BMC Research Notes

RESEARCH NOTE

Open Access

Genetic sex validation for sample tracking in next-generation sequencing clinical testing

Jianhong Hu[1], Viktoriya Korchina[1], Hana Zouk[3,4], Maegan V. Harden[5], David Murdock[1,2], Alyssa Macbeth[5], Steven M. Harrison[3,5], Niall Lennon[5], Christie Kovar[1], Adithya Balasubramanian[1], Lan Zhang[1], Gauthami Chandanavelli[1], Divya Pasham[1], Robb Rowley[6], Ken Wiley[6], Maureen E. Smith[7], Adam Gordon[7], Gail P. Jarvik[8], Patrick Sleiman[9], Melissa A. Kelly[10], Harris T. Bland[11], Mullai Murugan[1], Eric Venner[1], Eric Boerwinkle[1,12], the eMERGE III consortium, Cynthia Prows[13], Lisa Mahanta[3], Heidi L. Rehm[3,5], Richard A. Gibbs[1] and Donna M. Muzny[1*]

Abstract

Objective Data from DNA genotyping via a 96-SNP panel in a study of 25,015 clinical samples were utilized for quality control and tracking of sample identity in a clinical sequencing network. The study aimed to demonstrate the value of both the precise SNP tracking and the utility of the panel for predicting the sex-by-genotype of the participants, to identify possible sample mix-ups.

Results Precise SNP tracking showed no sample swap errors within the clinical testing laboratories. In contrast, when comparing predicted sex-by-genotype to the provided sex on the test requisition, we identified 110 inconsistencies from 25,015 clinical samples (0.44%), that had occurred during sample collection or accessioning. The genetic sex predictions were confirmed using additional SNP sites in the sequencing data or high-density genotyping arrays. It was determined that discrepancies resulted from clerical errors (49.09%), samples from transgender participants (3.64%) and stem cell or bone marrow transplant patients (7.27%) along with undetermined sample mix-ups (40%) for which sample swaps occurred prior to arrival at genome centers, however the exact cause of the events at the sampling sites resulting in the mix-ups were not able to be determined.

Keywords Next-generation sequencing (NGS), Clinical testing, Sex concordance, SNP genotyping

*Correspondence:
Donna M. Muzny
donnam@bcm.edu
[1] Baylor College of Medicine, Human Genome Sequencing Center (HGSC), Houston, TX, USA
[2] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
[3] Laboratory for Molecular Medicine (LMM), Mass General Brigham, Cambridge, MA, USA
[4] Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[5] Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6] Division of Genomic Medicine, National Human Genome Research Institute, Bethesda, MD, USA
[7] Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
[8] Division of Medical Genetics, Department of Medicine, University of Washington Medical Center, Seattle, WA, USA
[9] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[10] Genomic Medicine Institute, Geisinger, Danville, PA, USA
[11] Vanderbilt University Medical Center, Nashville, TN, USA
[12] Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA
[13] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Introduction

The implementation of next-generation sequencing (NGS) technologies in clinical laboratories [1–4] typically involves three phases: (i) the pre-analytic phase including sample collection, DNA extraction and shipment; (ii) the analytic phase of NGS library preparation, DNA sequencing, bioinformatics analysis; and (iii) a post-analytic phase including clinical report generation and delivery. Each phase is inherently subject to sample tracking and identification errors, with prior reports of more than 46% of errors occurring during the pre-analytical phase, caused by inappropriate test requests, order entry errors, patient misidentification, and labelling errors [5]. Validation and tracking of sample identity therefore is a basic and important aspect of effective clinical NGS testing.

DNA-based methods for sample tracking include genotyping of short tandem repeats (STRs) or single nucleotide polymorphisms (SNPs) [6–8]. STRs are generally located in non-coding regions, prone to high sequencing error rates, and often require longer than typical sequencing read lengths to precisely define the number of repeats, limiting their application. In contrast, SNPs are ubiquitous in the genome and simple to assay [9–11]. In this study, a 96-SNP panel was used to track samples through the clinical NGS workflow in the National Institute of Health's Electronic Medical Records and Genomics Phase III (eMERGE) program [12]. The network linked together 11 sample collection sites and 2 clinical genetic testing laboratories, the Human Genome Sequencing Center Clinical Laboratory at Baylor College of Medicine (BCM-HGSC-CL) and the Mass General Brigham Laboratory for Molecular Medicine (LMM) in partnership with the Clinical Research Sequencing Platform (CRSP) at the Broad Institute of MIT and Harvard. A total of 25,015 clinical DNA samples were processed. The 96-SNP panel-based procedure provided a robust method for sample tracking in the clinical NGS workflow and showed that the testing of sex can provide a valuable quality control tool.

Methods

Fluidigm SNP genotyping assay

Two clinical laboratories harmonized methods for the program [12] and utilized a 96-SNP panel but incorporated different selected SNPs to track samples and determine ancestry. Each 96-SNP panel contained one subset of SNPs on the sex-chromosomes. The autosome SNPs are within the target region of the capture design used in the eMERGE program (Additional files 1, 2) [12]. Assays were performed according to the manufacturer's recommendations.

The BCM-HGSC-CL's 96-SNP panel replaced 19 of the original Fluidigm SNPtrace 96 sites to match genomic regions specifically targeted in eMERGE III. The remaining sites included 3 SNPs on Chromosome X and 3 on Chromosome Y [13, 14]. At the Broad Institute, the chosen SNPs included 95 autosomal SNPs and 1 sex determining assay locus, covering the AMELX and AMELY gene (AMG_3B) with a sex-specific 6 base-pair insertion/deletion.

Illumina Infinium HumanCoreExome SNP array assays and NGS

The HumanCoreExome v1-3 BeadChips contain 500K variant sites, including more than 12,900 located on the X chromosome, that are informative for genetic sex prediction. Infinium SNP array assay were performed with 200 ng genomic DNA according to manufacturer's instructions. DNA sequencing for the eMERGE phase III program has been described previously [12].

Results

As a first step towards assessing sample swaps during the analytic phase in NGS testing, we tested the concordance between data generated from the 96-SNP panel genotyping and the DNA sequence data at each of the two Genome Characterization Centers. The BCM-HGSC-CL and LMM/Broad laboratories utilized the same analytical platform foundation, employing slightly different SNP sites for the assays, but generally similar workflows (Fig. 1). The average SNP call rates were 97.3% and 97.5% for the 25,015 samples processed at the BCM-HGSC-CL and the LMM/Broad, respectively. No sample swaps were identified during the analytic NGS testing phase. Next, we compared the 96-SNP panel genotype-based sex to reported sex at the time of sample accessioning, where a total of 110 (0.44%) non-concordant cases from two testing laboratories were identified. The two testing laboratories utilized slightly different workflows to technically validate the sex discrepancies.

At the BCM-HGSC-CL, of the 14,515 samples processed, 73 samples with sex discrepancies were re-tested with the same 96-SNP panel. Identical results were obtained for 70 of the re-tested samples (Table 1). For the remaining 3 cases, where the sex provided on test requisition was male, non-concordant or ambiguous data were observed between the initial and the repeated assays. For two of these samples, the automated software calls from one of each duplicate assays indicated that the DNA source was from individuals with Klinefelter Syndrome (47, XXY). However, further review of the SNP scatter plots for autosome and sex SNPs indicated that the inconsistent sex calls most likely resulted from sample contamination involving a mixture of male and female

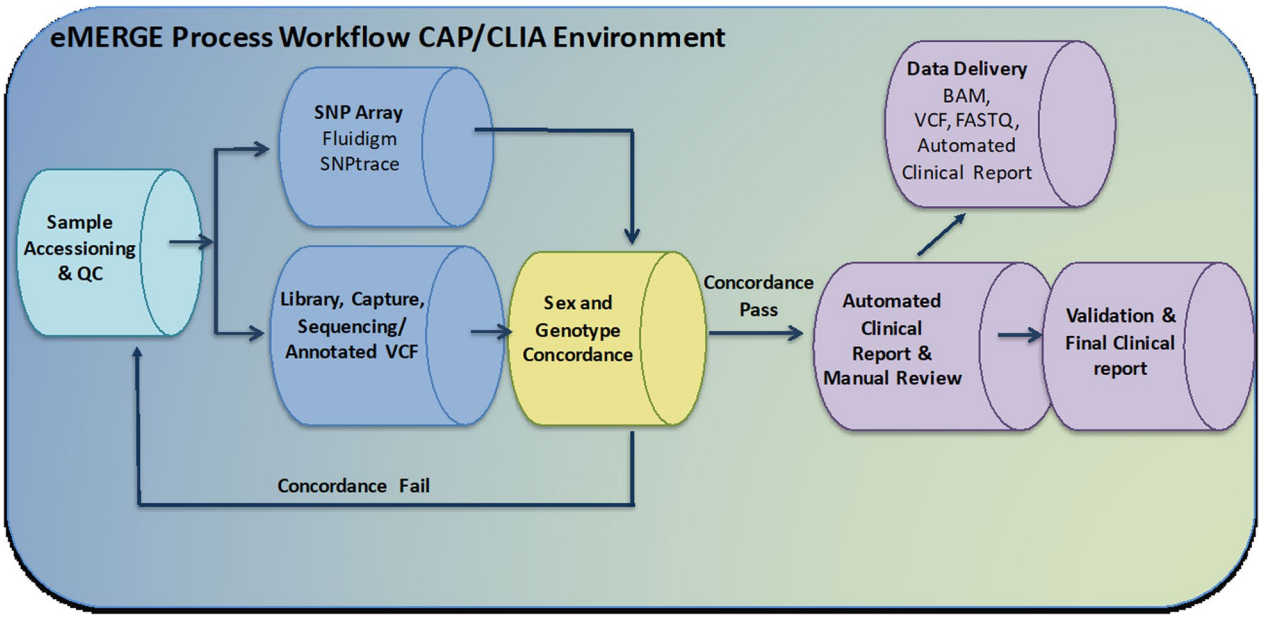

Fig. 1 eMERGE sample processing workflow. Steps indicating where aliquots of DNA are taken from samples that are presented to the Clinical DNA Sequencing Laboratory for accession, to test via the Fluidigm 96-SNP panel assay. Data from the Fluidigm 96-SNP panel assay are compared with DNA sequence data from the DNA sequencing pipeline as a quality control step, ahead of the Automated Clinical Reporting step

DNAs (Fig. 2). The third sample was called as female with lower confidence initially. In the repeated assay, one of the X SNPs failed to call due to localization in between clusters in plot analysis. This is most likely due to the female sample mixed up with some DNA sample from another female.

Next, Illumina HumanCore Exome Arrays were utilized as an orthogonal high-density hybridization genotyping assay to further test 71 of the 73 samples with sex inconsistencies except two samples which had insufficient genomic DNA (Table 1). HumanCore Exome Array results confirmed 96-SNP panel genotyping sex data, including the suspected two contaminated female samples with additional male or other female DNA.

At the Broad/LMM, the reported sex from the test requisition was compared with the genetic sex determined by both the Fludigm genotyping assay and the data from the eMERGE III sequencing panel. Of the 10,500 samples processed, 151 were initially either identified as discordant or had no sex determination. For 95 samples, the Fluidigm assay data could not return a sex determination, however the sequencing sex matched the reported sex for each and no further action was taken. For 19 of the remaining 56 samples, the sequencing and reported sex were concordant, but did not match the genotyping determined sex. Further review of these 19 samples showed that the genotyping assay calls were generally borderline or low confidence calls, suggesting sub-optimal performance of the single sex determining SNP as the reason for the data discrepancy, rather than either a sex reporting error at accession or sample mix-up in the testing laboratory. The remaining 37 samples had highly confident sex determination calls from both the SNP assay and the subsequent DNA sequencing that were concordant, but did not match the site reported sex (Table 1).

Internal tracking showed that none of the 110 confidently identified sex discrepant samples occur within the clinical DNA sequencing laboratories and that most errors were likely introduced prior to shipment of samples. Sampling sites identified handling errors from test requisitions, sample extraction, and sample handling procedures for 54 cases. Forty-six of these had information that was incorrectly or incompletely entered on the test requisitions and were resolved by examination of other records. In 6 other cases, it was determined that incorrect samples had been shipped from the sampling sites to the genome centers. Biological explanations for the discrepant tracking data were identified for an additional 12 cases. In 4 of these 12 cases, further examination of records revealed that the samples were provided by transgender participants. In addition, 8 sex discrepant samples were determined to be from individuals who had received stem cell or bone marrow transplants. Causes of the sample genetic vs. reported sex discrepancy are listed in Table 2.

Where possible, the information on test requisition forms was amended and correct clinical reports were

Table 1 Comparison of genetic sex determined in various assays and reported sex on test requisition

Sequencing site	Total	Sample providing site	Sex on test requisition	Sex from 1st Fluidigm array	Sex from 2nd Fluidigm array	Sex from Illumina array	Sex from sequencing data	Sample number
BCM-HGSC-CL	73	Site 1	Male	Female	Female	Female	–	5
			Female	Male	Male	Male	–	5
		Site 2	Male	Female	Female	Female	–	13
			Male	Female	Klinefelter	Female	–	1
			Female	Male	Male	Male	–	7
		Site 3	Male	Female	Female	Female	–	3
			Female	Male	Male	Male	–	3
			Female	Male	Male	NA[a]	–	1
		Site 4	Male	Female	Female	Female	–	7
			Female	Male	Male	Male	–	9
			Male	Klinefelter	Female	NA[a]	–	1
		Site 5	Male	Female	Female	Female	–	6
			Male	Female	No Call	Female	–	1
			Female	Male	Male	Male	–	4
		Site 6	Male	Female	Female	Female	–	4
			Female	Male	Male	Male	–	3
LMM/broad	37	Site 7	NA[b]	Female	–	–	Female	1
			Female	NA[c]	–	–	Male	1
			Male	Female	–	–	Female	16
			Female	Male	–	–	Male	13
		Site 8	Male	Female	–	–	Female	1
		Site 9	Male	Female	–	–	Female	1
			Female	Male	–	–	Male	2
		Site 10	Male	Female	–	–	Female	1
			Female	Male	–	–	Male	1

[a] Insufficient gDNA for Illumina array

[b] Sex not reported on requisition form

[c] Sex not called in assay; *NA* not available

issued for 45 cases processed at the BCM-HGSC-CL, or the incorrect samples were replaced and re-processed. Twelve cases sequenced at the BCM-HGSC-CL with sample mix-ups due to unknown causes were withdrawn from the study. Similarly, 32 unsolved cases sequenced at LMM/Broad were either withdrawn or remained under investigation.

Discussion

To identify sample swaps during the processing of 25,015 clinical samples in the NIH eMERGE III program, two clinical DNA sequencing laboratories first utilized a Fluidigm-based 96-SNP panel assay to track internal processes. These analyses indicated no sample swaps had occurred in the time interval between sample arrival at the testing laboratories and the delivery of the final DNA sequencing data. In contrast, when the test was expanded to predict the concordance between the self-reported sex of participants at the time of their initial enrollment, with a predicted sex-by-genotype, there were 110 discordant samples. A battery of follow-up tests indicated that these likely arose before the materials were received at the clinical DNA sequencing laboratories. The bases of the sample tracking errors at sample collection sites were determined in 66 of the 110 cases (60%), while leaving the remaining 44 cases unsolved and under investigation. Of these 66 resolved cases, the largest source for the initial discordance occurring in 54 cases (81%) arose from clerical or shipping errors. The remaining 12 cases (18% of the 66 solved) had biological underpinnings that explained the discordant results, as 8 were due to stem cell/bone marrow transplants while 4 were from transgender individuals. Future sample collecting procedures could be modified by including more informative test requisition options to ensure that participants are invited to note these types of events at the time of collection, so that this information is available for quality control.

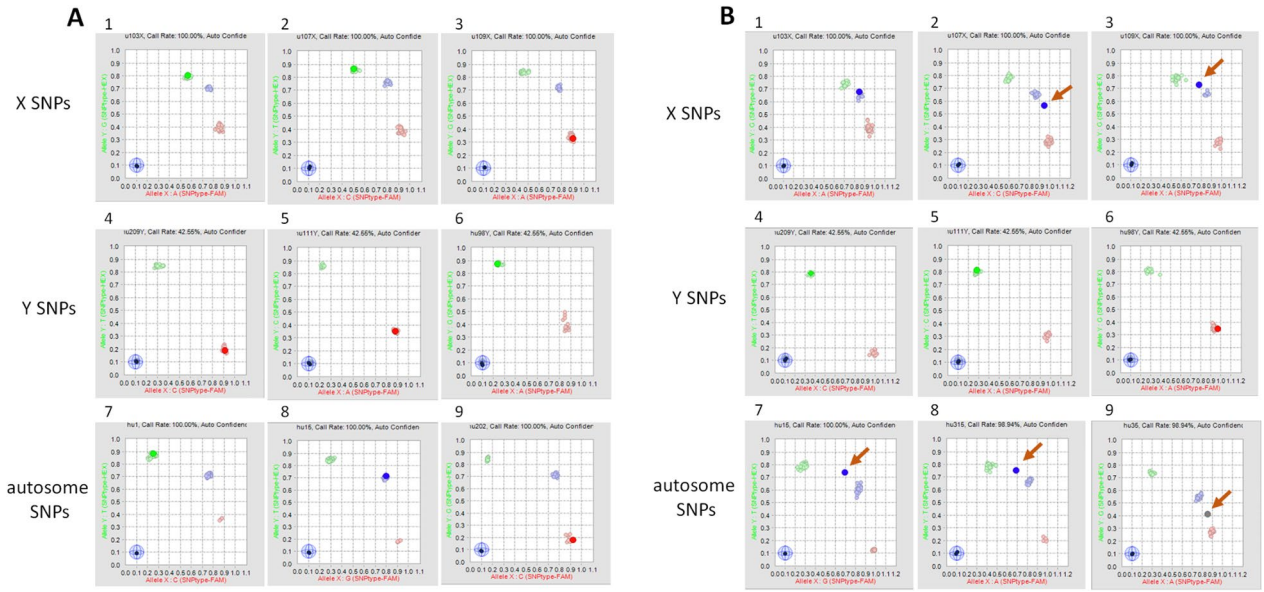

Fig. 2 Scatter plot analysis of 96-SNP panel reveals sample contamination. Scatter plot analysis from vendor software, showing a normal DNA male sample (**A**) or a contaminated sample containing a mixture of male and female DNAs (**B**). Panels 1–3 SNPs on X chromosome; panels 4–6 SNPs on Y chromosome; panels 7–9 autosomal SNPs. Each panel shows the data from a single SNP, as compared to clusters from all other SNPs. Clusters are shown as either homozygous (red or green), or heterozygous (blue) positions. In panels B2, 3, 7–9 single SNPS are represented as outside the expected (arrows) resulting in erroneous or 'no-call' from the software

Table 2 Causes of sample sex discrepancy

Sex discrepant categories	BCM-HGSC-CL samples	LMM/broad samples	Total
Sampling site errors			
Incorrect/incomplete information on test requisition	45	1	46
Error during DNA extraction	0	2	2
Incorrect sample shipped	6	0	6
Transgender	2	2	4
Stem cell/bone marrow transplant recipient	8	0	8
Not solved/under investigation	12	32	44
Total sex discrepancies	**73**	**37**	**110**

The 96-SNP panel has proven value for precise sample tracking [15]. In general, 20 informative SNP loci are sufficient for unique individual sample identification [16, 17]. Other SNP panels have been used for identification of human samples [9, 18, 19]. A low-density QC genotyping array launched by Illumina which includes 15,949 markers has been utilized in genomic-based clinical diagnostics [20]. Our studies showed that these two different SNP platforms exhibited consistent results when applied for sex identification. In comparison to the use of the Illumina Infinium array platform, the workflow for the 96-SNP panel assay is faster (1-day workflow vs 3-day workflow) and more cost-effective (chip price for SNPtrace is about 15% of HumanCoreExome Array per sample). However, the Illumina Infinium array platform provides more information on linkage analysis, HLA haplotyping, ethnicity determination and other genetic information in addition to fingerprinting and thus may be preferred in some scenarios. It may also take into account the sex prediction accuracy of the two methods, the error rate, albeit low, as well as the cost of re-testing that may be necessary in some cases due to low data quality. Other commercial systems are also available to substitute for the platforms described here if they provide cost-effective and precise data with similar qualities.

This level of tracking error is unacceptable for ongoing clinical practice, but the study does not represent the levels that will be expected in further clinical programs.

At least one laboratory declared their initial sample enrollments as 'research samples' and thus committed to later repeat assays under a fully compliant protocol, to verify any findings that may impact care. Others were able to quickly identify points of error and rectify their protocols to ensure faithful future sample handling. All sites committed to rechecking of records and reconciling actionable findings with orthogonal data, including family histories and biochemical tests, before returning results. The 'lessons learned' from these analyses ensure that a repeat of the same program would likely minimize any similar errors.

Limitations

While false positive rates are low for this application of SNP trace, false negative rates will be high. Here, the overall level of genetic and reported sex discordance of 0.44% is likely an underestimate of the true error rate in this study, as the misclassification of genetic sex from a random sample swap would be expected to result in incorrect, erroneous assignment, only 50% of the time. The true ratio may be skewed by factors introducing a sex-bias in the direction of misclassification. This could be caused by skewed phenotypes of individuals with sex chromosome anomalies or that gender obfuscation may be socially driven in an unequal manner, depending on the gender identity of the individual. Overall, the rate is likely higher than the 0.44% identified here, but not anticipated to be higher than twice that level.

Abbreviations

HGSC	Human Genome Sequencing Center
LMM	Laboratory for Molecular Medicine
NGS	Next-generation DNA sequencing
STR	Short tandem repeat
SNP	Single nucleotide polymorphism
PCR	Polymerase chain reaction
eMERGE	Electronic Medical Records and Genomics
EMR	Electronic medical record
HGSC-CL	Human Genome Sequencing Center Clinical Laboratory
BCM	Baylor College of Medicine
CRSP	Clinical Research Sequencing Platform
NHGRI	National Human Genome Research Institute
IRB	Institutional review board

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-024-06723-w.

Additional file 1: Table S1. 96-SNP panel design—BCM-HGSC-CL.

Additional file 2: Table S2. List of 96 SNPs in LMM/broad (CRSP) PCR panel design.

Acknowledgements

We thank all eMERGE Phase III Network participants for their engagement in this research effort.

EMERGE CONSORTIUM: Debra J. Abrams[9], Samuel E. Adunyah[14], Ladia H. Albertson-Junkans[15], Berta Almoguera[9], Paul S. Appelbaum[16,17], Samuel Aronson[3], Sharon Aufox[7], Lawrence J. Babb[5], Adithya Balasubramanian[1], Hana Bangash[18], Melissa A. Basford[19], Meckenzie Behr[9], Barbara Benoit[20], Elizabeth J. Bhoj[9], Sarah T. Bland[11], Eric Boerwinkle[1,12], Kenneth M. Borthwick[21], Erwin P Bottinger[22,23], Deborah J. Bowen[24], Mark Bowser[3], Murray Brilliant[25], Adam H. Buchanan[10], Andrew Cagan[26], Pedro J. Caraballo[27], David J. Carey[28], David S. Carrell[15], Victor M. Castro[26], Gauthami Chandanavelli[1], Rex L. Chisholm[7], Wendy Chung[29], Christopher G. Chute[30], Brittany B. City[19], Ellen Wright Clayton[19,31], Beth L. Cobb[32], John J. Connolly[9], Paul K. Crane[33], Katherine D. Crew[34], David R. Crosslin[35], Renata P. da Silva[9], Jyoti G. Dayal[6], Mariza De Andrade[36], Josh C. Denny[37], Ozan Dikilitas[18], Alanna J. DiVietro[19], Kevin R. Dufendach[38,96], Todd L. Edwards[19,39], Christine Eng[2], David Fasel[40], Alex Fedotov[41], Stephanie M. Fullerton[93], Birgit Funke[42], Stacey Gabriel[5], Vivian S. Gainer[26], Ali Gharavi[40], Richard A. Gibbs[1], Joe T. Glessner[9,43], Jessica M. Goehringer[10], Adam Gordon[7], Adam S. Gordon[7], Chet Graham[3], Heather S. Hain[9], Hakon Hakonarson[9,43], Maegan V. Harden[5], John Harley[44,94], Margaret Harr[9], Steven M. Harrison[3,5], Andrea L. Hartzler[35], Scott Hebbring[25], Jacklyn N. Hellwege[19,45], Nora B. Henrikson[15,46], Christin Hoell[7], Ingrid Holm[47], George Hripcsak[48], Alexander L. Hsieh[48], Jianhong Hu[1], Elizabeth D. Hynes[3], Gail P. Jarvik[8], Darren K. Johnson[10], Laney K. Jones[10], Yoonjung Y. Joo[49], Sheethal Jose[6], Navya Shilpa Josyula[50], Anne E. Justice[50], Elizabeth W. Karlson[51], Kenneth M. Kaufman[32,52], Jacob M. Keaton[19,53], Melissa A. Kelly[10], Eimear E. Kenny[54,55], Dustin L. Key[15], Atlas Khan[56], H. Lester Kirchner[50], Krzysztof Kiryluk[40], Terrie Kitchner[25], Barbara J. Klanderman[3], David C. Kochan[18], Viktoriya Korchina[1], Christie Kovar[1], Emily Kudalkar[3], Benjamin R. Kuhn[57], Iftikhar J. Kullo[18], Philip Lammers[14,58], Eric B. Larson[15,59], Matthew S. Lebo[3,60], Ming Ta Michael Lee[10], Niall Lennon[5], Kathleen A. Leppig[15,61], Chiao-Feng Lin[3], Jodell E. Linder[19], Noralane M. Lindor[62], Todd Lingren[63,64], Cong Liu[48], Yuan Luo[65], John Lynch[66], Alyssa Macbeth[5], Lisa Mahanta[3], Bradley A. Malin[19], Brandy M. Mapes[19], Maddalena Marasa[56], Keith Marsolo[67], Elizabeth McNally[7], Frank D. Mentch[9], Erin M. Miller[64,68], Hila Milo Rasouly[56], David Murdock[1,2], Shawn N. Murphy[69], Mullai Murugan[1], Donna M. Muzny[1], Melanie F. Myers[64,70], Bahram Namjou[71], Addie I. Nesbitt[9], Jordan Nestor[56], Yizhao Ni[63,64], Janet E. Olson[62], Aniwaa Owusu Obeng[72,73], Jennifer A. Pacheco[7], Joel E. Pacyna[74], Divya Pasham[1], Thomas N. Person[10], Josh F. Peterson[19], Lynn Petukhova[75,95], Cassandra Pisieczko[10], Siddharth Pratap[14], Cynthia Prows[13], Megan J. Puckelwartz[7], Alanna K. Rahm[10], James D. Ralston[15,61], Arvind Ramaprasan[15], Luke V. Rasmussen[65], Laura J. Rasmussen-Torvik[7,65], Heidi L. Rehm[3,5], Dan M. Roden[76], Elisabeth A. Rosenthal[77], Robb K. Rowley[6], Maya S. Safarova[18], Avni Santani[9,78], Juliann M. Savatt[10], Daniel J. Schaid[62], Steven Scherer[1], Baergen I. Schultz[6], Aaron Scrol[15], Soumitra Sengupta[48], Gabriel Q. Shaibi[79], Ning Shang[48], Himanshu Sharma[3], Richard R. Sharp[74], Yufeng Shen[48], Rajbir Singh[14], Patrick Sleiman[9], Maureen E. Smith[7], Jordan W. Smoller[80], Duane T. Smoot[14], Ian B. Stanaway[35], Justin Starren[65], Timoethia M. Stone[19], Amy C. Sturm[10], Agnes S. Sundaresan[81], Peter Tarczy-Hornoch[35,82], Casey Overby Taylor[10,83], Lifeng Tian[9], Sara L. Van Driest[84], Matthew Varugheese[3], Lyam Vazquez[9], David L. Veenstra[85,86], Digna R. Velez Edwards[11,87], Eric Venner[1], Miguel Verbitsky[88], Kimberly Walker[1], Nephi Walton[10], Theresa Walunas[49,89], Firas H. Wehbe[65], Wei-Qi Wei[11,19], Scott T. Weiss[90,91], Quinn S. Wells[92], Chunhua Weng[48], Ken L. Wiley Jr.[6], Marc S. Williams[10], Janet Williams[10], Leora Witkowski[3,42], Laura Allison B. Woods[19], Julia Wynn[29], Lan Zhang[1], Yanfei Zhang[10], Hana Zouk[3,4], Jodell Jackson[97**]

[14]Meharry Medical College, Nashville, TN; [15]Kaiser Permanente Washington Health Research Institute, Seattle, WA; [16]Department of Psychiatry, Columbia University, New York, NY; [17]NY State Psychiatric Institute, New York, NY; [18]Department of Cardiovascular Medicine, Mayo Clinic,Rochester, MN; [19]Vanderbilt University Medical Center, Nashville, TN; [20]Research IS and Computing, Laboratory for Molecular Medicine (LMM), Mass General Brigham, Cambridge, MA; [21]Hood Center for Health Research, Geisinger, Danville PA; [22]Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, NY; [23]Division of Nephrology and Hypertension, Department of Medicine; [24]Department of Bioethics and Humanities, School of Medicine, University of Washington, Seattle, WA; [25]Marshfield Clinic Research Institute, Marshfield, WI; [26]Research IS and Computing, Laboratory for Molecular Medicine (LMM), Mass General Brigham, Cambridge, MA; [27]Department of Medicine, Mayo Clinic, Rochester, MN; [28]Molecular and Functional Genomics, Geisinger, Danville PA; [29]Department of Pediatrics, Columbia University Medical Center, New York, NY; [30]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, MD; [31]Center for Biomedical Ethics and Society, Vanderbilt University, Nashville, TN; [32]Cincinnati Children's Hospital Medical Center, Cincinnati, OH; [33]Department

of Medicine, School of Medicine, University of Washington, Seattle, WA; [34]Department of Medicine and Epidemiology, Columbia University, New York, NY; [35]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; [36]Department of Health Science Research, Division of BioStatistics and Informatics, Mayo Clinic, Rochester, MN; [37]All of Us Research Program, National Institutes of Health, Bethesda MD; [38]Divions of Neonatology and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; [39]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [40]Department of Medicine, Columbia University, New York, NY; [41]Irving Institute for Clinical and Translational Research, Columbia University, New York, NY; [42]Harvard Medical School, Boston, MA; [43]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA; [44]Departments of Pediatrics and Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio; [45]Division of Genetic Medicine, Department of Medicine, Vanderbilt Genetics Institute; [46]Department of Health Services, School of Public Health, University of Washington; [47]Division of Genetics and Genomics and the Manton Center for Orphan Diseases Research, Boston Children's Hospital, and the Department of Pediatrics, Harvard Medical School, Boston, MA; [48]Department of Biomedical Informatics, Columbia University, New York, NY; [49]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; [50]Population Health Sciences, Geisinger, Danville, PA; [51]Department of Medicine, Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Boston, MA; [52]Cincinnati Veterans affairs; [53]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [54]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY; [55]Departments of Medicine and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; [56]Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY; [57]Pediatric Gastroenterology & Nutrition, Geisinger, Danville, PA; [58]Baptist Cancer Center, Memphis, TN; [59]Division of General Internal Medicine, University of Washington, Seattle, WA; [60]Brigham and Women's Hospital, Harvard Medical School, Boston, MA; [61]University of Washington Biomedical and Health Informatics, Seattle, WA; [62]Department of Health Sciences Research, Mayo Clinic, Rochester, MN; [63]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center; [64]College of Medicine, University of Cincinnati, Cincinnati, Ohio; [65]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; [66]University of Cincinnati, Cincinnati, Ohio; [67]Department of Population Health Sciences, School of Medicine, Duke University, Durham, NC; [68]Division of Cardiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [69]Department of Neurology, Massachusetts General Hospital, Boston, MA; [70]Division of Human Genetics, Cincinnati Children's Hospital, Cincinnati, Ohio; [71]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center (CCHMC), Cincinnati, Ohio; [72]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; [73]Departments of Pharmacy, Medicine and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; [74]Biomedical Ethics Research Program, Mayo Clinic, Rochester, MN; [75]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA; [76]Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; [77]Division of Medical Genetics, School of Medicine, University of Washington, Seattle, WA; [78]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; [79]Center for Health Promotion and Disease Prevention, Arizona State University, Phoenix, AZ; [80]Department of Psychiatry and Center for Genomic Medicine, Massachusetts General Hospital; [81]Population Health Sciences, Geisinger, Danville, PA; [82]Department of Pediatrics (Neonatology), University of Washington, Seattle, WA; [83]Department of Medicine, Johns Hopkins University, Baltimore, MD; [84]Departments of Pediatrics and Medicine, Vanderbilt University Medical Center, Nashville, TN; [85]Department of Pharmacy, University of Washington, Seattle, WA; [86]The Comparative Health Outcomes, Policy & Economics (CHOICE) Institute, Seattle, WA; [87]Department of Obstetrics and Gynecology, Division of Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN; [88]Division of Nephrology, Department of Medicine, Columbia University, New York, NY; [89]Center for Health Information Partnerships, Northwestern University, Chicago, IL; [90]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; [91]Department of Medicine, Harvard Medical School, Boston, MA; [92]Division of Cardiovascular Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [93]Department of Bioethics and Humanities, School of Medicine, University of Washington, Seattle, WA; [94]Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [95]Department of Dermatology, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY; [96]Department of Pediatrics, University of Cincinnati, Cincinnati, OH; [97]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA. **eMERGE Consortium representative.

Author contributions

JH, HLR, RAG, DMM contributed to the study concept and design; JH, VK, HZ, MVH, CK, MES annotated and compiled information regarding sample accessioning; HZ, MVH, DM, EV performed NGS data analysis; NL, MES, GJ, HLR, RAG, DMM provided funding support for the project; Investigation: JH, VK, HZ, MVH, AM, SMH, CK, MES, AG, PS, MK, HB, LM, HLR, RAG, DMM conducted the research and investigation process of sample verification; AB, LZ, GC, DP performed the 96-SNP panel and Illumina array genotyping assay; VK, CK, RR, KW, MM participated in the project administration; MES, AG, GJ, PS, MK, HB, CP provided eMERGE sample collections; JH, MM, EV, HLR, RAG, DMM supervised the studies; JH, HZ, MVH, HLR, RAG, DMM were the major contributors in original draft writing; JH, HZ, MVH, DM, AM, SMH, NL, RR, KW, AG, GJ, PS, MK, HB, MM, EV, EB, CP, LM, HLR, RAG, DMM participated in manuscript revision. All authors read and approved the final manuscript.

Funding

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).

Availability of data and materials

Data are available in dbGaP for controlled public access (phs001616.v1.p1).

Declarations

Ethics approval and consent to participate

The Electronic Medical Records and Genomics (eMERGE) Network is a National Human Genome Research Institute (NHGRI)-funded consortium tasked with developing methods and best practices for utilization of electronic medical record (EMR) as a tool for genomic research. 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. The protocol number for Baylor College of Medicine was (#H-40455).

Consent for publication

Not applicable.

Competing interests

JH, DM, MM, RAG, DMM disclose that the Baylor Genetics Laboratory is co-owned by Baylor College of Medicine. EV is cofounder of Codified Genomics, which provides variant interpretation services. DM has received consulting fees from Illumina. The remaining authors disclose they have no competing interests.

Received: 28 August 2023 Accepted: 16 February 2024
Published online: 03 March 2024

References

1. Norton N, Li D, Hershberger RE. Next-generation sequencing to identify genetic causes of cardiomyopathies. Curr Opin Cardiol. 2012;27(3):214–20.
2. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012;71(1):5–14.
3. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. N Engl J Med. 2013;369(16):1502–11.
4. Hayeems RZ, Dimmock D, Bick D, Belmont JW, Green RC, Lanpher B, Jobanputra V, Mendoza R, Kulkarni S, Grove ME, et al. Clinical utility of genomic sequencing: a measurement toolkit. NPJ Genom Med. 2020;5(1):56.
5. Hammerling JA. A review of medical errors in laboratory diagnostics and where we are today. Lab Med. 2012;43(2):41–4.
6. Butler JM. Chapter 14: Short tandem repeat analysis for human identity testing. In: Current protocols in human genetics. New York: Wiley; 2004. p. 18.
7. Butler JM. Short tandem repeat typing technologies used in human identity testing. Biotechniques. 2007;43(4):ii–v.
8. Butler JM, Coble MD, Vallone PM. STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci Med Pathol. 2007;3(3):200–5.
9. Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med. 2013;5(9):89.
10. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van't Hof P, van Iterson M, Zhernakova DV, Claringbould A, Franke L, et al. A SNP panel for identification of DNA and RNA specimens. BMC Genom. 2018;19(1):90.
11. Gurkan C, Bulbul O, Kidd KK. Editorial: Current and emerging trends in human identification and molecular anthropology. Front Genet. 2021;12: 708222.
12. eMerge C. Harmonizing clinical sequencing and interpretation for the eMERGE III Network. Am J Hum Genet. 2019;105(3):588–605.
13. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, Kidd KK. SNPs for a universal individual identification panel. Hum Genet. 2010;127(3):315–24.
14. Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet. 2009;10:39.
15. Liang-Chu MM, Yu M, Haverty PM, Koeman J, Ziegle J, Lee M, Bourgon R, Neve RM. Human biosample authentication using the high-throughput, cost-effective SNPtrace(TM) system. PLoS ONE. 2015;10(2): e0116218.
16. McGuire AL, Gibbs RA. Genetics. No longer de-identified. Science. 2006;312(5772):370–1.
17. Lin Z, Altman RB, Owen AB. Confidentiality in genome research. Science. 2006;313(5786):441–2.
18. Miller JK, Buchner N, Timms L, Tam S, Luo X, Brown AM, Pasternack D, Bristow RG, Fraser M, Boutros PC, et al. Use of Sequenom sample ID Plus(R) SNP genotyping in identification of FFPE tumor samples. PLoS ONE. 2014;9(2): e88163.
19. Castro F, Dirks WG, Fahnrich S, Hotz-Wagenblatt A, Pawlita M, Schmitt M. High-throughput SNP-based authentication of human cell lines. Int J Cancer. 2013;132(2):308–14.
20. Ponomarenko P, Ryutov A, Maglinte DT, Baranova A, Tatarinova TV, Gai X. Clinical utility of the low-density Infinium QC genotyping Array in a genomics-based diagnostics laboratory. BMC Med Genom. 2017;10(1):57.

Publisher's Note