Supplementary Information

Table of Contents

1. SV callset quality assessment

We performed a host of analyses to assess callset quality in terms of sensitivity, accuracy and genotyping error. These include: (i) an analysis of validation rate using deep coverage PacBio sequencing of nine samples included in our study, (ii) an analysis of SV detection sensitivity using a comprehensive long-read SV callset from the Human Genome Structural Variation Consortium (HGSVC)[1], and (iii) an analysis of Mendelian inheritance among families included in our study. Where possible, we also compared our callset to the two most comprehensive short-read callsets published thus far: the 1000 Genomes Project (1KG) Phase3 callset based on 2504 individuals and 9 algorithms[2], and the HGSVC short-read callset based on 9 individuals and 13 algorithms[1].

Taken together, the results described below demonstrate that the SV callsets reported here have (i) high variant detection accuracy; (ii) levels of sensitivity that are low compared to the HGSVC long-read SV map but comparable to best-in-class short-read maps from 1KG and HGSVC; and (iii) excellent genotyping accuracy that is equivalent if not superior to prior large-scale short-read WGS studies from 1KG[2] and GTEx[3]. Notably, the sensitivity and accuracy of our callset is weakest at small and repetitive variants that are least likely to be functional, and strongest at the subset of variants that are most important for our findings and conclusions: variants that are rare, large, or predicted to be functionally relevant based on variant impact scores.

1.1. SV validation using long-read whole-genome sequencing data

We assessed the accuracy of our SV callset by measuring the validation rate of SV calls in nine samples for which deep coverage (>52-85X) long-read Pacific Biosciences (PacBio) data was available from an unrelated project at our genome center (**Supplementary Table 2**). We followed the same strategy used in our prior work[4,5]: we used split-read mappings from long-reads to validate SV calls made from short-reads (see **Methods**). We judged a short-read SV call to be validated if the breakpoint coordinates and orientations were confirmed by split-read mappings from multiple long-reads. We assessed the validation rate according to SV type, frequency, and size. The overall validation rate in the nine genomes with long-read data is 84% (**Supplementary Table 3a**). This compares favorably to the 74% validation rate obtained for the HGSVC short-read callset (**Supplementary Table 3b**), which includes a total of three trios, the offspring of which are included in our long-read dataset. Importantly, in our callset the validation rate is substantially higher for the specific variant classes that are most crucial for our findings and conclusions: deletions (87%), rare SV (90%) and singleton SV (95%). Moreover, whereas higher-validating rare and singleton variants comprise a small minority of SVs in the nine genomes assessed by long-read WGS (1.9% and 0.33% of sites, respectively), they comprise the vast majority of SVs in the larger B38 callset including all samples (35% and 58% of sites). Based on the validation rate of different SV frequency classes and their relative abundance in the full dataset, the B38 dataset has an SV false discovery rate (FDR) of 7.0%. We believe that this is likely to be an overestimate of FDR considering that true variants in repetitive regions can be missed by long-read alignment; for example, 3 of 6 singleton variants failing automated long-read validation (of 133 in total) appear to be true variants based on manual review of raw alignment signals (see **Methods**).

1.2. SV inheritance patterns in families

A broad measure of callset quality is the extent to which SV calls and genotypes among related samples are consistent with the laws of Mendelian inheritance. We first measured Mendelian error (ME) rate within the parent-offspring trios that were included in the final public versions of the B37 (n=452 trios) and B38 (n=36 trios) callsets. For most variant types in both callsets, the ME rate was acceptably low (≤5%) using our default variant filtering approach[6] (see **Methods**; **Extended Data Figs. 2c and 3g**). However, the ME rate was higher for tandem duplications (DUP), inversions (INV) and unclassified (BND) variants in the B38 callset. This slightly elevated ME rate is in our view tolerable for DUPs since they are often multi-allelic and difficult to genotype, and because all DUP variants were detected by at least two independent sources of evidence including breakpoint spanning alignments (read-pair and/or split-read) plus read-depth evidence, making false positives unlikely. For the INV and BND types, ME rate correlates well with the Mean Sample Quality (MSQ) field used to judge site-level confidence (**Extended Data Fig. 3g**); we therefore redefined "High Confidence" variant calls by additionally filtering on the MSQ field at a threshold that yielded ME rate ≤5%.

We next examined inheritance patterns within a set of large 3-generation CEPH pedigrees comprising 576 individuals and 142 founders, which have 4-14 grandchildren in the F2 generation (median=8) (**Extended Data Fig. 6a**) and include a total of 409 trios. We calculated the transmission rate for different SV frequency classes, and found that all classes exhibit rates that are close to the expected 50% (**Extended Data Fig. 6b**). We examined the ME rate by frequency class and found acceptably low ME rates for all classes (**Extended Data Fig. 6c)**. Notably, of the 3,359 variants that were private to a single family and were observed in the 329 samples from the third generation (F2), we observed only 82 Mendelian errors (**Extended Data Fig. 6d**), and 100% of these were caused by predicted *de novo* mutations. These could potentially be caused by three sources: (1) true *de novo* mutations, (2) false positive SV calls, (3) an ultra-rare inherited SV with a genotyping error in the parental carrier. To distinguish these, we examined the transmission of mutation predictions in the F1 generation, and found that 18 of 21 (86%) predicted *de novo* SVs from the F1 were transmitted to at least one individual in the F2. This demonstrates that most *de novo* mutation calls – and, by extension, most singleton variants in the larger set of unrelated samples – are *bona fide* germline variants rather than false positives. The third explanation, genotyping error, is unlikely considering the size of the families examined here, where a heterozygote parental variant will typically be transmitted to more than one progeny, and have more opportunities to be detected as an inherited rather than spontaneous variant. Consistent with this, of the 775 family private variants carried by at least two siblings in the F2 generation, only 1.7% were not carried by a single parent from the P0 generation, suggesting that the genotype "undercall" rate is low (i.e., <1.7%).

1.3. SV detection sensitivity based on comparison to long-read SV maps

A key challenge for measuring sensitivity is the availability of a high quality "truthset" in which the entire set of true variants in some individual(s) are known. The most comprehensive SV map published thus far is from the recent HGSVC study[1], which used deep long-read sequencing and local reference-guided assembly to report 72,297 autosomal variant calls calls among 3 samples. These same 3 samples were included in our short-read callset, enabling a direct comparison. A challenge is that there are false positives in the HGSVC callset. To reduce the effects of false positives, we derived a higher-confidence set of variant calls by imposing the additional requirement that HGSVC-derived calls must be validated by split-read mapping analysis of the WashU PacBio data, which was generated and analyzed independently from the HGSVC study (see **Methods**). Of the 72,297 variant calls reported by the HGSVC (mean 24,099 per person), we were able to validate 66,239 (91.6%) using this strategy. These calls comprise an initial truthset. A caveat is that the majority of SV calls from HGSVC correspond to short tandem repeat (STR) variants caused by expansion or contraction of smaller repeat units, which are more conventionally referred to in the literature as microsatellites, minisatellites or variable number tandem repeats (VNTRs). These STR variants are more repetitive, more difficult to detect, and functionally less potent than SVs reported by microarray or short-read WGS studies, which typically affect ≥50bp of relatively unique non-STR sequence. Indeed, sensitive detection of STR variants from short-read WGS data requires highly specialized algorithms (e.g., see [7-11]). To distinguish these abundant STR variants from other SVs, we derived a second truthset that includes the 21,566 SV calls at which ≥50% of the structurally variable interval is not covered by annotated STRs. We report results using both truthsets (**Supplementary Tables 4a and 4b**) but focus the discussion below on the latter more restricted set which is more relevant to the goals and methods of our study.

Based on this truthset, our callset achieves an overall sensitivity of 49%. This is inferior to the most comprehensive short-read callset generated to date – HGSVC – which achieves a sensitivity of 63%. It is not surprising that our callset is less sensitive than HGSVC considering that HGSVC used a compendium approach involving 13 variant callers (one of which is the same used for our study, LUMPY), which improves sensitivity but limits scalability for large datasets, and also incurs a greater burden of false positives as apparent by the lower overall validation rate of HGSVC calls in section 1.1 above (74% vs. 84%). However, the difference in sensitivity between our study and HGSVC is due primarily to small and repetitive SVs such as Alu insertions, and the sensitivity of our callset improves when restricting to the subset of variant calls that are most crucial for our results. For example, if we limit the truthset to variants that are most likely to be functional by restricting to >1 kb in size (n=3728), our study achieves superior sensitivity to HGSVC (63% vs. 53%). Similarly, if we restrict the truthset to SVs in the top 10% of predicted impact scores, which are more likely to be functionally relevant, the two studies achieve similar sensitivity (82% vs. 86%). Our callset also achieves higher overall sensitivity than the 1KG callset (52% vs. 36%) to detect in at least one parent variants in the two individuals subjected to long-read WGSs (1KG did not analyze the children) (**Supplementary Tables 4c and 4d**), which is notable considering that the 1KG callset has served as the gold standard SV reference resource for several years.

1.4. Additional evidence supporting SV callset quality

A number of additional lines of evidence support the quality of our SV callsets. First, the number and types of variants that we report are roughly consistent across samples, cohorts and sequencing center, and with previous studies that used similar methods, including 1KG[2] and GTEx[3] (**Extended Data Figs. 2 and 3**). Second, the distribution of linkage disequilibrium between an SV and it's most tightly linked SNV (**Extended Data Fig. 2e**) is consistent with our callset from the GTEx study[3], which used very similar methods and characterized the SV calls and genotypes extensively in the context of eQTL mapping, including a comparison to 1KG data. Third, there is a broadly consistent number of ultra-rare singleton variants per sample, and the number of singleton variants is lowest in Finnish samples and highest in African American samples (controlling for sample size), as expected (**Extended Data Fig 3d**). Fourth, principal components analysis using SV calls reveals population structure that is consistent with self reported ancestry (**Extended Data Fig. 4**). Fifth, the site frequency distribution of SVs is broadly similar to that of SNPs (**Fig. 1e and Extended Data Fig. 2d**). Sixth, the SV size distribution shows increased mean length with decreased frequency, which is expected considering the strength of selection against large variants. Seventh, assessment of these tools in a recent publication demonstrated that they achieve expected variant detection performance on 1KG project samples, as compared to the 1KG project callset[6]. Eighth, the dosage sensitivity analyses reported in **Fig. 4** are consistent with independent measures of functional constraint including pLI and dosage sensitivity scores from ExAC[12], haploinsufficiency scores from DECIPHER[13], evolutionary conservation scores from PHASTCONS[14], and non-coding impact prediction scores from LINSIGHT[15].

Finally, comparison of SV genotypes at variants discovered by both our study and 1KG shows that the two datasets are broadly consistent, with a genotype concordance of 91% (kappa=0.852) across 2,643 variants and 13,201 observations, with most discordant genotypes occurring at mobile element insertions (MEIs) (**Extended Data Fig. 7c and 7d**). These discordant genotype observations are more likely to be caused by errors in 1KG than our dataset for two reasons: (i) MEIs in our callset are well-tagged by neighboring SNVs based on the distribution of max R^2 values plotted in **Extended Data Fig. 2e**, which should not be possible with a high genotyping error rate; and (ii) at discordant sites we observe a strong correlation in our dataset between SV genotypes determined by breakpoint-spanning reads and independently derived copy number estimates based on read-depth information, and this correlation is much less strong when using genotypes from the 1KG dataset (**Extended Data Fig. 7e**).

2. Supplementary References

1 Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784, doi:10.1038/s41467-018-08148-z (2019).

2 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

3 Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nature genetics* **49**, 692-699, doi:10.1038/ng.3834 (2017).

4 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

5 Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**, 966-968, doi:10.1038/nmeth.3505 (2015).

6 Larson, D. E. *et al.* svtools: population-scale analysis of structural variation. *Bioinformatics*, doi:10.1093/bioinformatics/btz492 (2019).

7 Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**, 1154-1162, doi:10.1101/gr.135780.111 (2012).

8 Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14**, 590-592, doi:10.1038/nmeth.4267 (2017).

9 Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol* **19**, 121, doi:10.1186/s13059-018-1505-2 (2018).

10 Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res* **28**, 1709-1719, doi:10.1101/gr.235119.118 (2018).

11 Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* **27**, 1895-1903, doi:10.1101/gr.225672.117 (2017).

12 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

13 Bragin, E. *et al.* DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic acids research* **42**, D993-D1000, doi:10.1093/nar/gkt937 (2014).

14 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

15 Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics* **49**, 618-624, doi:10.1038/ng.3810 (2017).

3. Supplementary Tables

See accompanying Supplementary Table files provided in excel format

Supplementary Table 1. Description of which callset and sample subsets were used for each of the major analyses in the study.

Supplementary Table 2. Description of PacBio long-read datasets used for SV validation analyses. Sequencing data can be found in SRA using the accessions shown at bottom.

Supplementary Table 3. SV validation rate analysis using split-read mapping with deep coverage (>60x) PacBio long-read WGS data. **(a)** validation rate for our CCDG B38 callset ("ccdg") and the HGSVC short-read callset[27] ("hgsvc") for 3 genomes analyzed in both studies. Various subsets of variants are shown according to the value in the "Group" column, such as SV length ("len"), SV type, and frequency class. The frequency classes are as follows: singletons SVs are present in 1 family; "rare" SVs are present at <1% frequency; "low" are low-frequency SVs at 1-5% frequency; common SV are >5% frequency. Note that allele frequencies were not available for SVs reported by HGSVC since that study analyzed only 9 samples, and we therefore defined rare SVs as those observed in 1 of the 6 trio founders and <1% of samples from the 1KG Phase3 callset. The "quality bin" column refers to whether the SVs are classified as high confidence ("PASS") or low-confidence ("LOW") (see **Methods**).

Supplementary Table 4. SV detection sensitivity analysis based on long-read SV calls from the Human Genome Structural Variation Consortium (HGSVC)[27]. **(a)** Sensitivity for non-STR structural variants, where STRs were defined as SV calls at which >50% of the SV interval and/or breakpoint sequences correspond to annotated STR sequences in the reference genome. Sensitivity is shown for short-read calls from our study ("ccdg") and the HGSVC study ("hgsvc") by SV length ("len"), SV impact prediction score percentile ("impact") and SV type ("type"), as indicated by the "comparison" and "group" columns at left. The "Total" column refers to the total number of observations accounting for each variant reported by the HGSVC long-read callset in each of the three genomes in the truthset. Note that SV types used here correspond directly to the SV types reported by the HGSVC for long-read calls, where the "INS" class includes both MEI and DUP variants (see **Methods**). **(b)** Sensitivity at all SVs, including STRs. **(c)** Sensitivity comparison to 1KG[4] at non-STR SVs using the two sets of parents of the samples used in parts (a) and (b), under the logic that variants in the truthset derived from children should also be present in one of their parents. Here, sensitivity measures the fraction of variants from the offspring that were detected in either of the two parents. This approach was necessary to enable a comparison to 1KG because none of the offspring and only two sets of parents were included in the 1KG Phase3 callset. **(d)** Sensitivity analysis as in part (c), including STRs.

Supplementary Table 5. Number of variants represented in **Fig 2d**. (**a**) Number of rare (<1% MAF) gene-altering variants in each variant subclass. (**b**) Number of gene-altering SV in each variant subclass, excluding genes with pLI<0.1.

Supplementary Table 6. Number of variants for each category in **Fig. 3a and 3b**. For SNV and indel and for each of the functional annotations, the number of variants above each impact score quantile are shown.

Supplementary Table 7. Data Availability. **(a)** dbGaP accession identifiers for CRAM files and their locations on the AnVIL. **(b)** dbGap accessions identifiers for joint callsets.

4. Supplementary Files

Supplementary File 1. Site frequency map for the B38 shareable callset in vcf and bedpe format (https://github.com/hall-lab/sv_paper_042020/blob/master/Supplementary_File_1.zip).

Supplementary File 2. Site frequency map for the B37 shareable callset in vcf and bedpe format (https://github.com/hall-lab/sv_paper_042020/blob/master/Supplementary_File_2.zip).

Supplementary File 3. HGSVC SV calls used in sensitivity analyses. HGSVC assembly-only vcfs annotated with status of validation by >=2 PacBio long reads (see Methods; https://github.com/hall-lab/sv_paper_042020/blob/master/Supplementary_File_3.zip).

Supplementary File 4. DEL and DUP sensitivity scores for each gene based on the observed frequency of CNV in the combined dataset of 17,795 samples (https://github.com/hall-lab/sv_paper_042020/blob/master/Supplementary_File_4.zip).

5. Consortium Members

NHGRI Centers for Common Disease Genomics

List of collaborators:

Sample contributors and cohort PIs

Raphael A. Bernier[1], Julie Baker[2], Michael Boehnke[3], Erwin P. Bottinger[4], Steven R. Brant[5], Eric Boerwinkle[6,7], Esteban G. Burchard[8], Carlos D. Bustamante[2], Judy H. Cho[4,9,10], Rajiv Chowdhury[11], Michael J. Cutler[12], Scott M. Damrauer[13], Evan E. Eichler[14,15], Andres M. Estrada[16], Tatiana Foroud[17], Nelson B. Freimer[18], Christopher A. Haiman[19], Lynn B. Jorde[20], John Kane[21], Eimear E. Kenny[4,10,22,23], Charles Kooperberg[24], William E. Kraus[25], Subra Kugathasan[26], Markku Laakso[27], Ruth J.F. Loos[4], Loic Le Marchand[28], Gregory M. Marcus[29], Richard P. Mayeux[30], Dermot P.B. McGovern[31], Karla S. Mendoza[16], Rodney D. Newberry[32], Kari E. North[33], Aarno Palotie[34-36], Ulrike Peters[24], Clive Pullinger[21], Aaron Quinlan[20], Daniel J. Rader[37], Dan M. Roden[38], Stephen S. Rich[39], Samuli Ripatti[34-36], Veikko Salomaa[40], Svati H. Shah[25], M. Benjamin Shoemaker[38], Marja-Riitta Taskinen[41], Stephan R. Targan[31]

Broad Institute CCDG

Eric Banks[36], Mark J. Daly[34-36], Yossi Farjoun[36], Stacy Gabriel[36], Namrata Gupta[36], Patrick T. Ellinor[36,42], Daniel Howrigan[35,36], Sek Kathiresan[36,42,43], Amit Khera[36,42,43], Eric S. Lander[36,44,45], Robert Maier[35,36], Benjamin M. Neale[35, 36], Christine Stevens[36], Kathleen Tibbetts[36], Charlotte Tolonen[36]

Baylor College of Medicine Human Genome Sequencing Center CCDG

Eric Boerwinkle[6,7], Paul De Vries[6], Huyen Dinh[7], Harsha Doddapaneni[7], Richard A. Gibbs[7], Megan L. Grove[7], Yi Han[7], Jianhong Hu[7], Goo Jun[6], Ziad Khan[7], Olga Krasheninina[7], Vipin Menon[7], Ginger A. Metcalf[7], Zeineen Momin[7], Donna M. Muzny[7], Caitlin Nessner[7], Jireh Santibanez[7], William J. Salerno[7], Kimberly Walker[7], Bing Yu[6]

McDonnell Genome Institute at Washington University in St. Louis CCDG

Haley Abel[46,47], Elizabeth Appelbaum[46], Lei Chen[46], Ryan Christ[46], Lisa Cook[46], Matthew Cordes[46], Laura Courtney[46], Tracie Deluca[46], Susan K. Dutcher[46,47], Nelson B. Freimer[18], Catrina Fronick[46], Lucinda Fulton[46], Robert Fulton[46], Liron Ganel[46], Ira M. Hall[46-48], Bo Ji[46], Chul Joo Kang[46], Krishna Kanchi[46], David Larson[46,47], Adam E. Locke[46,48], Amy Ly[46], Joanne Nelson[46], Jennifer Ponce[46], Nathan O. Stitziel[46-48], Jason Waligorski[46], Richard K. Wilson[46,49], Erica Young[46,48]

New York Genome Center CCDG

Toby Bloom[50], Esteban G. Burchard[8], Robert B. Darnell[50-52], Evan E. Eichler[14,15], Shailu Gargeya[50], Goren Germer[50], Daniel H. Geschwind[53-55], David B. Goldstein[56,57], Ivan Iossifov[58], Eimear E. Kenny[4,10,22,23], Lily Khaira[50], Tuuli Lappalainen[50,59], Tom Maniatis[50,60], Guiseppe Narzisi[50], Catherine Reeves[50], Tychele Turner[14], Michael Wigler[50,58], Lara Winterkorn[50], Michael C. Zody[50]

Rutgers GSP Coordinating Center

Goncalo R. Abecasis[3], Carlos D. Bustamante[2], Steve Buyske[61], Hyun Min Kang[3], Tara Matise[62], Kari E. North[33], Genevieve Wojcik[2], Jinchuan Xing[62], Yeting Zhang[62]

NHGRI Program Staff

Adam Felsenfeld[63], Carolyn Hutter[63], Vivian Ota Wang[63], Heidi Sofia[63], Taylorlyn Stephan[63]

Affiliations

[1] Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, USA

[2] Department of Genetics, Stanford University, Stanford, CA, USA

[3] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, School of Public Health, Ann Arbor, MI, USA

[4] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[5] Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

[6] Human Genetics Center and Department of Epidemiology, University of Texas Health Science Center, Houston, TX, USA

[7] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

[8] Department of Bioengineering, University of California, San Francisco, CA, USA

[9] Department of Medicine, Icahn School of Medicine at Mt. Sinai, New York, NY, USA

[10] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mt. Sinai, New York, NY, USA

[11] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[12] Intermountain Heart Institute, Intermountain Medical Center, Murray, UT, USA.

[13] Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[14] Department of Genome Science, University of Washington, Seattle, WA, USA

[15] Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

[16] National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV, Irapuato, Guanajuato, Mexico

[17] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

[18] Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA

[19] Department of Preventative Medicine, University of Southern California, Los Angeles, CA, USA

[20] Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

[21] Cardiovascular Research Institute, University of California, San Francisco CA, USA

[22] The Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[23] Center for Statistical Genetics, Icahn School of Medicine at Mt Sinai, New York, NY, USA

[24] Fred Hutchinson Cancer Research Center, Seattle, WA, US

[25] Department of Medicine, Duke University, Durham, NC, USA

[26] Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA

[27] Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland

[28] Cancer Center, University of Hawaii, Honolulu, HI, USA

[29] Department of Medicine, University of California, San Francisco CA, USA

[30] Department of Neurology, Columbia University, New York, NY, USA

[31] F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[32] Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA

[33] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA

[34] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

[35] Analytical and Translational Genetics Unit, Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[36] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[37] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[38] Department of Medicine, Vanderbilt University, Nashville, TN, USA

[39] Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA

[40] National Institute for Health and Welfare, Helsinki, Finland

[41] Research Programs Unit, Diabetes & Obesity, University of Helsinki, and Heart and Lung Centre, Helsinki University Hospital, Helsinki, Finland

[42] Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[43] Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[44] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

[45] Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[46] McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

[47] Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

[48] Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA

[49] current address: Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA

[50] New York Genome Center, New York, NY, USA

[51] Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, NY, USA

[52] Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA

[53] Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[54] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[55] Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA

[56] Institute for Genomic Medicine, Columbia University Medical Center, New York, NY, USA

[57] Department of Genetics and Development, Columbia University Medical Center, New York, NY, USA

[58] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[59] Department of Systems Biology, Columbia University, New York, NY, USA

[60] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

[61] Department of Statistics, Rutgers University, Piscataway, NJ, USA

[62] Department of Genetics, Rutgers University, Piscataway, NJ, USA

[63] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA