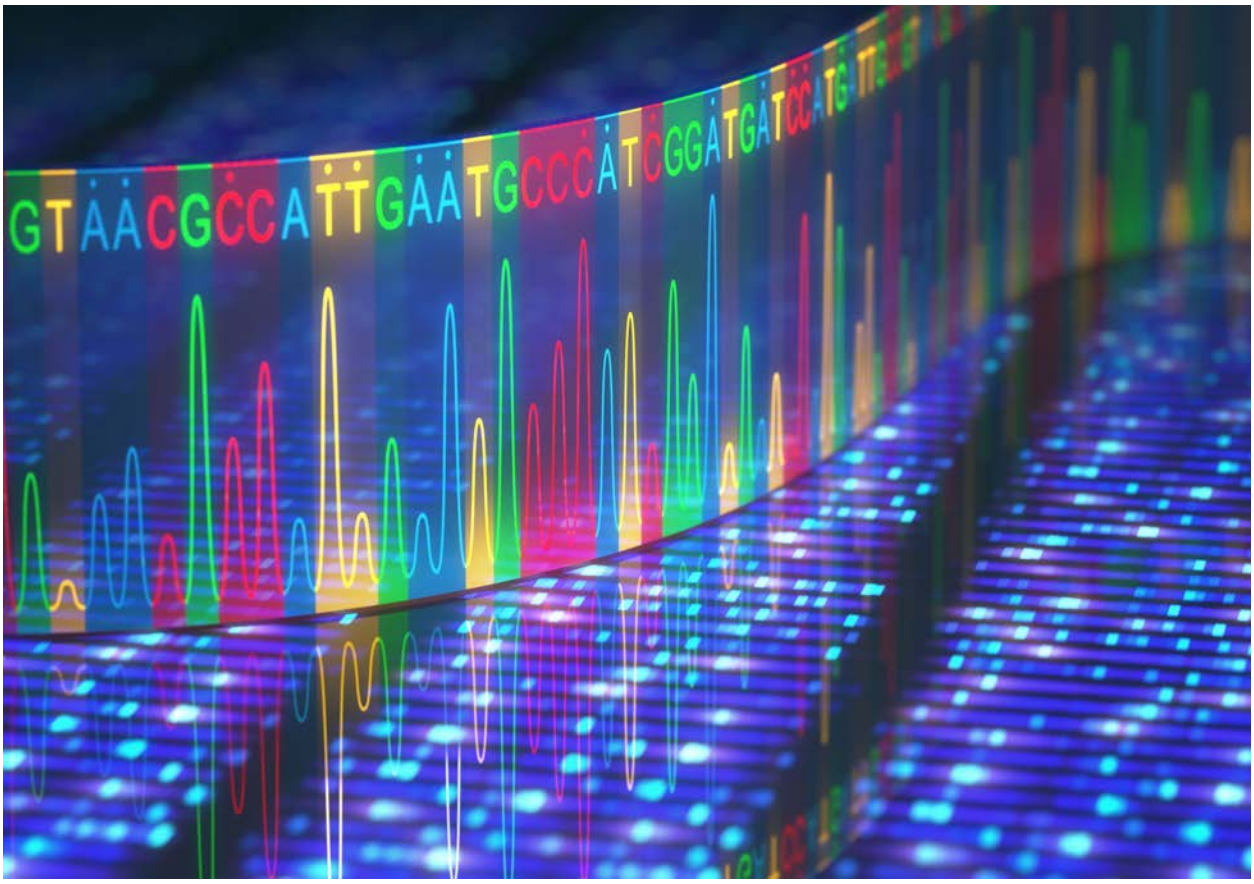

Considerations Toward a Comprehensive Genomics Strategy

The Genomics Working Group Report to the *All of Us* Research Program Advisory Panel

December 5, 2017

Roster

Lon Cardon, Ph.D. (co-chair)
BioMarin Pharmaceutical Inc. (*All of Us* Advisory Panel Member)

Debbie Nickerson, Ph.D. (co-chair)
University of Washington

Gonçalo Abecasis, D.Phil.
University of Michigan

Wendy Chung, M.D., Ph.D.
Columbia University

Joshua Denny, M.D.
Vanderbilt University

Caroline Fox, M.D., M.P.H.
Merck Research Laboratories

Eric Green, M.D., Ph.D.
National Human Genome Research Institute

Joel Hirschhorn, M.D., Ph.D.
Boston Children's Hospital

Howard Jacob, Ph.D.
HudsonAlpha Institute for Biotechnology

Sekar Kathiresan, M.D.
Eli and Edythe L. Broad Institute of MIT and Harvard

Eimear Kenny, Ph.D.
Icahn School of Medicine at Mount Sinai

Christopher O'Donnell, M.D.
VA Boston Healthcare System

Brad Ozenberger, Ph.D.
All of Us Research Program

Heidi Rehm, Ph.D.
Partners HealthCare

Jay Shendure, M.D., Ph.D.
University of Washington

Alan Shuldiner, M.D.
Regeneron Pharmaceuticals, Inc.

Karl Surkan, Ph.D.
Massachusetts Institute of Technology

Sharon F. Terry, M.A.
Genetic Alliance (*All of Us* Advisory Panel Member)

Table of Contents

Report of Genomics Working Group
of the *All of Us* Research Program Advisory Panel

Executive Summary

The Genomics Working Group (GWG) of the *All of Us* Research Program Advisory Panel was tasked with considering possible strategies for deploying a comprehensive genomics platform for the program. The group discussed technologies available for large-scale genomic data generation and their costs, capturing genomic data on the diverse base of participants in the program, and the *All of Us* Research Program's responsibility to make information collected in the study available to participants. It was the opinion of the members of the GWG that the program may benefit from a phased approach to developing a genomics strategy. Accordingly, the group contemplated a pilot study that would allow *All of Us* to assess processes and methods to evaluate return of information burdens in preparation for its scale-up to at least 1 million participants.

The GWG considered the major genomic data generation platforms—whole genome genotyping (WGG), whole exome sequencing (WES), and whole genome sequencing (WGS)—and their potential application for the full scale of the program. The GWG determined that both WGG *and* WGS would be important for the program to evaluate to enable subsequent determination of its most appropriate data platforms and data generation timelines. In incorporating research value and potential for return of information, the group also agreed that WGS offers sufficient research advantages over WES, at costs only incrementally greater, to warrant evaluation in a pilot phase without investment in WES.

After examining the advantages of different models of scale for a potential pilot phase, the GWG landed on an ideal size of approximately 5% of the eventual 1 million–participant goal, or 50,000 individuals. The GWG considered this sample size sufficient for testing pipelines, evaluating data types, capturing variants across the diverse participant base of the program, and assessing return-of-information strategies. This approach would provide a sufficient scale for beginning to examine the power of the *All of Us* platform to integrate genomic data with environmental, lifestyle, and medical record data, thereby demonstrating the program's potential to advance our understanding of health and disease.

Goals of the *All of Us* Research Program

The Precision Medicine Initiative defined precision medicine as "an approach to disease treatment and prevention that seeks to maximize effectiveness, accounting for individual variability in genomes, environment, and lifestyle." Precision medicine seeks to refine our understanding of causes of disease, disease onset, treatment response, disease progression, and other health outcomes through the precise measurement of molecular, environmental, and behavioral factors. This understanding may lead to more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of new treatments. Coincident with advancing the science of medicine is a changing culture of medical research that engages individuals not just as patients or research subjects but as active partners. The *All of Us* Research Program believes that a combination of highly engaged participants and rich biological, health, behavioral, and environmental data will usher in new possibilities in precision medicine.

The mission of *All of Us* is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all. The ultimate objective of the program is to build a robust research resource that can facilitate exploration of biological, clinical, social, and environmental determinants of health and disease. *All of Us* will collect and curate health-related data and biospecimens from 1 million or more individuals who reflect the diversity of the United States. Through a transparent process that ensures confidentiality and proper use, these data and biospecimens will be made broadly available for research use. The program seeks to achieve this through nurtured relationships with participant partners and by delivering the largest, richest biomedical data set ever, catalyzing a robust ecosystem of researchers and funders eager to use and support it. The *All of Us* Research Program should have sufficient scale and overall scope to enable research for a wide range of diseases, providing insights into individualized decision-making for patients of varying health status and environment.

Purpose, Approach, and Activities of the Working Group

Purpose

The Genomics Working Group (GWG) was convened as an advisory group to the *All of Us* Research Program Advisory Panel. The GWG first met on June 27, 2017, when *All of Us* director Eric Dishman presented the group with its charge, one that reflected the multiple goals of *All of Us*:

1. Develop a data set of genomic variation for all stakeholders, including participants and researchers.
2. Create a valuable genomics resource for discovery, advancing knowledge and of direct utility to the participant should s/he opt to receive individual genomic information.
3. Set a foundation for future genomic testing.

Approach

In their consideration of what a comprehensive genomics strategy for *All of Us* may entail, the members of the GWG agreed that a pilot study may be the best approach to inform the program's ability to scale while simultaneously enriching its data platform for discovery. This approach would also allow the program to test participants' needs and perspectives regarding the return of individual genomic information. Given the evolving nature of genome sequencing technologies and the annotation and interpretation of sequence variation, members of the GWG also agreed that a pilot study would inform optimization of future genomic strategies in the program.

While considered in context and acknowledged through the group's work, specific focus on major ethical, legal, and societal implications (ELSI) was deferred to a separate committee or task force within the program's governance structure. The GWG also deferred specific discussions on return of genomic results to participants.

Activities

The working group discussed the rapidly evolving nature of genomic sequencing and variant annotation approaches while considering the commitment of *All of Us* to return information to participants. The GWG evaluated genomic technology platforms, types of data to be generated, analysis approaches, and clinical and research utility to inform the development of a comprehensive genomic strategy for the program. The group considered the following questions:

1. What is the current state of sequencing and array-based assays at scale?
2. Which genomic assay has the best value for participants and researchers? Which has the best balance of features for both?
3. What are potential approaches for incorporating genomic sequencing, considering different cost models?

The GWG met seven times through WebEx from July 2017 through October 2017, creating two task groups—pilot study proposal and DNA genotyping array—which met separately to accomplish and coordinate their work.

Options for a Genomics Strategy at Scale

The Need for a Pilot Study: Rationale and Goals

Genomic data will be a key component of the *All of Us* Research Program's overall platform and discovery potential. This research should prove valuable to various stakeholders and will be an engagement tool for participant partners, providing them with access to their own research data, including genomic information. Considering these goals and the expansive scale of the project, group members agreed that an initial pilot study could provide the necessary information to design an optimal genomics strategy for the full program.

Genomics remains a field of rapidly evolving technologies, and currently there are multiple options with respect to the generation of genomic data: whole genome genotyping (WGG) using a high-density microarray, whole exome sequencing (WES), and whole genome sequencing (WGS). Each approach yields a different set of data and comes at a different cost. The group felt that these options should be considered at a sufficient scale of participants to inform the value of genomic data to the diverse needs of the stakeholders before any ultimate strategy decisions were made for the entire 1 million–participant cohort.

To account for genomic data options, sample sizes, implications for data analysis and sharing, and possible designs for a phased approach, the GWG considered the following high-level goals for a potential pilot phase:

1. Create a genomic data set for both genotyping and sequencing data types (within a CLIA-certified laboratory).
2. Develop a workflow involving genomic data generation, genome interpretation, data visualization, limited return of genomic results, and potential hand-off to clinical care.
3. Develop a substantive research resource for internal and external investigators to examine resultant genomic data and consider value of different types.

Questions Informing a Pilot Study

Given these goals for a potential pilot study, the GWG contemplated several key questions to inform a pilot design.

1. What are the major strategies for detecting variation in the human genome?

Broadly speaking, there are four commonly used approaches for detecting variation in the human genome (**Table 1**):

a. Whole genome genotyping (WGG) analyzes human DNA by using genome-wide genotyping arrays together with imputation methods that leverage reference haplotypes like those in the Haplotype Reference Consortium. WGG allows for identification of most of the common DNA sequence variants in a given genome (i.e., the variants in each genome that are commonly shared with other individuals). Imputation is a statistical approach that uses the correlation structure among single nucleotide variants (SNVs) to infer genotype at a site not directly assayed.

b. Whole exome sequencing (WES) uses hybrid selection to isolate the exome and short-read DNA sequencing to identify both common and rare genomic variants in the captured regions. The exome represents approximately 1% of the genome (coding exons and splice sites, as well as untranslated exons and proximal promoters, if desired).

c. Whole genome sequencing (WGS) uses short-read DNA sequencing technologies to determine both common and rare genomic variants across the whole genome and enables comprehensive assessment of SNVs, insertion/deletion polymorphisms, and structural

variants (SVs) across the genome, except for the most repetitive and least accessible portions of the genome.

d. Whole genome sequencing plus (WGS+) employs more expensive DNA sequencing technologies to determine genomic variants, like larger SNVs, that are currently not well detected by the approaches above. Since WGS+ costs tens of thousands of dollars per sample, this approach was not considered further as an option for a pilot study.

Below are the group's considerations of the first three approaches with respect to detection of different types of genomic variants and cost:

- WGG with imputation (using extant whole genome sequence data) can accurately characterize SNVs in coding or non-coding sequence down to 0.5% frequency and, in some ancestral populations, as low as 0.05%. If used alone, WGG with imputation will miss rare variants in populations currently underrepresented by whole genome sequence data. Commercially available WGG arrays can be supplemented with custom variant content of interest to *All of Us*, including pharmacogenetics markers, clinically relevant variants, and others.

- WES can accurately characterize SNVs across the full range of allele frequencies, including those private to an individual (i.e., not yet observed in prior studies), but analysis is limited to coding sequence representing roughly 50 million of the 3.2 billion bases of human genome sequence.

- WGS can accurately characterize SNVs across the full range of allele frequencies in both coding and non-coding sequence. At present, our ability to interpret variation in coding sequences for clinical and biologic meaning far exceeds that for variation in non-coding sequences. However, the availability of WGS and phenotypes in a large, diverse cohort such as *All of Us*, combined with experimental data sets of epigenetic and gene expression, may allow the research community to narrow this knowledge gap.

SVs—small insertions/deletions, larger insertions/deletions, rearrangements, and duplications—are more challenging than SNVs to identify with all three approaches (**Table 1**), but they are most effectively determined by WGS methods.

Cost assessments were gathered only for initial genomic data generation and did not include costs for laboratory validation of variants, interpretation, or return of results. Costs of WGG (i.e., microarray chip and sample processing) range from $30 to $100 per sample. Costs of WES in a U.S. Clinical Laboratory Improvement Amendments (CLIA)–certified laboratory range from $350 to $1,000 per sample. Costs of WGS in a CLIA-certified laboratory range from $1,000 to $2,000 per sample. On average, the cost of WGS is currently three to four times more than WES; however, the GWG anticipates that this difference will narrow considerably in the next few years, to the point that the pricing of WGS per sample might approach the cost of WES.

Table 1. Approaches to characterizing DNA sequences in an *All of Us* participant			
Approach	WGG	WES	WGS
Cost range per participant for genomic data generation in a CLIA facility	$ $30–$100	$$ $350–$1,000	$$$ $1,000–$2,000
Single Nucleotide Variants			
Common	Yes	Yes	Yes
Rare (<1% frequency), coding	No	Yes	Yes
Rare (<1% frequency), non-coding	No	No	Yes
Structural Variants			
Small, common	Partial	Partial	Partial
Small, rare	No	No	Partial
Large	Partial	Partial	Yes

2. What are the options for the design of a custom genotyping array?

The group considered the current WGG products on the market and options for design of custom variant content tailored to the goals of the program. It was noted that in large contemporary cohorts like the UK Biobank, the Million Veteran Program, and 23andMe, WGG technology has been the initial genetic modality of choice for several reasons. In comparison to current genome sequencing platforms, WGG technology has shorter data generation times and considerably lower informatics overhead. The quick turnaround time could enable identification of problems with data flow (e.g., issues with sample tracking or data quality) early in the program. WGG arrays are low cost—approximately five to 10 times cheaper than WES or WGS at the moment—and have been tested in tens of millions of participants. Ongoing advancements in imputation strategies have rapidly improved the recovery of rare variants in WGG data sets, and ancestry detection algorithms can be an important means of engaging the public with genetic data.

Several array platforms were highlighted by members of the group as designed for optimal performance in multiethnic populations, representative of the diverse participants expected to enroll in *All of Us*. It was noted that WGG arrays are increasingly closing the gap between research and clinical utility; several industry and academic groups have partnered with array vendors to design custom clinical content. By targeted improvement of oligonucleotide design and calling pipelines, WGG arrays can reliably call many pharmacogenetic variants, human leukocyte antigen alleles, variants in the American College of Medical Genetics and Genomics (ACMG) list of clinically actionable genes, and other sites that are of high value for research and clinical applications. For example, 23andMe is currently approved by the U.S. Food and Drug Administration for return of results of customized content on the Illumina Global Screening Array (GSA) for 10 diseases, including hereditary hemochromatosis, celiac disease, and Alzheimer's disease. It should be noted, however, that WGG technology poses several challenges for calling very rare variants and copy-number sites. Although arrays can be prepared with custom content, adding new content increases costs and adds significant time for design and

production. For this reason, the GWG considered it more expedient for a potential pilot to pursue an existing array technology without significant custom content. However, customization of an existing array platform could be beneficial for the larger implementation of the program.

3. What are considerations for the return of results, given different platforms and strategies?

The GWG identified several factors to consider in choosing a platform and strategy for the return of genomic results:

- Total number of participants receiving results: Prior studies report a 1% to 3% frequency of findings from a set of 59 genes designated by ACMG as clinically actionable.
- Pathogenic versus likely pathogenic (LP) variants: The ACMG recommendations suggest the return of pathogenic variants only, not LP, although many programs also include LP. A decision should balance increased sensitivity and ability to offer more results to participants, with the downside of participants receiving a higher proportion of variants that may later be revealed to be benign.
- Expert review versus manual review: To reduce the burden of manual variant review to determine pathogenicity and return-of-results support, consideration for use of only variants approved as three-star status in ClinVar could be considered, but this may lead to the return of fewer results to participants from minority populations. Furthermore, return of results could focus on the variants designated most actionable by the ACMG (e.g., cancer or cardiomyopathy).
- CLIA versus non-CLIA platforms: Use of a CLIA platform would allow more results to be returned directly, potentially without requiring costly validation. This is particularly true for future considerations of high-volume return of results, including carrier status, pharmacogenomics, complex trait polygenic risk scores, and more.
- Arrays versus sequencing platforms: There are now well-established quality control metrics for validating the accuracy of sequencing data for variant detection. Arrays for rare variant detection have not been as widely validated for rare variant detection, so more data are needed to understand the accuracy of array data for rare variant calling. Arrays may require independent validation unless specific variants have been validated.

The proportion of participants who have an actionable, returnable result is something that *All of Us* would need to calibrate based on the above trade-off considerations.

4. What is the right sample size for a pilot study?

One major question the GWG considered was the appropriate sample size for a potential pilot study. The study would have to be large enough to assess the value of different options for generating genomic data and to test various aspects of workflow, logistics, and the creation of usable data sets. On the other hand, it could not be so large that it overwhelmed the program or prevented the early learning and evaluation required to make the necessary adjustments to

produce the most useful data at scale. Each of these represent challenges for an ambitious program enrolling 1 million or more participants.

The genomic data set that a pilot study could generate would represent one of the first substantive *All of Us* resources with broad utility for the research community. As WGG, WES, and WGS each have different capabilities and costs, GWG members felt that it would be prudent to compare the relative value of all three in the same cohort to determine the best strategy for *All of Us* at scale. If properly generated and made easily accessible, this resource would likely attract researchers interested in utilizing the program's platform to study genetic research questions. For this reason, any initial genomics pilot the program may initiate must be compelling and viewed as a substantive advance. In contrast, if a genomics pilot were seen as unimpressive in scale or poor in quality, this could harm the reputation of the *All of Us* at a critical early stage.

Two additional considerations included the potential for scientific learnings within a potential pilot study and the ability to test the return-of-results workflow at a range of recruitment centers. For example, the Advisory Committee to the NIH Director Precision Medicine Working Group delivered a report, *The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine*, which identified determining the clinical impact of loss-of-function mutations as a scientific opportunity. The report quoted an approximate carrier rate of 1 in 250 for loss-of-function mutations. The number of carriers for a given loss-of-function variant would scale with a pilot study sample size (e.g., 20, 200, and 400 carriers for study sample sizes of 5,000, 50,000, and 100,000).

From a return-of-results perspective, if 1% of participants are expected to carry actionable mutations in ACMG genes, return would be expected in 50, 500, and 1,000 participants for sample sizes of 5,000, 50,000, and 100,000, respectively. If the 500 participants are distributed across 10 or more sites, each site would be able to test the return-of-results workflow in approximately 50 participants.

Table 2: Current and emerging biobanks to facilitate genome–phenome studies				
Biobank	**Enrollment Locations**	**Enrollment Start Date**	**Enrollment to Date**	**Genomics Strategy**
Commercial Funding				
deCODE Genetics (Amgen)	Iceland	1996	> 300,000	WGG, WGS
DiscovEHR (Regeneron Genetics Center)	Geisinger Health System (Danville, Pennsylvania)	2007	> 200,000	WGG, WES
Government Funding				
China Kadoorie Biobank	China	2004	> 500,000	WGG
UK Biobank	United Kingdom	2006	> 500,000	WGG, WGS
Million Veteran Program	VA Hospitals	2011	> 500,000	WGG, WGS
Electronic Medical Records and Genomics (eMERGE) Network	Various academic medical centers in the United States	2007	>100,000	WGG, targeted sequencing of 59 ACMG genes
Institutional Funding				
BioVU	Vanderbilt University Medical Center (Nashville, Tennessee)	2007	> 215,000	WGG
Kaiser Permanente Research Bank	United States	2016	> 250,000	WGG

Based on the above considerations, GWG members agreed that a potential pilot study should be at least comparable in size to previous projects with similar data and should aim to achieve goals that were not readily feasible with similarly sized data sets. **Table 2** provides a summary of selected large-scale genotype–phenotype data resources. The group noted that WGG data are available in population samples on the order of hundreds of thousands of individuals with phenotypic data, so use of array information would need to be focused on facilitating goals (such as return of results) that had not been attempted at scale in other projects. With regard to sequence data, the 1000 Genomes Project provided genomic sequence data on about 2,500 individuals, but without any associated phenotypic information. A more analogous project to *All of Us* is the National Human Genome Research Institute's (NHGRI) eMERGE Network, which has access to data in electronic health records, nearly 100,000 individuals with WGG, and about 25,000 participants with sequencing data for the ACMG clinically actionable genes. For *All of Us* to generate a data set (even at a pilot phase) that would represent a true advance over other existing programs, it should aim to exceed eMERGE in scope and scale.

Using this logic, members of the GWG reasoned that a potential pilot study of 5,000 individuals would be too small, 10,000 to 25,000 would be acceptable, and 50,000 would be impressive. Based on this reasoning, the GWG believes that a potential pilot with a target of 50,000 participants could generate as much enthusiasm for the program as possible at an early and critical stage. This size would also provide for robust testing of data generation and return-of-information workflows at a substantive scale, as noted previously.

Final Considerations for a Pilot Study

Should the program decide to pursue the design of a pilot, the GWG presents the following goals for consideration: (1) test the generation, processing, analysis, interpretation, and sharing of genomic data in *All of Us* at a scale sufficient to enable planning for a genomic strategy for the full cohort; (2) pilot the return of clinically significant results to participants at a scale sufficient to inform the development and refinement of the return-of-results strategy and process; and (3) provide a genotype and phenotype resource that will add value above other existing genomic data sets and make this resource widely available to the research community in an efficient and non-burdensome manner.

A sample size representing approximately 5% of the expected 1 million participants (i.e., a pilot study size of 50,000 participants), incorporating genomic data generation using both WGG and WGS in each participant, may be the best option for a pilot study. Additionally, the program may consider using an existing array technology, perhaps with modular customization if expedient. As discussed above, WGG will allow for rapid data generation at lower cost and testing of a scenario in which all 1 million participants receive an array. WGS of the same samples will allow for the evaluation of the incremental scientific and clinical value of both coding and non-coding sequence variants beyond what is obtained through WGG, especially for rare variants across diverse populations, and could be valuable for discovery and for return of results.

While considering the design of a potential pilot study, the GWG discussed a workflow including at least five elements: genomic data generation, genome interpretation, data visualization, return of results to participants, and a handoff to clinical care for participants with clinically actionable results. Implementation of this workflow might require the assignment of these elements to different partners based on capabilities; some contributors may be capable of integrating multiple elements.

In its evaluation of costs, the GWG determined that a potential pilot study of 50,000 individuals would likely exceed $100 million. This cost could represent a valuable investment, as exploration of the interplay among genome, environment, and lifestyle has the potential to transform our understanding and treatment of disease and engage *All of Us* participant–partners in their health.

References

1. Precision Medicine Initiative. (2015). The Precision Medicine Initiative. Retrieved from https://obamawhitehouse.archives.gov/precision-medicine
2. Precision Medicine Initiative. (2015). Precision Medicine Initiative Cohort Program—Building a Research Foundation for 21st Century Medicine. Retrieved from https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf
3. *All of Us* Research Program. (2017). Genomics Working Group of the *All of Us* Research Program Advisory Panel. Retrieved from https://allofus.nih.gov/genomics-working-group-all-us-research-program-advisory-panel
4. Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., … Biesecker, L. G.; American College of Medical Genetics and Genomics. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine, 15*(7), 565–574.
5. Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., … Carey, D. J. (2016). Distribution and clinical impact of functional variants in 50,726 whole exome sequences from the DiscovEHR study. *Science, 354*(6319). pii: aaf6814.
6. Natarajan, P., Gold, N. B., Bick, A. G., McLaughlin, H., Kraft, P., Rehm, H. L., … Green, R. C. (2016). Aggregate penetrance of genomic variants for actionable disorders in European and African Americans. *Science Translational Medicine, 8*(364), 364ra151.
7. Khera, A.V., & Kathiresan, S. (2017). Genetics of coronary artery disease: Discovery, biology and clinical translation. *Nature Reviews: Genetics, 18*(6), 331–344.
8. European Bioinformatics Institute. (2016). About ISGR and the 1000 Genomes Project. Retrieved from http://www.internationalgenome.org/about