


SHORT COMMUNICATION

Thousands of missing variants in the UK Biobank are recoverable by genome realignment

Tongqiu Jia¹  | Brenton Munson¹ | Hana Lango Allen² | Trey Ideker¹ | Amit R. Majithia¹

¹Department of Medicine, University of California San Diego, La Jolla, California

²Medical Research Council Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, UK

Correspondence

Amit R. Majithia and Trey Ideker, Department of Medicine, University of California San Diego, Biomedical Research Facility II (3A19) 9500 Gilman Drive, La Jolla, CA 92093-0688. Email: amajithia@ucsd.edu; tideker@ucsd.edu

Funding information

UCSD/UCLA Diabetes Research Center grant; National Institute on Drug Abuse, Grant/Award Number: P50 DA037844; National Human Genome Research Institute, Grant/Award Number: R01 HG009979; National Institute for Diabetes, Digestive and Kidney diseases; UCSD/UCLA Diabetes Research Center grant; National Institute on Drug Abuse, Grant/Award Number: P50 DA037844; National Human Genome Research Institute, Grant/Award Number: R01 HG009979

Abstract

The UK Biobank is an unprecedented resource for human disease research. In March 2019, 49,997 exomes were made publicly available to investigators. Here we note that thousands of variant calls are unexpectedly absent from this dataset, with 641 genes showing zero variation. We show that the reason for this was an erroneous read alignment to the GRCh38 reference. The missing variants can be recovered by modifying read alignment parameters to correctly handle the expanded set of contigs available in the human genome reference. Given the size and complexity of such population scale datasets, we propose a simple heuristic that can uncover systematic errors using summary data accessible to most investigators.

KEYWORDS

DNA, exome, genetics, sequence alignment, sequence analysis

1 | INTRODUCTION

The UK Biobank (UKB) is a resource of unprecedented size, scope, and openness, making available to researchers deep genetic and phenotypic data from approximately half a million individuals (Bycroft et al., 2018). The genetic data released thus far include array-based genotypes on 488,000 individuals and exome sequencing on 49,997 of these, with further exome sequences to be released in 2020. Such comprehensive cataloging of protein-coding variation across the entire allele frequency spectrum attached to extensive clinical phenotyping has the potential to accelerate biomedical discovery, as evidenced by recent successes with other exome biobanks (Abul-Husn et al., 2018). Given

the scale of the data (the current exomes release contains approximately 120 TB of aligned sequence), few investigators have the computational infrastructure or knowledge to identify and curate genetic variants and instead rely on releases of accompanying pre-processed variant calls (variant call format [VCF], approximately 5 GB). Specifically, the UKB has released pre-processed VCFs from two different variant analyses, called the Regeneron Seal Point Balinese (SPB) (Van Hout et al., 2019) and functionally equivalent (FE) pipelines (Regier et al., 2018). Although these pipelines are still evolving, studies have already made use of the released exome variants mainly for comparison with previous UKB genotyping data or variant databases (Weedon et al., 2019). However, a recent report pointed out an error in

duplicate read marking in the SPB pipeline that could lead to false variant calls (<http://www.ukbiobank.ac.uk/wp-content/uploads/2019/08/UKB-50k-Exome-Sequencing-Data-Release-July-2019-FAQs.pdf>), resulting in the removal of the SPB release from the UKB data repository. Thus, the FE pipeline is currently the only source of variant calls available for downstream research. Here we identify an error in the FE pipeline that results in a systematic lack of variant calls for thousands of genes, and we provide a solution to patch this bug.

2 | METHODS

2.1 | UK Biobank whole exome sequencing (WES) and genotype array data

We analyzed the sample-level aligned sequence data (CRAM files) from the FE pipeline (Bycroft et al., 2018). A total of 49,960 individuals had both exome-sequencing data and genotype array data as of November 26, 2019, out of which 49,909 individuals pass standard genotype array quality control. As the exome data are in coordinates relative to GRCh38, but the genotype array data are in coordinates relative to GRCh37, we used the UCSC genome browser LiftOver tool (Hinrichs et al., 2006) to update genotype data coordinates to GRCh38. To facilitate direct comparison of the exome to array genotype data (Figure 1), we filtered to select variants on the genotyping array present at a minor allele frequency (MAF) > 0.01 that were also covered by the exome-sequencing regions.

2.2 | Variant comparison to gnomAD

We obtained targeted exome capture regions for both UKB and gnomAD (Karczewski et al., 2019) (v2.1, https://storage.googleapis.com/gnomad-public/intervals/exome_calling_regions.v1.interval_list).

The exome calling regions from gnomAD were converted to GRCh38 coordinates using the UCSC genome browser LiftOver tool (Hinrichs et al., 2006) to facilitate comparison to UKB. We used BEDTools (Quinlan & Hall, 2010) to extract shared regions between UKB and gnomAD. Using BEDOPS (Neph et al., 2012), we further annotated the common genomic regions to a total of 23,040 genes based on the Ensembl 85 gene model (Yates et al., 2016). For each gene, we aggregated variants from the UKB FE pipeline project-level variant calls and compared the number of variants per gene to those in gnomAD (Figure 2a and 2b). To evaluate whether population structure contributes to the difference in variant distribution (Figure 2c), we tallied the number of variants in gnomAD when subdividing individuals into six population groups: African, Latino, East Asian, European, South Asian, and other (population not assigned).

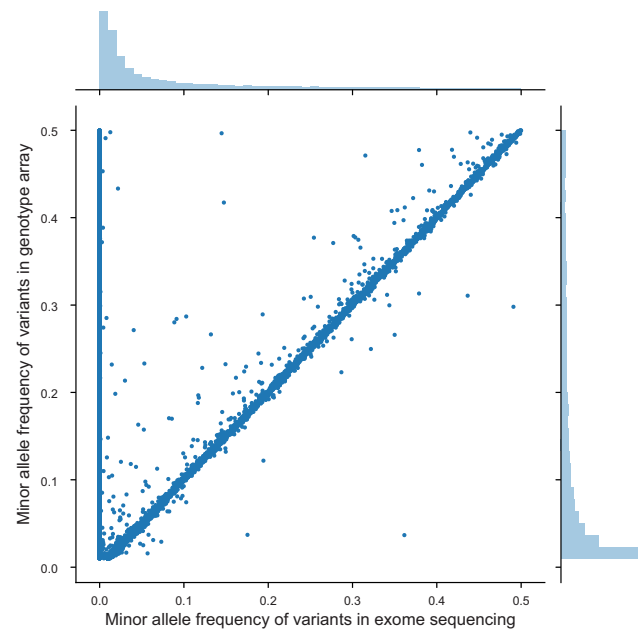


FIGURE 1 Variant allele frequencies for individuals in the UK Biobank called by analysis of genotyping arrays versus exome sequencing. The minor allele frequency (MAF) determined by each method is plotted, covering a total of 30,979 common variants measured by both methods over 49,909 individual samples. The distribution of variant allele frequencies is shown for each method by histograms above (exome) and to the right of (genotyping array) the main scatterplot [Colour figure can be viewed at wileyonlinelibrary.com]

2.3 | Extraction and reprocessing of raw unmapped reads

Using SAMtools (Li et al., 2009), we query-name sorted the aligned sequence reads in the UKB CRAM files and losslessly extracted the raw unmapped reads into FASTQ files. Using BWA-MEM (Li, 2013), these reads were mapped to the full version of the GRCh38 genome reference, which contains both the primary assembly and all alternative contigs (Figure 3c). We generated all bwa-required index files locally except the “.alt” index file, which we downloaded from the NCBI (ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa.alt). We marked duplicates and recalibrated base quality scores following GATK best practices (DePristo et al., 2011). To produce the scenario in which alternative contigs are not properly referenced (Figure 3a), we used BWA-MEM command-j to specify the aligner to ignore the “.alt” index file (Figure 3b).

3 | RESULTS

In our initial investigations of protein-coding variation in the UKB exomes, we noted a complete absence of variation in a

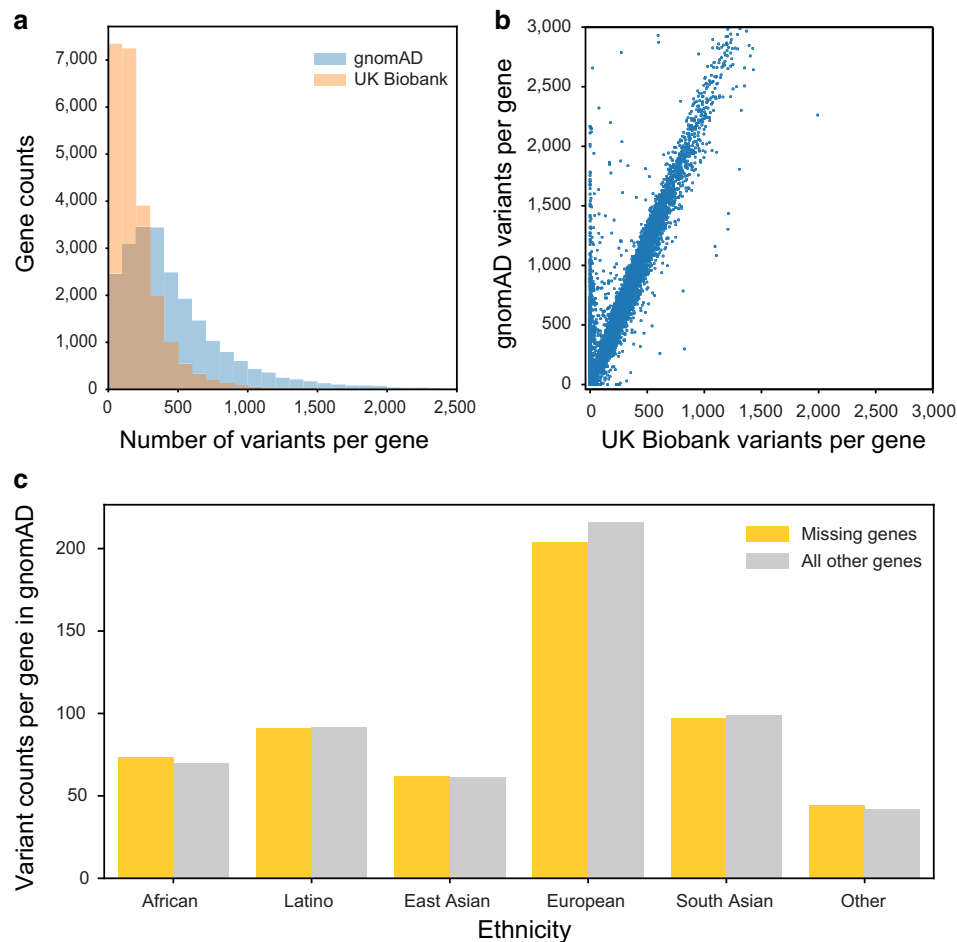


FIGURE 2 Evaluation of exome variants called by the UK Biobank (UKB) against the Genome Aggregation Database (gnomAD). (a) Histogram of variant counts for each of the 23,040 human genes commonly annotated in UKB (orange) and gnomAD (blue), at a fixed bin size of 100. (b) Scatterplot of variant counts for each gene in gnomAD versus UKB. (c) Counts for variants in the 641 genes that have variant calls in gnomAD but none in the UKB (yellow), divided into six subpopulations by ethnicity. Counts for all other human genes are shown as a reference (gray) [Colour figure can be viewed at wileyonlinelibrary.com]

number of genes of interest, including *CLIC1*, *HRAS*, *TNF*, and *MYH11* (one of the ACMG 59 genes in which incidental sequencing findings should be reported) (Green et al., 2013). Such absence was unexpected given the UKB exome sample size, as these genes were not under severe evolutionary constraint (Samocha et al., 2014), and protein-coding variants had been called for these genes in other databases (Lek et al., 2016), some of which were present at sufficiently high frequency to be included on genotyping arrays. We reasoned that the lack of variant calls in these genes was unlikely to be explained by ascertainment of a unique population in the UKB (i.e., the variants truly did not exist), and was instead caused by a technical error in sequencing, data processing, variant calling, or a combination of these.

To prove that the missing variants are indeed present in the UKB population, we first evaluated the internal consistency between the genotyping and exome-sequencing data that had been collected for the same UKB samples. We identified a

total of 30,979 common variants (MAF > 0.01) in the UKB dataset that overlapped the sequenced exons that had also been ascertained in 49,909 samples by genotyping arrays (see Methods). While the majority of variants had been called by both methods (24,614 variants, 79.5%), a substantial minority (6,365 variants, 20.5%) were called by the genotyping arrays but not by exome sequencing (Figure 1). This discrepancy included many common variants with MAFs close to 0.5 (i.e., that were present in almost 50% of the array samples) providing strong evidence that the exome-sequencing genotype calls are missing variants that are actually present in the sequenced samples and should have been detected in this population.

We next examined variant calls aggregated per gene in the UKB exomes in comparison to the Genome Aggregation Database (gnomAD v2.1.1, 125,748 sequenced exomes; see Methods) (Karczewski et al., 2019). Our analysis focused on the exons sequenced both in UKB and gnomAD, which encompasses 23,040 human genes (see Methods; Figure 2a).

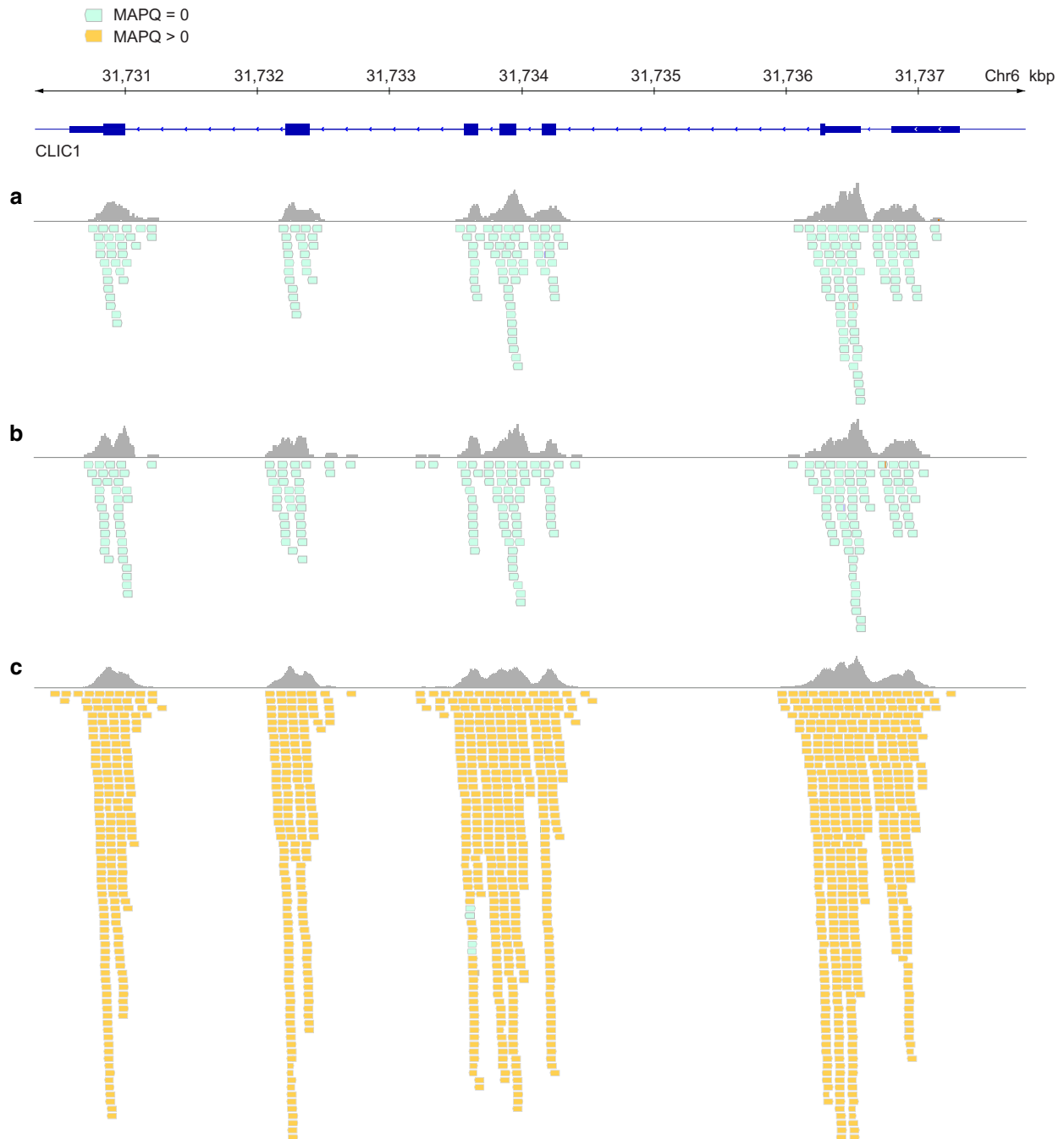


FIGURE 3 UK Biobank (UKB) exome read alignments at the CLIC1 genomic locus. (a) Alignments obtained from the UKB exome release. (b) Realignment of reads without alternative contigs index file. (c) Corrected alignments with proper indication of alternative contigs. Read alignments to the human genome are visualized with an Integrated Genomics Viewer (GRCh38, IGV v2.6.2) [Colour figure can be viewed at wileyonlinelibrary.com]

We found that, for most genes, the number of variants in gnomAD was well predicted by the number in UKB, with expected 1:2.3 proportionality given the larger gnomAD sample size (Figure 2b). However, this analysis also highlighted 641 genes with 0 variants called in the UKB exomes, versus a median of 286 variants (range of 1 to 14,291) in gnomAD (Supporting Information Table S1). Using the aggregate

observed variant frequency per gene in gnomAD, we calculated the probability for at least one variant being observed in the UKB exome sample for each gene. Of the 641 genes, 598 (93%) should have had at least one variant identified (95% CI one-tailed binomial distribution). Given that the UKB is a predominantly European ancestry population and the gnomAD dataset contains a more diverse population,

we performed ancestry-specific analysis (Figure 2c) of these genes in gnomAD. The largest number of variants in these genes were found in the European ancestry samples as expected by their majority representation in the gnomAD dataset. This excluded the possibility that some or all of the genes lacking variation in the UKB was due to ancestry-specific variation.

To understand the reason for these missing variant calls in the UKB, we analyzed the sequencing read data, provided by the FE pipeline, for individual exomes at the 641 loci. Our analysis indicated that, despite having reads mapped to these genes (Figure 3a), the mapping quality (MAPQ) score was zero in many cases, causing these reads to be eliminated from the downstream procedures for variant calling. The MAPQ field in the SAM specification (Li et al., 2009) is the PHRED scaled probability (Ewing & Green, 1998) of the alignment being erroneous. In practice, however, each aligner treats the MAPQ field differently. With the aligner BWA-MEM (Li, 2013) used in the FE pipeline, a MAPQ score of zero is given to reads that align equally well to more than one genomic location. Thus, it is typically an indicator of reads that come from duplicated or repetitive regions of genomic DNA. However, many of the loci we individually examined were not known to harbor repetitive elements or reside in regions of genome duplication. Investigating further, we found that the zero MAPQ scores were due to the reads showing multiple alignments to the GRCh38 genome reference, not to repetitive elements but to so-called “alternative contigs” (ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa). As of this genome release, alternative contigs are used frequently to represent divergent haplotypes that cannot be easily captured by a single linear sequence. Indeed, of the 598 genes with high probability of missing variation, 568 (95%) had alternative contigs represented in the genome reference (Supporting Information Table S1).

Starting from the raw reads available from CRAM files, we found that the original read alignment provided by the UKB (Figure 3a) was most closely reproduced when performing the alignment under default alignment parameters (BWA-MEM; see Methods). This alignment (Figure 3b) does not take into consideration alternative contigs in the absence of an index file specifically marking these contigs; it treats them instead as independent genomic regions equal to primary contigs. Reads that map to both primary and alternative contigs are therefore interpreted as mapping to multiple genomic locations at these loci. We found that realigning the raw reads while providing the alternative contig index file for the genome reference resulted in a dramatic increase in the number of reads that properly mapped to those loci and increased their MAPQ scores well above zero (Figure 3c).

4 | DISCUSSION

We have found that genetic variants documented in the UKB FE release are conspicuously absent from certain genes in a manner that is best explained by errors of read alignment. Furthermore, while our analysis has focused on 641 genes with an absolute lack of variant calls, additional genes may have partially duplicated or repetitive sequences such that they are missing substantial (but above zero) variation beyond those identified in our short study (2391 genes are currently contained within alternative contig representations of the genome). Thus, the variant calls in the current UKB exome data should not be used for large-scale genomic analyses, as only genes without alternative haplotypes are unaffected by the erroneous alignment. We notified the UKB bioinformatics team of the bug and our proposed patch, and UKB acknowledged the error, accepted the solution, and provided a comprehensive list of affected regions (<https://www.ukbiobank.ac.uk/wp-content/uploads/2019/12/Description-of-the-alt-aware-issue-with-UKB-50k-WES-FE-data.pdf>). The UKB further retracted the exome data release and will implement the correction in an upcoming larger release of 150,000 exomes in 2020. For investigators using the current dataset, we provide a description of and protocol for read realignment (Supplemental File) that we hope others will find useful for generating corrected alignment files, which can then be used to generate accurate genotype calls with downstream variant calling pipelines. For researchers in general, utilizing population scale genetic data like the UKB for wide-ranging applications, including evaluation of variants at a single locus, we propose the simple heuristic of comparing unique variant counts per gene across all genes in the new data release with a large reference data set (as we did with gnomAD; Figure 2b) from summary variant count files, which will quickly reveal systematic biases in variant identification. This heuristic should be robust for large samples in outbred populations across ethnic groups given the common, recent origin of most protein-coding variants (Tennessen et al., 2012).

This study highlights the need for the community of genetics investigators to continually evaluate data processing protocols for the UKB and other large genomic resources, sharing concerns in a transparent and timely manner. To facilitate quality control of data processing and releases, we recommend that a README file is attached to all processed data detailing the processing commands and parameter settings for data generation. The datasets and README files would benefit from version control, which will enhance communication and reproducibility. The now retracted exome-sequencing data was available to approved researchers for 9 months since its first release in March 2019, impacting the work of at least 30 research groups around the world

who obtained access and likely many more local scientific collaborators of those groups. Noting that other researchers (Michael Weedon, personal communication) had flagged a systematic lack of variant calls months earlier in the UKB mailing list archives, we believe that use of a more open, searchable community forum could have led to identification and patch of the exome data-processing error earlier. As tasks like sequence alignment and variant calling are very computationally expensive, robust centralized sequence data-processing protocols are critical for enabling the use of such resources by the wide-ranging research community—particularly as UKB prepares to expand the initial exome release to 500,000 whole genomes over the next few years.

ACKNOWLEDGMENTS

We are grateful to Dr. Olivia Osborne for helpful discussions and to William Markuske for support with high-performance computing. This work was funded by grants from the National Institute on Drug Abuse and the National Human Genome Research Institute (P50 DA037844 and R01 HG009979 to T.I.) as well as the National Institute for Diabetes, Digestive and Kidney diseases (K08 DK102877-01 and R03 DK113328-01 to A.R.M.) and a UCSD/UCLA Diabetes Research Center grant (P30 DK063491 to A.R.M.).

CONFLICT OF INTEREST

T.I. is cofounder of Data4Cure, is on the Scientific Advisory Board, and has an equity interest; in addition, T.I. is on the Scientific Advisory Board of Ideaya BioSciences, has an equity interest, and receives sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies.

AUTHOR CONTRIBUTIONS

T.I. and A.R.M. designed and directed the study. T.J. executed all analyses and made the figures with help from B.M. T.J., T.I., and A.R.M. wrote the manuscript with assistance from H.L.A. All authors read, edited, and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Tongqiu Jia  <https://orcid.org/0000-0002-6012-6436>

REFERENCES

- Abul-Husn, N. S., Cheng, X., Li, A. H., Xin, Y., Schurmann, C., Stevis, P., ... Dewey, F. E. (2018). A protein-truncating HSD17B13 variant and protection from chronic liver disease. *The New England Journal of Medicine*, 378(12), 1096–1106.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498.
- Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II: Error probabilities. *Genome Research*, 8(3), 186–194.
- Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., ... American College of Medical Genetics and Genomics. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 15(7), 565–574.
- Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., ... Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research*, 34(Database issue), D590–D598.
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210. Retrieved from <https://doi.org/10.1101/531210>
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997. Retrieved from <http://arxiv.org/pdf/1303.3997.pdf>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
- Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., ... Stamatoiyannopoulos, J. A. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics*, 28(14), 1919–1920.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.
- Regier, A. A., Farjoun, Y., Larson, D. E., Krasheninina, O., Kang, H. M., Howrigan, D. P., ... Hall, I. M. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications*, 9(1), 4038.
- Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., ... Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9), 944–950.
- Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., ... on behalf of the NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding

- variation from deep sequencing of human exomes. *Science*, 337(6090), 64–69. Retrieved from <https://doi.org/10.1126/science.1219240>
- Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. X., Ye, B., Pandey, A. K., ... on behalf of the Regeneron Genetics Center. (2019). Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv*, 572347. Retrieved from <https://doi.org/10.1101/572347>
- Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T., & Wright, C. F. (2019). Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: Implications for direct-to-consumer genetic testing. *bioRxiv*, 696799. Retrieved from <https://doi.org/10.1101/696799>
- Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., ... Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, 44(D1), D710–D716.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Jia T, Munson B, Lango Allen H, Ideker T, Majithia AR. Thousands of missing variants in the UK Biobank are recoverable by genome realignment. *Ann Hum Genet.* 2020;1–7. <https://doi.org/10.1111/ahg.12383>