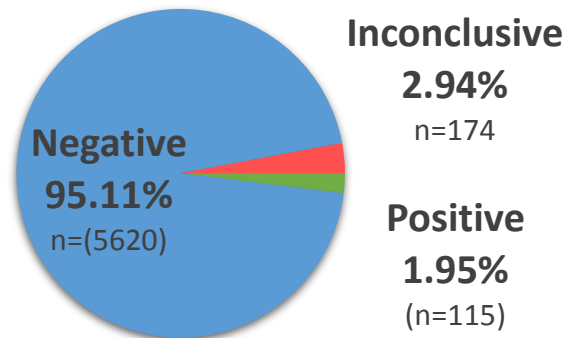


Figure 6.

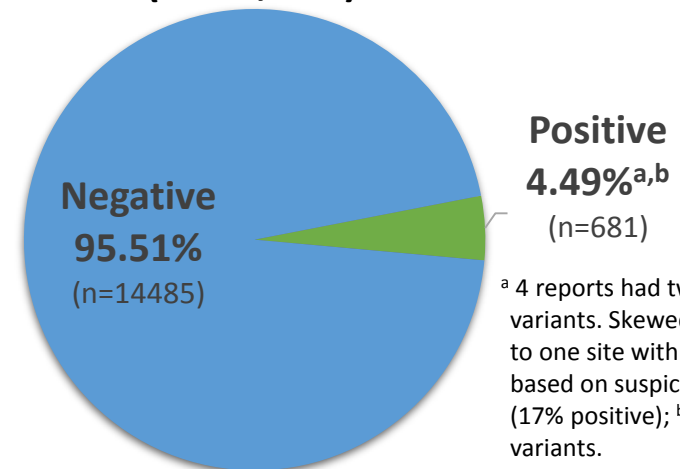
Indication-based returnable results

(n=5,909)



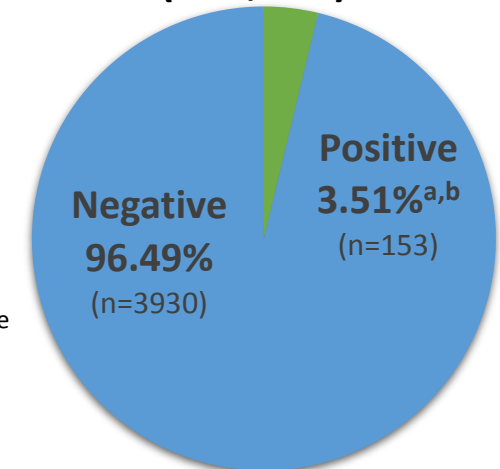
Non indication-based consensus returnable results

(n=15,166)



Non indication-based site-specific returnable results

(n=4,073)



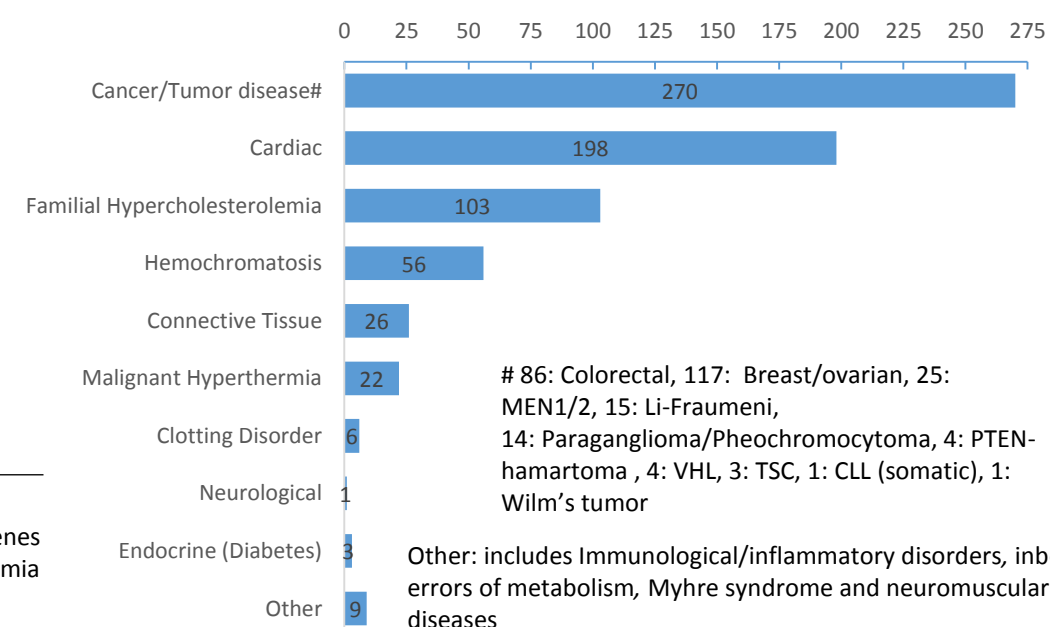
^a 4 reports had two pathogenic variants. Skewed positive rate due to one site with sample selection based on suspicious genotype (17% positive); ^b 8 patients had 2 variants.

^a 5 reported as carrier status
^b 10 patients have two site specific variants or indication/non-indication based variant.

Indication	Total	Positive	Negative	Inconclusive
Colorectal cancer/Polyps*	2358	33	2151	174
Breast/Ovarian Cancer ^a	145	25	120	n/a
Arrhythmia	71	0	71	n/a
Asthma	56	0	56	n/a
Cardiomyopathy	7	1	6	n/a
Chronic Kidney Disease	31	0	31	n/a
Obesity	20	0	20	n/a
Pulmonary Hypertension	17	0	17	n/a
Tubular Sclerosis Complex	5	5	0	n/a
Abnormality of pain sensation	584	0	584	n/a
Autistic Behavior	57	0	57	n/a
Ehlers-Danlos Syndrome	72	1	71	n/a
Hyperlipidemia* ^b	2612	50	2562	n/a
Pediatric Migraine	461	0	461	n/a
TOTAL	5909	115	5620	174

* 1 report had an additional secondary finding; ^aFindings from 68 consensus genes except for two in CHEK2; ^b587 patients had colorectal cancer and hyperlipidemia

Returnable findings per disease area



86: Colorectal, 117: Breast/ovarian, 25: MEN1/2, 15: Li-Fraumeni, 14: Paraganglioma/Pheochromocytoma, 4: PTEN-hamartoma, 4: VHL, 3: TSC, 1: CLL (somatic), 1: Wilm's tumor

Other: includes Immunological/inflammatory disorders, inborn errors of metabolism, Myhre syndrome and neuromuscular diseases

Path and Lpath site-specific variants	Total
<i>CHEK2</i>	51
<i>ATM</i>	22
<i>SERPINA1</i> [^]	13
<i>MC4R</i>	3
<i>F11, KCNE2, BMPR2, JAK2</i>	7
<i>KCNE1 (risk allele)</i>	54
<i>CFTR</i> [^]	3
Total	153

[^] reported as carrier status