

Using genotype array data to compare multi- and single-sample variant calls and improve variant call sets from deep coverage whole-genome sequencing data

Supplementary Information

Suyash S. Shringarpure¹, Rasika A. Mathias^{2,3}, Ryan D. Hernandez^{4,5,6}, Timothy D. O'Connor^{7,8,9}, Zachary A. Szpiech⁴, Raul Torres¹⁰, Francisco M. De La Vega¹, Carlos D. Bustamante¹, Kathleen C. Barnes^{2,3} and Margaret A. Taub¹¹, on behalf of the CAAPA consortium¹²

¹Genetics Department, Stanford University School of Medicine, Stanford, CA USA

²Department of Medicine, Johns Hopkins University, Baltimore, MD.

³Department of Epidemiology, Bloomberg School of Public Health, JHU, Baltimore, MD.

⁴Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA

⁵Institute for Human Genetics, University of California, San Francisco, San Francisco, CA.

⁶California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA.

⁷Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD.

⁸Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore.

⁹Department of Medicine, University of Maryland School of Medicine, Baltimore, MD.

¹⁰Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, CA.

¹¹Department of Biostatistics, Bloomberg School of Public Health, JHU, Baltimore, MD.

¹²See Supplementary Materials for full listing of consortium contributors.

1 Supplementary Tables

Sample	Average Depth	Fraction of Missing Calls	Percent YRI Ancestry	Population/Sample Site
Sample 1	36.74	0.070	0.72	Atlanta
Sample 2	34.96	0.074	0.76	Wake Forest U.
Sample 3	35.21	0.072	0.65	Honduras
Sample 4	28.77	0.079	0.25	Columbia
Sample 5	34.91	0.088	0.97	Barbados
Sample 6	34.60	0.073	0.84	Johns Hopkins U.
Sample 7	31.00	0.078	1.00	African
Sample 8	30.79	0.076	0.71	Chicago
Sample 9	31.99	0.074	0.93	Jamaica
Sample 10	38.28	0.070	0.88	Nashville
CAAPA average	35.04	0.070	0.70	—

Supplementary Table 1: Summary of characteristics of samples included in analysis, including average sequencing depth, fraction of missing genotype calls from sequencing, estimated percent Yoruban (YRI) ancestry and population or sample site from the CAAPA consortium.

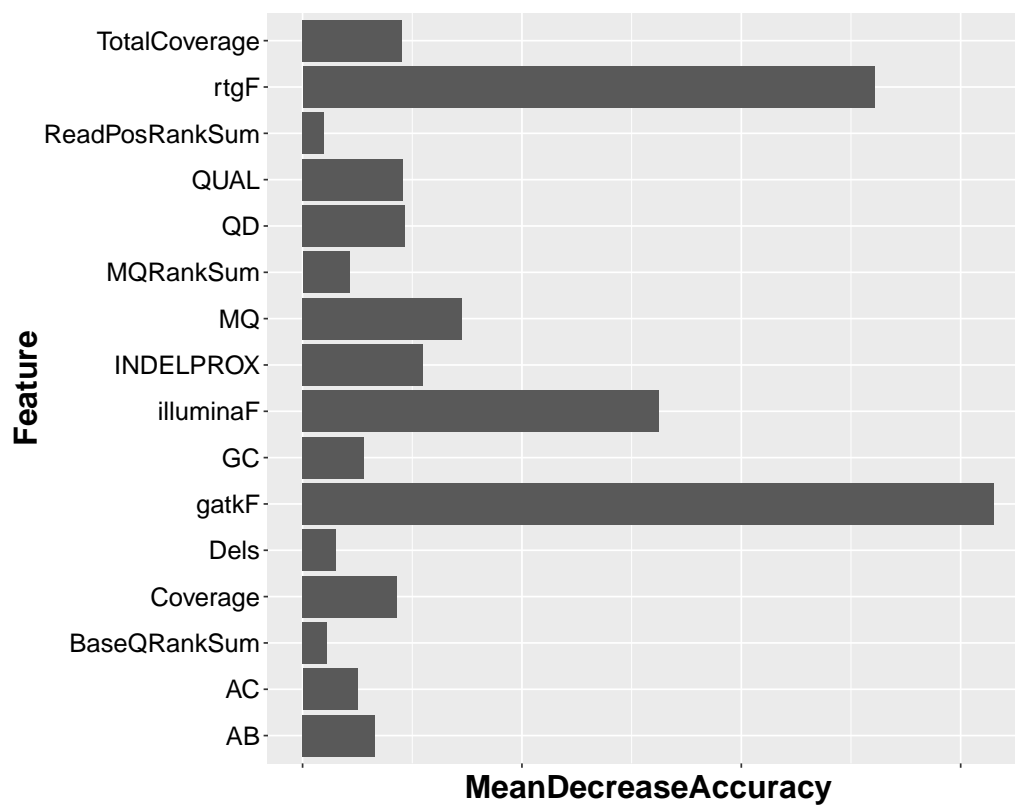
Feature	Description	Membership	Call quality
AB	Allele balance: fraction of reads carrying alternate alleles	Y	Y
Coverage	Number of reads covering the site	Y	Y
TotalCoverage	Total number of reads covering the site summed across all samples	Y	Y
INDELPROX	Proximity to indels: 1 if indel within 10 base pairs of site, 0 otherwise	Y	Y
ReadPosRankSum	Bias in allele positioning within reads comparing reference and alternate alleles	Y	N
MQRankSum	Bias in mapping quality for the reads containing reference and alternate alleles	Y	N
BaseQRankSum	Bias in base-call quality comparing reference and alternate alleles	Y	N
QUAL	Phred-scaled indicator of probability that a site is variable in the sample or set of samples	Y-Full	Y
Dels	Indicator of whether site has a spanning deletion: 1 if yes, 0 otherwise	Y	Y
MQ	Mapping Quality: average mapping quality for all reads covering the site	Y	Y
QD	Quality by depth: QUAL divided by number of reads covering the site	Y-Full	Y
AC	Allele count in genotypes, for each ALT allele, in the same order as listed	Y-Full	Y
GC	Fraction of G or C nucleotides in 400bp window centered on site	Y	Y
FS	Strand bias comparing reference and alternate alleles calculated using Fisher's exact test	N	Y
HaplotypeScore	Strength of evidence for more than two segregating haplotypes	N	Y
illuminaF	Frequency of the SNV in the Illumina callset	Y-Full	Y
rtgF	Frequency of the SNV in the RTG callset	Y-Full	Y
gatkF	Frequency of the SNV in the GATK callset	Y-Full	Y

Supplementary Table 2: Features used for classifying variants according to either call set membership (Membership column = Y for limited classifier, Y-Full for full classifier) or call quality (Call quality = Y). Each feature is specific to a particular site. With the exception of the frequencies, all values are generated by running the GATK on the full set of potential sites. Frequencies were generated from calls made on the full set of samples, excluding the particular sample being considered.

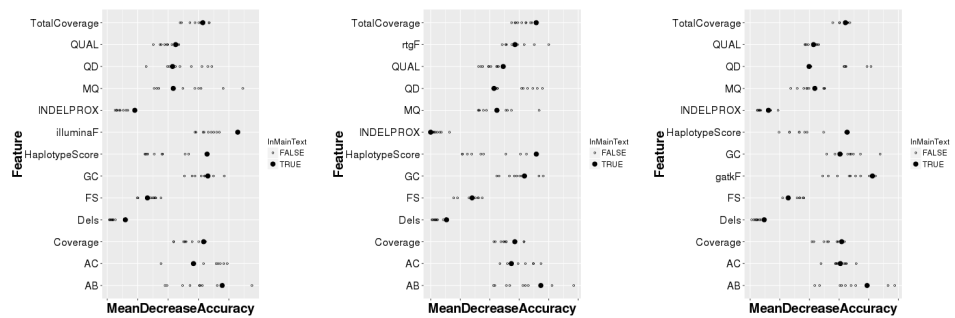
	Illumina only	Illumina + RTG	Illumina + RTG + GATK	Illumina + GATK	RTG only	RTG + GATK	GATK only	Error Rate
Illumina only	4787	44	4	15	79	0	0	3%
Illumina+RTG	4	3326	1	0	229	0	0	7%
Illumina+RTG+GATK	0	0	43290	1	0	0	0	0%
Illumina+GATK	0	0	16	6019	0	0	10	0%
RTG only	19	201	0	0	6424	0	0	3%
RTG+GATK	0	0	95	0	4	681	0	13%
GATK only	0	0	0	73	3	2	1007	7%

Supplementary Table 3: Confusion matrix for classifying variants. Rows show the “true” labels of variants depending on which methods find a variant. Columns show the predictions from our Random Forest classifier. The overall error rate of the classifier is 1.2%.

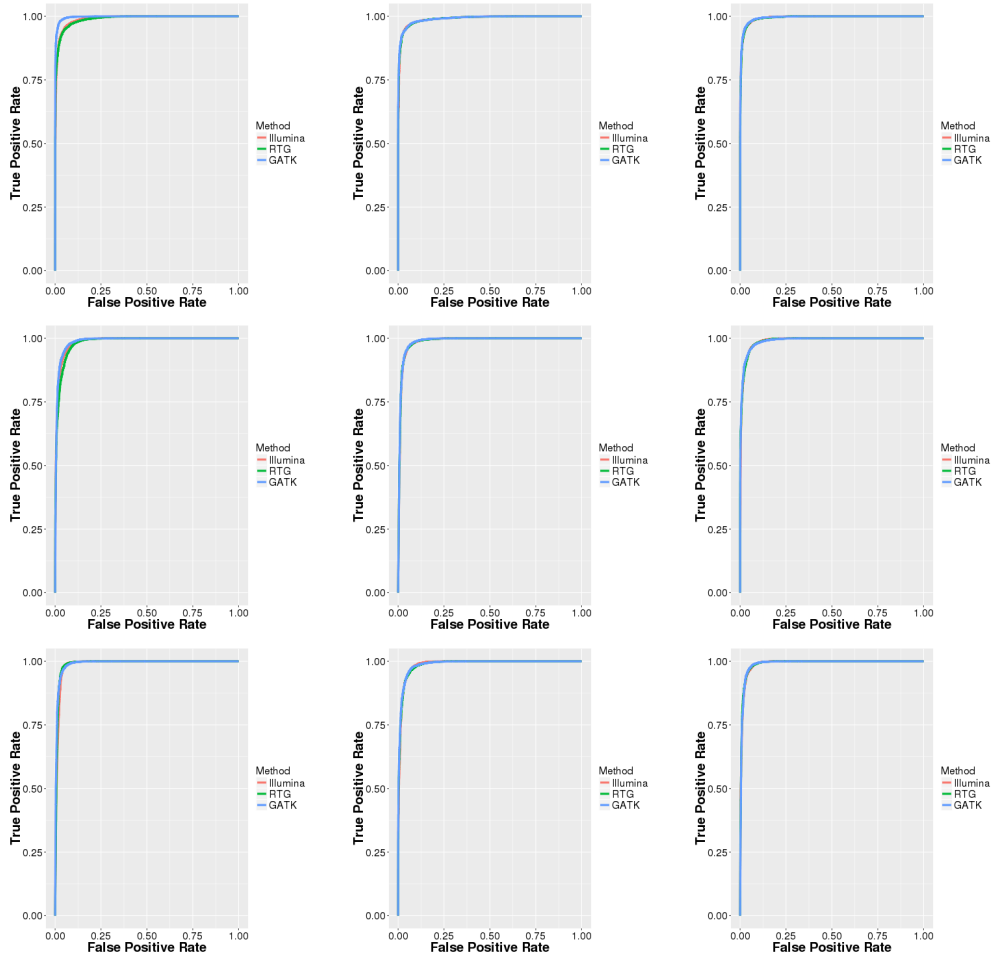
2 Supplementary Figures



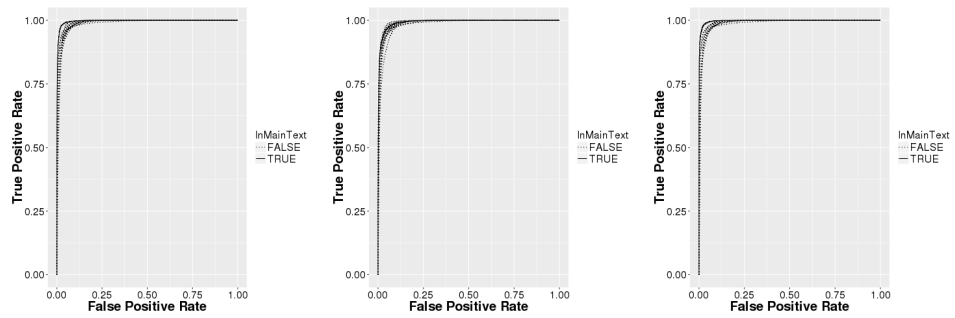
Supplementary Figure 1: Feature importance for the Random Forest classifier distinguishing calls made by different calling algorithms, including the full set of features. Scale on the x-axis is unitless but indicates relative importance of the different features.



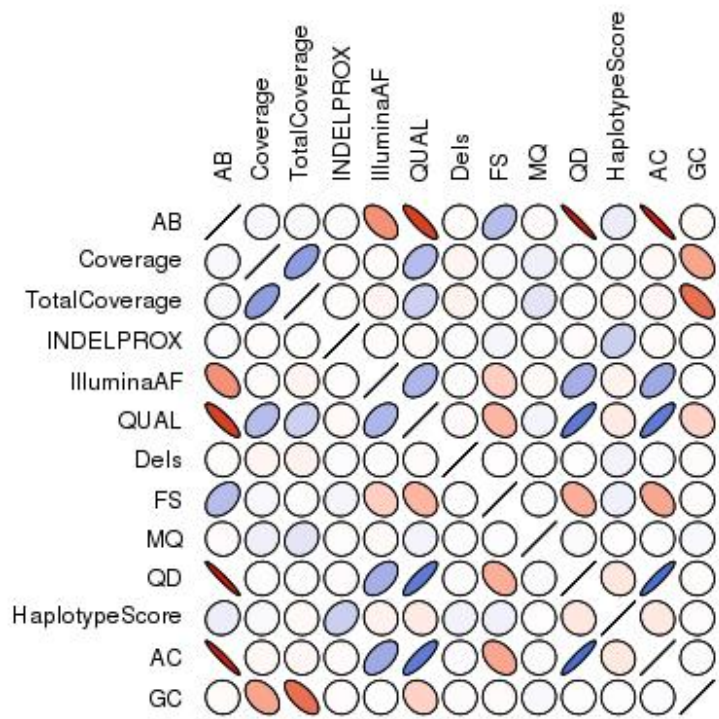
Supplementary Figure 2: Feature importance for call set-specific classifiers based on Omni genotype data across all 10 samples considered. Note that the frequency features refer to the estimates of the allele frequency from the call set being studied. Shown are results from classifiers for Illumina (left), RTG (center) and GATK (right). Bold points indicate the values for Sample 1, presented in the main text.



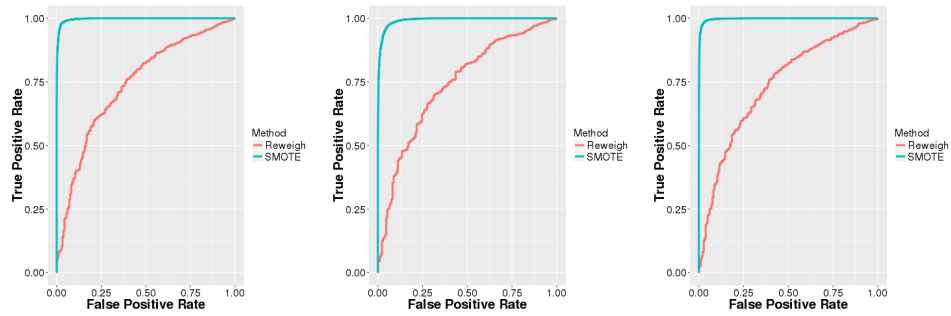
Supplementary Figure 3: ROC curves for call set-specific classifiers based on Omni genotype data across additional 9 samples considered.



Supplementary Figure 4: ROC curves for call set-specific classifiers based on Omni genotype data across all 10 samples considered, zoomed in to highlight region of interest. Shown are results from classifiers for Illumina (left), RTG (center) and GATK (right). Bold lines indicate the curves for Sample 1, presented in the main text.



Supplementary Figure 5: Correlation between features. Blue denotes positive correlation and red denotes negative correlation. Darker shades represent stronger correlation.



Supplementary Figure 6: ROC curves for call set-specific classifiers based on Omni genotype data comparing results using SMOTE to reweighting. Shown are results from classifiers for Illumina (left), RTG (center) and GATK (right).

Complete list of CAAPA consortium members and their affiliations

Kathleen C. Barnes, PhD^{1,2}, Terri H. Beaty, PhD², Meher Preethi Boorgula, MS¹, Monica Campbell, BS¹, Sameer Chavan, MS¹, Jean G. Ford, MD^{2,3}, Cassandra Foster, CCRP¹, Li Gao, MD, PhD¹, Nadia N. Hansel, MD, MPH¹, Edward Horowitz, BA¹, Lili Huang, MPH¹, Rasika Ann Mathias, ScD^{1,2}, Romina Ortiz, MA¹, Joseph Potee, MS¹, Nicholas Rafaels, MS¹, Ingo Ruczinski, PhD⁴, Alan F. Scott, PhD¹, Margaret A. Taub, PhD⁴, Candelaria Vergara, PhD¹, Jingjing Gao, PhD⁵, Yijuan Hu, PhD⁶, Henry Richard Johnston, PhD⁶, Zhaohui S. Qin, PhD⁶, Albert M. Levin, PhD⁷, Badri Padhukasahasram, PhD⁸, L. Keoki Williams, MD, MPH^{8,9}, Georgia M. Dunston, PhD^{10,11}, Mezbah U. Faruque, MD, PhD¹¹, Eimear E. Kenny, PhD^{12,13}, Kimberly Gitzen, PhD¹⁴, Mark Hansen, PhD¹⁴, Rob Genuario, PhD¹⁴, Dave Bullis, MBA¹⁴, Cindy Lawley, PhD¹⁴, Aniket Deshpande, MS¹⁵, Wendy E. Grus, PhD¹⁵, Devin P. Locke, PhD¹⁵, Marilyn G. Foreman, MD¹⁶, Pedro C. Avila, MD¹⁷, Leslie Grammer, MD¹⁷, Kwang-Youn A. Kim, PhD¹⁸, Rajesh Kumar, MD^{19,20}, Robert Schleimer, PhD²¹, Carlos Bustamante, PhD¹², Francisco M. De La Vega, DS¹², Chris R. Gignoux, MS¹², Suyash S. Shringarpure, PhD¹², Shaila Musharoff, MS¹², Genevieve Wojcik, PhD¹², Esteban G. Burchard, MD, MPH^{22,23}, Celeste Eng, BS²³, Pierre-Antoine Gourraud, PhD²⁴, Ryan D. Hernandez, PhD^{22,25,26}, Antoine Lizee, PhD²⁴, Maria Pino-Yanes, PhD^{23,27}, Dara G. Torgerson, PhD²³, Zachary A. Szpiech, PhD²², Raul Torres, BS²⁸, Dan L. Nicolae, PhD^{29,30}, Carole Ober, PhD³¹, Christopher O Olopade, MD, MPH³², Olufunmilayo Olopade, MD²⁹, Oluwafemi Oluwole, MSc²⁹, Ganiyu Arinola, PhD³³, Timothy D. O'Connor, PhD^{34,35,36}, Wei Song, PhD^{34,35,36}, Goncalo Abecasis, DPhil³⁷, Adolfo Correa, MD, MPH, PhD³⁸, Solomon Musani, PhD³⁸, James G. Wilson, MD³⁹, Leslie A. Lange, PhD⁴⁰, Joshua Akey, PhD⁴¹, Michael Bamshad, MD⁴², Jessica Chong, PhD⁴², Wenqing Fu, PhD⁴¹, Deborah Nickerson, PhD⁴¹, Alexander Reiner, MD, MSc⁴³, Tina Hartert, MD, MPH⁴⁴, Lorraine B. Ware, MD^{44,45}, Eugene Bleecker, MD⁴⁶, Deborah Meyers, PhD⁴⁶, Victor E. Ortega, MD⁴⁶, Pissamai Maul, BSc, RN⁴⁷, Trevor Maul, RN⁴⁷, Harold Watson, MD^{48,49}, Maria Ilma Araujo, MD, PhD⁵⁰, Ricardo Riccio Oliveira, PhD⁵¹, Luis Caraballo, MD, PhD⁵², Javier Marrugo, MD⁵³, Beatriz Martinez, MSc⁵², Catherine Meza, LB⁵², Gerardo Ayestas⁵⁴, Edwin Francisco Herrera-Paz, MD, MSc^{55,56,57}, Pamela Landaverde-Torres⁵⁵, Said Omar Leiva Erazo⁵⁵, Rosella Martinez, BSc⁵⁵, Al-

varo Mayorga, MD⁵⁶, Luis F. Mayorga, MD⁵⁵, Delmy-Aracely Mejia-Mejia, MD^{56,57}, Hector Ramos⁵⁵, Allan Saenz⁵⁴, Gloria Varela⁵⁴, Olga Marina Vasquez⁵⁷, Trevor Ferguson, MBBS, DM, MSc⁵⁸, Jennifer Knight-Madden, MBBS, PhD⁵⁸, Maureen Samms-Vaughan, MBBS, DM, Ph⁵⁹, Rainford J. Wilks, MBBS, DM, MSc⁵⁸, Akim Adegnika, MD, PhD^{60,61,62}, Ulysse Ateba-Ngoa, MD^{60,61,62}, Maria Yazdanbakhsh, PhD⁶²

¹ Department of Medicine, Johns Hopkins University, Baltimore, MD.

² Department of Epidemiology, Bloomberg School of Public Health, JHU, Baltimore, MD.

³ Department of Medicine, The Brooklyn Hospital Center, Brooklyn, NY.

⁴ Department of Biostatistics, Bloomberg School of Public Health, JHU, Baltimore, MD.

⁵ Data and Statistical Sciences, AbbVie, North Chicago, IL.

⁶ Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA.

⁷ Department of Public Health Sciences, Henry Ford Health System, Detroit, MI.

⁸ Center for Health Policy & Health Services Research, Henry Ford Health System, Detroit, MI.

⁹ Department of Internal Medicine, Henry Ford Health System, Detroit, MI.

¹⁰ Department of Microbiology, Howard University College of Medicine, Washington, DC.

¹¹ National Human Genome Center, Howard University College of Medicine, Washington, DC.

¹² Department of Genetics, Stanford University School of Medicine, Stanford, CA.

¹³ Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York.

¹⁴ Illumina, Inc., San Diego, CA.

¹⁵ Knome Inc., Cambridge, MA.

¹⁶ Pulmonary and Critical Care Medicine, Morehouse School of Medicine, Atlanta, GA.

¹⁷ Department of Medicine, Northwestern University, Chicago, IL.

¹⁸ Department of Preventive Medicine, Northwestern University, Chicago, IL.

¹⁹ Department of Pediatrics, Northwestern University, Chicago, IL.

²⁰ The Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago.

²¹ Department of Medicine, Northwestern Feinberg School of Medicine, Chicago,

IL.

²² Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA.

²³ Department of Medicine, University of California, San Francisco, San Francisco, CA.

²⁴ Department of Neurology, University of California, San Francisco, San Francisco, CA.

²⁵ Institute for Human Genetics, Institute for Human Genetics, University of California, San Francisco, San Francisco.

²⁶ California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA.

²⁷ CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid.

²⁸ Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, CA.

²⁹ Department of Medicine, University of Chicago, Chicago, IL.

³⁰ Department of Statistics, University of Chicago, Chicago, IL.

³¹ Department of Human Genetics, University of Chicago, Chicago, IL.

³² Department of Medicine and Center for Global Health, University of Chicago, Chicago, IL.

³³ Department of Chemical Pathology, University of Ibadan, Ibadan, Nigeria.

³⁴ Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD.

³⁵ Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore.

³⁶ Department of Medicine, University of Maryland School of Medicine, Baltimore, MD.

³⁷ Department of Biostatistics, SPH II, University of Michigan, Ann Arbor, MI.

³⁸ Department of Medicine, University of Mississippi Medical Center, Jackson, MS.

³⁹ Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS.

⁴⁰ Department of Genetics, University of North Carolina, Chapel Hill, NC.

⁴¹ Department of Genomic Sciences, University of Washington, Seattle, WA.

⁴² Department of Pediatrics, University of Washington, Seattle, WA.

⁴³ University of Washington, Seattle, WA.

⁴⁴ Department of Medicine, Vanderbilt University, Nashville, TN.

- ⁴⁵ Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville.
- ⁴⁶ Center for Human Genomics and Personalized Medicine, Wake Forest School of Medicine, Winston-Salem, NC.
- ⁴⁷ Genetics and Epidemiology of Asthma in Barbados, The University of the West Indies.
- ⁴⁸ Faculty of Medical Sciences Cave Hill Campus, The University of the West Indies.
- ⁴⁹ Queen Elizabeth Hospital, Queen Elizabeth Hospital, The University of the West Indies.
- ⁵⁰ Immunology Service, Universidade Federal da Bahia, Salvador, BA.
- ⁵¹ Laboratrio de Patologia Experimental, Centro de Pesquisas Gonalo Moniz, Salvador, BA.
- ⁵² Institute for Immunological Research, Universidad de Cartagena, Cartagena.
- ⁵³ Instituto de Investigaciones Immunologicas, Universidad de Cartagena, Cartagena.
- ⁵⁴ Faculty of Medicine, Universidad Nacional Autonoma de Honduras en el Valle de Sula, San Pedro Sula.
- ⁵⁵ Facultad de Medicina, Universidad Catolica de Honduras, San Pedro Sula.
- ⁵⁶ Centro de Neumologia y Alergias, San Pedro Sula.
- ⁵⁷ Faculty of Medicine, Centro Medico de la Familia, San Pedro Sula.
- ⁵⁸ Tropical Medicine Research Institute, The University of the West Indies.
- ⁵⁹ Department of Child Health, The University of the West Indies.
- ⁶⁰ Centre de Recherches Mdicales de Lambarn.
- ⁶¹ Institut fr Tropenmedizin, Universitt Tbingen.
- ⁶² Department of Parasitology, Leiden University Medical Center, Netherlands.