Published in final edited form as:
Hum Genet. 2019 July ; 138(7): 691–701. doi:10.1007/s00439-019-02033-5.

Mind the Gap: Resources required to receive, process, and interpret research-returned whole genome data

Dana C. Crawford[1,2,3,*], **Jessica N. Cooke Bailey**[1,3,*], **Farren B. Briggs**[1,3,*]

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio 44106

[2]Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, Ohio 44106

[3]Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio 44106.

Abstract

Most genotype-phenotype studies have historically lacked population diversity, impacting the generalizability of findings and thereby limiting the ability to equitably implement precision medicine. This well-documented problem has generated much interest in ascertainment of new cohorts with an emphasis on multiple dimensions of diversity, including race/ethnicity, gender, age, socioeconomic status, disability, and geography. The most well-known of these new cohort efforts is arguably All of Us, formerly known as the Precision Medicine Cohort Initiative Program. All of Us intends to ascertain at least one million participants in the United States representative of the multiple dimensions of diversity. As an incentive to participate, All of Us is offering the return of research results, including whole genome sequencing data, as well as the opportunity to contribute to the scientific process as non-scientists. The scale and scope of the proposed return of research results are unprecedented. Here, we briefly review possible return of genetic data models, including the likely data file formats and modes of data transfer or access. We also review the resources required to access and interpret the genetic or genomic data once received by the average participant, highlighting the nuanced anticipated barriers that will challenge both the digitally, computationally literate and illiterate participant alike. This inventory of resources required to receive, process, and interpret return of research results exposes the potential for access

Corresponding Author: Dana C. Crawford, Case Western Reserve University, 2103 Cornell Road, Wolstein Research Building, Suite 2-527, Cleveland, OH 44106, (216)-368-5546, dana.crawford@case.edu.

[*]These authors contributed equally to this work

Author Contributions
JNCB, DCC, and FBB contributed equally to this work, including conception of the review topic, drafting the work, and substantively revising the work. All authors approve the submitted version and all authors agree to be personally accountable for the work.

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF hie of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Conflicts of Interest
The authors declare no conflict of interest.

disparities and warns the scientific community to mind the gap so that all participants have equal access and understanding of the benefits of human genetic research.

Keywords

Precision medicine; return of research results; All of Us; diversity; digital literacy; whole genome sequencing; third-party interpretation services; direct-to-consumer genetic testing

Genetics is mainstream

The idea of genetic data is popular among non-scientists, but its understanding and consequences are poorly understood. Coverage of core genetic concepts in science, technology, engineering, and math (STEM) education is highly variable (Dougherty et al. 2011; Lontok et al. 2015) and genetic literacy is persistently low among the United States (US) general population (Haga et al. 2013). Despite the technical chasm, persistent pop cultural references to genetics and genomics in both entertainment and the media have elevated and maintained the voguishness of these data. In the movies, the *Jurassic Park* franchise launched in the early 1990s spurred an interest in ancient DNA and genetic engineering while the now cult classic *Gattaca* (1997) dealt with genetic determinism and genetic discrimination among other bioethical issues (Müller and Dalzotto 2018), prescient of today's genetic testing and gene editing capabilities. Examples in television include the popular "CSI: Crime Scene Investigation" (2000-2015), which since the mid-oughts has popularized forensics and spurred debates on its lasting impact on real-life jurors (Schweitzer and Saks 2007). More recently, genetic ancestry and genealogy have taken a starring role as a tool for self-discovery on both television ("Who Do You Think You Are" in partnership with Ancestry.com (2010-12; 2013-present) and "Finding Your Roots" (2012-present)) and social media (e.g., YouTube).

In contrast to the idea of "genetics," most non-scientists are not yet aware of "precision medicine" research (Ray 2018). Non-scientists, however, are aware of the availability of tailored prevention strategies or treatment plans from celebrity-penned op-eds and testimonials such as Angelina Jolie and her decision to have a prophylactic double mastectomy based, in part, on family history and genetic testing results (Desai and Jena 2016; Jolie 2013; Liede et al. 2018). Like the terms "precision medicine" or "personalized medicine", the Precision Medicine Initiative Cohort Program, now known as All of Us, is unknown to most surveyed (Ray 2018). In comparison, awareness of personal genetic testing through direct-to-consumer (DTC) genetic tests has steadily increased since becoming available in 2006, albeit with noted differences by geography, age, race/ethnicity, education, and income (Agurs-Collins et al. 2015; Apathy et al. 2018; Finney Rutten et al. 2012; Salloum et al. 2018). Now that All of Us has launched as of May 2018, the specific program as well as precision medicine in general promise to become more prominent in the general conversation given the expected exposure to national advertisement campaigns and mostly positive narratives offered by the media (Marcon et al. 2018).

The promise of return of research results

This backdrop of pop cultural curiosity of genetics, mixed with a growing but variable awareness for at least DTC genetic testing, is making return of genetic and genomic results a popular incentive for participation in precision medicine research. Local and national surveys demonstrate that the majority of potential participants would like to receive genetic results (Cooke Bailey et al. 2018), including genetic ancestry data (Kaufman et al. 2016), about themselves. In a study of US veterans, the majority of those surveyed indicated that they would be less likely to participate in a Veterans Affairs biobank if individual research results were not returned (Kaufman et al. 2009). Regarding what results should be returned, most surveyed want “all” results when presented with multiple options (Cooke Bailey et al. 2018; Kaufman et al. 2009), including whole genome sequencing data (Middleton et al. 2015). However, within the genomics research community, there is ongoing dialogue as to what should be returned, by whom and to whom (Jarvik et al. 2014).

The new All of Us program is promising unprecedented access to individual-level research data by a study participant. According to the recent funding announcements for the All of Us Research Program Genome Centers (Health 2018a), awardee laboratories are expected to be Clinical Laboratory Improvement Amendments (CLIA)-certified and are expected to have the capacity for high-throughput whole genome genotyping and sequencing in this regulatory environment. It is likely that the All of Us program will develop custom genotyping arrays to target genetic variants on the American College of Medical Genetics (ACMG) secondary findings list (Kalia et al. 2016) as well as select pharmacogenomic variants. For whole genome sequencing, variant calling and minimal first-pass functional and clinical annotation efforts leveraging existing data in repositories such as ClinVar (Landrum et al. 2014; Rehm et al. 2015) are expected with a focus on genes on the ACMG secondary findings list and pharmacogenes deemed important by Clinical Pharmacogenetics Implementation Consortium (CPIC) standards (Relling et al. 2017). All pathogenic or potentially pathogenic variants identified from the ACMG secondary findings gene list will trigger clinical validation by one of the awardee laboratories, and results will be documented in a formal clinical report.

The gap

The return of these clinically important research results to the participant is not necessarily unexpected or novel given that other medical centers with research biobanks have or had programs to return medically important research results to participants through a medical provider (Pulley et al. 2012; Schwartz et al. 2018). The novelty of the proposed All of Us return of results program is the scale, with at least one million participants being genotyped or sequenced. The other novelty, which is the focus of this brief review, is the scope. That is, in addition to medically important or actionable results, the All of Us program promises to return whole genome data to participants either un-interpreted or variant-called, a task tackled recently by only a handful of research groups (Thorogood et al. 2018). Here we review the resources required to receive, access, and understand whole genome sequencing data from the perspective of the non-expert with brief commentaries on the potential privacy consequences and other issues that could further widen the gap in research disparities in this era of big biomedical data and precision medicine.

How might the data be returned?

Beyond the funding announcement language for the Genome Centers (Health 2018a; Karow 2018), the All of Us program has not yet released a final protocol detailing how whole genome genotype and sequencing data will be returned to participants. Several data delivery mechanisms can be envisioned modeled from individual-level data access or sharing within research consortia (Baker 2010) and DTC genetic testing companies, including 1) granting participants access to a secure website portal and 2) returning data to participants via an encrypted storage device. For participants interested in accessing cohort data to conduct research as non-scientists, data will likely be available via a cloud-based infrastructure and computing environment as a component of the NIH Data Commons (Health 2018b). While the receipt of personal data and access to cohort data is technically feasible, the assumed requirements will not be easily met by all participants, thereby creating disparities in who can access genomic information and who cannot, as well as who will be able to engage in non-scientist research initiatives.

Should All of Us return whole genome results via a secure website portal, an in-home internet connection will be required. The majority of US residents do have a broadband internet subscription (~65-77%) and a desktop or laptop (~75%) (Center 2018; Ryan and Lewis 2017). However, there are marked differences in internet access and availability of a desktop or laptop in households by geography, age, race/ethnicity, education, and income. For example, it is well documented that a digital divide persists between rural and urban America, and as of 2015, more than 50% of rural US residents lack access to high-speed internet service. Internet access is the worst among Americans living on Tribal lands and US territories (2015), the latter's situation further exacerbated by Hurricanes Irma and Maria's landfall in 2017 (Thieme 2018). Based on demographics, in 2015 households that were African American alone (65%), led by an adult 65 years or older (62%), with limited English speaking skills (55%), had a household income less than $25,000 (51%), or where the educational attainment of the householder was at most a high school graduate (60%) had reduced rates of broadband internet subscriptions compared with other groups (Ryan and Lewis 2017). Delivery of All of Us program data via an encrypted storage device will not address these broadband access issues as data access will *still* require a desktop or laptop, and their ownership is associated with similar patterns of accessibility as described for internet access.

What might the data look like?

The popular DTC genetic testing companies such as 23andMe currently allow consumers to download their array-based whole genome genotype call data as compressed (zipped) text files, for which there are several readily available desktop tools to decompress such files. These files are 5 MB to 30 MB in size and can be opened in a text editor such as WordPad or Microsoft Excel, the latter of which has a maximum of 1,048,576 data rows. All of Us has not yet finalized how whole genome genotype or sequence data will be returned, though it is unlikely that general-purpose text compression methods such as gzip (Gailly and Adler 1992) and bzip2 (Seward 1996) will be used for whole genome sequence data. As described in the All of Us Genome Centers funding announcement OT-PM-18-002 (Health 2018a; Karow 2018), it is very likely the data will be in the form of compressed/binary SAM (known as BAM), CRAM, VCF or BCF files as opposed to the unmapped short reads and quality scores stored in FASTQ files (Cock et al. 2010). All of these files are generated by specialized algorithms that take advantage of repetitive subsequences inherent in genomic sequences to achieve a higher compression ratio compared to general-purpose text compression methods (Zhu et al. 2015).

Even more challenging than the file format is the fact that these algorithms and related tools work almost exclusively within command-line (i.e., Unix or Linux operating systems) environments which are largely unknown to the general population. Most of the freely available resources for “viewing” or working with whole genome data require, at minimum, baseline knowledge of working in a commandline environment. For All of Us participants unfamiliar with working in such an environment, online educational resources do exist that are hypothetically accessible to those with time and the means to access them. As an example, a quick internet search reveals “Command Line Tools for Genomic Data Science” Coursera class offered by Johns Hopkins University - course 5 of 8 in the Coursera Genomic Data Science Specialization. To sign up, users must have an e-mail address (and thus, an internet connection). Users can access Coursera via a 7-day free trial, and can continue to access course materials for this and all courses in the specialization for $39 per month, or users can choose the option of auditing the course. The course information page (University 2016) estimates that the course will take approximately 26 hours to complete, with a suggested commitment of 7 hours per week. The course is available only in English with English subtitles.

Once familiar with command-line environment, we can consider the possible file formats All of Us may use to return whole genome sequence data. Sequence data are commonly presented in the Sequence Alignment/Map (SAM) format, which was designed to harmonize data generated by distinct sequencing technologies. The SAM format is generated by the Burrows-Wheeler Alignment tool and is a common alignment structure applicable to all sequence types and aligners, supporting single- and paired-end reads as well as combining distinct read types/sources (Li et al. 2009). SAM format functions as a common interface between the alignment/mapping process and such downstream analyses as variant detection, genotyping, and assembly. SAM formatted files are composed of tab-delimited header and alignment sections, including 11 mandatory fields containing information pertinent to the sequence data contained in the file. SAM formatted files can average 500 gigabytes per file (which is equivalent to the total storage available for the average desktop computer) depending on coverage and number of reads (Li et al. 2009). To improve performance speed, the Binary Alignment/Map (BAM) format was designed, and it contains SAM format information with additional compression developed to achieve faster random access. A BAM file of a single whole genome sequence data file requires approximately 100 gigabytes. The CRAM format is a compressed alignment file of a BAM file designed by the European Bioinformatics Institute as a highly space-efficient format that uses reference-based compression (Zalunin 2012), wherein information stored includes only base calls that differ from the designated reference sequence, resulting in a 40-50% space savings in comparison with a BAM file (Hsi-Yang Fritz et al. 2011). The CRAM format requires reference sequences maintained by the Global Alliance for Genomics and Health (GA4GH) Large Scale Genomics workstream and made available in a CRAM reference registry. Users must note that the specific reference sequence used to generate the CRAM file must be used for compression and decompression.

The other likely data format would be the Variant Call Format (VCF) which was developed for the 1000 Genomes Project (Clarke et al. 2012) as a generic text file format that stores DNA variant information including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants, combined with related annotation information with respect to a reference genome sequence (Danecek et al. 2011). The annotation includes genomic position, reference and alternate alleles, quality of read, filter status (dependent upon whether a filter was applied), and additional information that varies depending upon the tool used to generate the VCF file (Danecek et al. 2011; McKenna et al. 2010). The VCF file is expected to be much smaller than the BAM file as it retains information on variants alone (at most a few hundred megabytes), with respect to a reference file (which may be a gigabyte in size). There is also BCF, which is binary version of a VCF file and requires even less storage space, analogous to the BAM version of a SAM file. VCF and related files are considered “interpreted” data files and may require an investigational device exemption with the Food and Drug Administration (FDA) before they can be returned to participants (Karow 2018).

Working with likely file formats

In general, software used to generate and manipulate the mapped (SAM, BAM, and CRAM) and annotated (VCF/BCF) sequencing data files is open source and free to use. For the average non-scientist participant, identifying, locating, and implementing the various tools can be daunting. For the format files, the SAMtools library and software package provides universal tools for working with SAM, BAM, and CRAM formats (Li 2011; Li et al. 2009). SAMtools is available in various programming languages, including C, C++, Java, Perl, Python, Ruby, Common Lisp, and Haskell, and tutorials for varying levels of expertise are available on various websites (Table 1). In addition to SAMtools, there are several other command-line packages which support these file formats, including Scramble (Bonfield 2014), CRAMtools (Hsi-Yang Fritz et al. 2011; Zalunin 2012), and Picard (Table 1).

If participant data are returned as unannotated but mapped file formats (SAM, BAM, and CRAM), participants can use a variety of available software to annotate their data. In bioinformatics, one of the most popular toolkits available for calling and annotating germline variation is the Genome Analysis Toolkit (GATK), a set of tools for working with high-throughput sequencing data, including variant calling, processing, and quality control, maintained by the Broad Institute (Van der Auwera et al. 2013). GATK is available via download and is supported in Unix, Linux, and MacOSX environments (not Microsoft Windows), or to run on FireCloud, a web-based, open platform for secure and scalable analysis in the cloud. Implementing GATK takes considerable bioinformatics expertise and time and requires a computing infrastructure not available to most non-scientists (Hansen 2016). There are alternative pipelines such as SpeedSeq (Chiang et al. 2015), Freebayes

(Garrison and Marth 2012) and “BAM to most probable genotype” (bam2mpg) (Hansen 2016; Teer et al. 2010). While these alternative pipelines may be more manageable compared with GATK, they still require a reference sequence dataset (such as from the 1000 Genomes Project) and access to a relatively expensive computing environment. For participants without these resources, an attractive alternative could be Galaxy (Afgan et al. 2018), a freely available, internet-accessible platform with a graphical user interface (GUI) designed to enable a variety of biomedical research analysis needs, including the manipulation of next generation sequencing files as described here.

If participant data is returned as annotated (“interpreted”) files such as VCF files, they can be manipulated using a variety of tools, most commonly by VCFtools (Danecek et al. 2011) which is a Perl or C++ program. There are also a handful of GUI-based applications, such as VCF-Miner (Hart et al. 2016), VCF.Filter (Muller et al. 2017), VarAFT (Desvignes et al. 2018), and the aforementioned Galaxy platform (Afgan et al. 2018). These GUI applications aim to provide bioinformaticians, medical geneticists, and researchers who have limited computing resources or are less comfortable with working at command-line alternative tools to readily interrogate VCF files, assuming ample user genomic literacy. As for BCF files, they can be manipulated using depracated versions of SAMtools or a contemporary version of BCFtools (Table 1).

For those participants with the resources to pay for commercially available tools, various are available, including Geneious (Kearse et al. 2012), BaseSpace (Illumina 2018), DNAnexus (Anderson 2017; DNAnexus 2019), among others (Milicchio et al. 2016; Smith 2015). Unlike the open source tools listed in Table 1, most of the commercially available tools are all-in-one and offer slick, easy-to-use GUIs that require no computer programming skills or command-line know-how (Milicchio et al. 2016; Smith 2015). Many of these commercial applications also take advantage of cloud computing resources (Navale and Bourne 2018), a major advantage for participants who do not want or cannot afford to invest in the hardware necessary for bioinformatics. A major disadvantage of these commercially available tools is the pricing, where licenses can range from several hundred to thousands of US dollars. Also, it is not clear how pricing will differ for participants not affiliated with the usual customer base typically associated with these software (e.g., academics, government, and industry).

Third party interpretation services

Considering the limited options available to the general populace for interrogating BAM, CRAM, VCF, or BCF files, it is likely that commercial website-based companies, already coined “third party interpretation (TPI)” sites (Badalato et al. 2017), will arise to offer services to store, manipulate, and translate the whole genome genotype and sequencing data returned to research participants. Flere, “interpretation” refers to offering the understanding of genetic variant data in the context of the customer’s health status, physical or physiologic traits, and/or disease risk. Currently, there are several TPI services that readily accept the downloadable genotype call data from the popular DTC genetic testing companies. These yet-to-be-regulated TPI services offer a smorgasbord of health-related services including fitness insights, nutrigenomic guidance, longevity reports, and a wide range of disease/ health-trait associations, providing an alternate venue for interested consumers as regulatory policies have limited the range of disease/health-trait associations that can be reported by DTC genetic testing companies (Borry et al. 2012; Branch 2013). The ethical concerns leading to the regulation of DTC genetic testing companies also apply to TPI services, including deficient informed consent, unclear privacy and data storage practices, uncertain clinical validity and utility of reported associations, and a clear absence of medical supervision (Badalato et al. 2017).

To gauge the current landscape of TPI services, we conducted an informal internet search of TPI services (using a collection of generic search terms such as “what to do with raw 23andMe data” and “interpret raw 23andMe data”) and have summarized 15 of these services which ranged from free of cost to a few hundred dollars (Supplementary Table 1). Two services accept whole genome sequence data (i.e., VCF files): Promethease and Enlis Genomics (Supplementary Table 1). All other identified TPI services generally require compressed/decompressed text files of genotypic data downloaded from DTC genetic testing services such as 23andMe, AncestryDNA, MyHeritage, FamilyTreeDNA, and several others. Many but not all services have clearly identifiable privacy policies and some degree of information on data storage/ownership. Most services report health-related information, but only a few outlined the resources used to generate the reports returned to consumers. For example, Promethease reports genetic associations available in SNPedia (Cariaso and Lennon 2012) and GENOtation reports genetic associations available in the National Human Genome Research Institute - European Bioinformatics Institute (NHGRI-EBI) Genome Wide Association Studies (GWAS) catalog (MacArthur et al. 2017); however many TPI services do not state the studies or references used to develop health-related reports (Supplementary Table 1).

The variability of the information relayed and data included in genetic test reports is a general problem as no standards exist even for clinically-ordered genetic testing. In the DTC space, 23andMe is currently the only service to have received marketing authorization by the FDA that allows the service to offer consumers genetic risk health reports for a limited number of genetic variants and conditions, including breast/ovarian cancer and *BRCA1/BRCA2,* Bloom Syndrome, and now pharmacogenomic-related variants, among others (Bates 2018). To fulfill part of the requirements established by FDA, 23andMe is required to use CLIA-certified labs for genotyping and to create genetic risk health reports using an easily understood format. No guidelines are given for how to report individual-level risk, population-level measures of association, or other classifications such as genotypes or variant annotations. Like DTC companies, TPI services, and clinical testing entities, large studies returning genomic results such as All of Us will have the opportunity to create their own genetic risk reports, potentially innovating the way genetic risk is communicated to consumers, patients, and participants.

With regard to the variability in how genetic information is relayed, we make special note that there are also no universal standards for reporting race-specific genetic results. None of the TPI services evaluated here had an option for race-specific reported genetic risk. The importance of race/ethnicity or genetic ancestry is well recognized when interpreting genetic data. As an example, CPIC provides guidelines that consider race or genetic ancestry if the data support clinically-relevant differences (e.g., race- or ancestry-specific genetic variants to include in warfarin dosing models) (Johnson et al. 2017). The lack of race-specific results offered by genetic testing services is in large part due to the limited genetic association studies and reference data in populations of non-European ancestry (Sirugo et al. 2019). As concerted efforts are underway to fill these diversity data gaps (notably, a major goal of All of Us), clinically-oriented groups such as the Clinical Genome Resource (ClinGen) have already mobilized an Ancestry and Diversity Working Group charged with investigating how best to scientifically and ethically use race, ancestry and genomic data (Popejoy et al. 2018).

A major challenge for consumers engaging in TPI services, including All of Us participants receiving raw genomic data, is the understanding of the various limitations and assumptions that underlie both the genetic data itself as well as the interpretation of those data. Reports are already beginning to surface describing DTC consumers using TPI services that then identify pathogenic variants outside the official FDA-approved DTC genetic risk health reports. When followed-up in a clinical setting, one report described that 40% of the DTC-generated genetic variants (genotype calls) were incorrect (false positives) (Tandy-Connor et al. 2018). Equally troubling are the reports that verified genotypes were misclassified as pathogenic by TPI services when clinically-contracted labs classified them as benign (Moscarello et al. 2019; Tandy-Connor et al. 2018). Per the funding announcement (Health 2018a; Karow 2018), All of Us Genome Centers include a CLIA-certified laboratory supported specifically to verify genetic variants considered clinically actionable and worthy of a clinical report to be returned to the participant. While this plan addresses potential false positive genetic variants for the returned clinical reports, it in no way addresses nor can it address the allure of extra annotation and the inevitable mutation misclassifications that will follow.

Evolving privacy concerns

Individuals who have downloaded their raw DTC genetic data are also joining crowdsourcing platforms such as openSNP (opensnp.org) (Greshake et al. 2014) and GEDmatch (www.gedmatch.com). These platforms allow users to openly share their genetic information downloaded from DTC genetic testing sites. OpenSNP collects phenotypic information, while GEDmatch is focused on reconstructing genealogies, and users who have shared their genetic information can also access the genetic information from other users. The possibility of re-identification via these services grabbed international headlines recently with the arrest of the Golden State Killer by criminal investigators (Guerrini et al. 2018). In general, public and private databases are biased towards genomes of European-descent (Ram et al. 2018), and recent work suggests that in the near future, re-identification of individuals of European ancestry in the US who have publicly released their genetic information is likely (Erlich et al. 2018). It is interesting to note that an unintended consequence of today’s cohort ascertainment efforts focused on recruitment of underrepresented groups may be the increased population of these private and open-source databases with genomes of diverse ancestries, widening privacy concerns to an even greater number participants and their un-genotyped relatives (Hazel et al. 2018).

Potential to widen research disparities

Return of research results to participants in of itself will not likely widen existing health disparities. Rather, return of research results will likely highlight existing health disparities in who can follow-up on the genetic results with additional medical screening, interventions, and treatments. Return of research results will also emphasize existing and noted deficiencies in genomic research and knowledge ranging from data on penetrance and risk for individual variants to lack of genotype-phenotype data for understudied populations.

The return of these genomic data may also impact genomic research in unexpected ways. That is, a major motivation to participate in genetic studies is the potential to receive results, and this motivation in turn is used as an incentive to participate in genetic studies. Previous sequencing studies have noted that early adopters of this technology on average were European-descent, highly educated, and relatively affluent (Facio et al. 2012; Yu et al. 2014). It may be that the offering of whole genome sequencing data to participants in All of Us will attract more participants with similar early adopter profiles (Lewis et al. 2015), an *irony* given that a major goal of All of Us is the recruitment of a diverse study population defined broadly by age, socioeconomic status, health status, race/ethnicity, and geography.

The offering of whole genome genotype and sequencing data as a returned result to all is intended to promote participant autonomy as well as trust between the participant and the scientific community (Angrist 2011). Increasingly, return of results in a program such as All of Us is also meant to promote science conducted by non-scientists where participants champion specific research questions or areas of scientific inquiry (Aungst et al. 2017; Beck et al. 2018; Woolley et al. 2016). For groups historically underrepresented in and distrustful of biomedical research, return of whole genome sequence data and other pro-offerings of data and research transparency may be insufficient to foster the necessary trust and relationship. Early reactions from Native American groups (Hansen and Keeler 2018) to the Tribal Collaboration Working Group Report (Report 2018) recently submitted to the All of Us Research Program Advisory Panel underscore this challenge, which will impact both participation and the research conducted that could benefit specific populations. More generally, differences in trust, genomic and health literacy, resources, skill sets, and opportunities to engage in auspicious science as a non-scientist may promote non-response bias in this cohort, which, while not designed to be population-based, is designed with the intent to represent as many US groups as possible.

Mind the gap

Bearing in mind the diversity of potential participants, and in the spirit of building trust, cohorts such as All of Us are beholden to ensuring equity in access to the returned research data. The least involved process for returning data might be through a secure website portal; however, considerations must first be made for those with little to no digital literacy. Sixteen percent of American adults between the ages of 16 and 65 years are digitally illiterate and 5% have no computer experience (Mamedova and Pawlowski 2018). Highest digital illiteracy rates exist among special populations that are of particular interest to All of Us, including 35% of Hispanics, 22% of Blacks, and 28% of individuals 55-65 years

(Mamedova and Pawlowski 2018). For those who are digitally literate, considerations for access to private computing resources and/or internet access are needed, particularly instances that might result in the downloaded genetic data being accidentally left on a public computer. One possible solution to the internet divide would be mailing encrypted hard-drives, but that would only resolve access for those with a personal computer. Thus, access to these large genetic files is in it of itself nuanced and will require a multifaceted approach.

Once access to the data has been granted, it is unlikely that the average *digitally literate* scientist or non-scientist would be capable of navigating these large and complex data files without a suite of detailed tutorials for existing tools and/or the expansion of the unregulated TPI market. Here lies an opportunity for the development of novel and effective solutions for returning data to genomic participants, such as a customized web-based tool that would allow individuals to securely store, navigate, and manipulate their data, and possibly interact with other participants as a means to facilitate science by non-scientists. A template for such a web-based tool for non-scientists is My46 (Tabor et al. 2017), a tool developed by the University of Washington to help individuals manage their genetic testing results. My46 provides accessible information on genetics and genetic testing, generates several health summary reports, and offers access to genetic counselors to answer questions on the returned health reports (Tabor et al. 2017). While an attractive template, web-based services such as My46 will not provide data access for the digitally illiterate. Thus, the need for a similarly thoughtful, interactive, and user-friendly tool that incorporates the spectrum of diversity (race, ethnicity, age, access to computational and online resources, education, geography, as well as health, genetic, and digital literacy) warrants intentional stakeholder investment by All of Us and any other cohort in which the offer of return of research results is intended for all participants as opposed to the privileged few who already have the ability or means to access these data.

Conclusions

Surveys of potential and current research participants make it clear that return of research results is both desired and expected. Large-scale studies such as All of Us are generating and offering genome-wide data beyond the scale and scope of previously debated return of results discussions. The responsible and equitable return of research results will require substantial investment in the development of multiple strategies to disseminate individual-level data to participants so that all who want them can access and potentially act on them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was funded by institutional funds from the CWRU Cleveland Institute for Computational Biology as well as the Clinical and Translational Science Collaborative of Cleveland UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCATS or NIH. JNCB is supported by the Clinical and Translational Science Collaborative of Cleveland (KL2TR002547).

References

(2015) 2015 Broadband Progress Report. Federal Communications Commission https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2015-broadband-progress-report. Accessed 01/02/2019 2019

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46: W537–W544. doi: 10.1093/nar/gky379 [PubMed: 29790989]

Agurs-Collins T, Ferrer R, Ottenbacher A, Waters EA, O'Connell ME, Hamilton JG (2015) Public Awareness of Direct-to-Consumer Genetic Tests: Findings from the 2013 U.S. Health Information National Trends Survey. Journal of Cancer Education 30: 799–807. doi: 10.1007/sl3187-014-0784-x [PubMed: 25600375]

Anderson C (2017) AZ Partners with DNAnexus for 2 Million Patient Sequencing Project. Clinical OMICs 4: 32–32. doi: 10.1089/clinomi.04.04.23

Angrist M (2011) You never call, you never write: why return of 'omic' results to research participants is both a good idea and a moral imperative. Personalized Medicine 8: 651–657. doi: 10.2217/pme.ll.62 [PubMed: 22199990]

Apathy NC, Menser T, Keeran LM, Ford EW, Harle CA, Fluerta TR (2018) Trends and Gaps in Awareness of Direct-to-Consumer Genetic Tests From 2007 to 2014. American Journal of Preventive Medicine 54: 806–813. doi: 10.1016/j.amepre.2018.02.013 [PubMed: 29656919]

Aungst H, Fishman JR, McGowan ML (2017) Participatory Genomic Research: Ethical Issues from the Bottom Up to the Top Down. Annual Review of Genomics and Human Genetics 18: 357–367. doi: 10.1146/annurev-genom-091416-035230

Badalato L, Kalokairinou L, Borry P (2017) Third party interpretation of raw genetic data: an ethical exploration. European Journal Of Human Genetics 25: 1189. doi: 10.1038/ejhg.2017.126 [PubMed: 28832567]

Baker M (2010) Next-generation sequencing: adjusting to data overload. Nature Methods 7: 495. doi: 10.1038/nmeth0710-495

Bates M (2018) Direct-To-Consumer Genetic Testing: Is the Public Ready for Simple, At-Home DNA Tests to Detect Disease Risk? IEEE Pulse 9: 11–14. doi: 10.1109/MPUL.2018.2869315 [PubMed: 30452341]

Beck S, Berner AM, Bignell G, Bond M, Callanan MJ, Chervova O, Conde L, Corpas M, Ecker S, Elliott HR, Fioramonti SA, Flanagan AM, Gaentzsch R, Graham D, Gribbin D, Guerra-Assunção JA, Hamoudi R, Harding V, Harrison PL, Herrero J, Hofmann J, Jones E, Khan S, Kaye J, Kerr P, Libertini E, Marks L, McCormack L, Moghul I, Pontikos N, Rajanayagam S, Rana K, Semega-Janneh M, Smith CP, Strom L, Umur S, Webster AP, Williams EH, Wint K, Wood JN, Consortium P-U (2018) Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. BMC Medical Genomics 11: 108. doi: 10.1186/s12920-018-0423-1 [PubMed: 30482208]

Bonfield JK (2014) The Scrmble conversion tool. Bioinformatics 30: 2818–2819. doi: 10.1093/bioinformatics/btu390 [PubMed: 24930138]

Borry P, van Hellemondt RE, Sprumont D, Jales CF, Rial-Sebbag E, Spranger TM, Curren L, Kaye J, Nys H, Howard H (2012) Legislation on direct-to-consumer genetic testing in seven European countries. Eur J Hum Genet 20: 715–21. doi: 10.1038/ejhg.2011.278 [PubMed: 22274578]

Branch M (2013) The FDA and me. Nature 504: 7–8.

Cariaso M, Lennon G (2012) SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res 40: D1308–12. doi: 10.1093/nar/gkr798 [PubMed: 22140107]

Center PR (2018) Internet/Broadband Fact Sheet. http://www.pewinternet.org/fact-sheet/internet-broadband/. Accessed 01/02/2019 2019

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM (2015) SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature Methods 12: 966. doi: 10.1038/nmeth.3505 [PubMed: 26258291]

Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, Sherry S, Flicek P, The Genomes Project C (2012) The 1000 Genomes Project: data management and community access. Nature Methods 9: 459. doi: 10.1038/nmeth.l974 [PubMed: 22543379]

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQfile format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767–1771. doi: 10.1093/nar/gkp1137 [PubMed: 20015970]

Cooke Bailey JN, Crawford DC, Goldenberg A, Slaven A, Pencak J, Schachere M, Bush WS, Sedor JR, O'Toole JF (2018) Willingness to participate in a national precision medicine cohort: Attitudes of chronic kidney disease patients at a Cleveland public hospital. Journal of Personalized Medicine 8: 21. doi: 10.3390/jpm8030021

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Group GPA (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158. doi: 10.1093/bioinformatics/btr330 [PubMed: 21653522]

Desai S, Jena AB (2016) Do celebrity endorsements matter? Observational study of BRCA gene testing and mastectomy rates after Angelina Jolie's New York Times editorial. BMJ 355: i6357. doi: 10.1136/bmj.i6357 [PubMed: 27974323]

Desvignes J-P, Bartoli M, Delague V, Krahn M, Miltgen M, Béroud C, Salgado D (2018) VarAFT: a variant annotation and filtration system for human next generation sequencing data. Nucleic Acids Research 46: W545–W553. doi: 10.1093/nar/gky471 [PubMed: 29860484]

DNAnexus (2019) DNAnexus. https://www.dnanexus.com/. Accessed 05/20/2019 2019

Dougherty MJ, Pleasants C, Solow L, Wong A, Zhang H (2011) A Comprehensive Analysis of High School Genetics Standards: Are States Keeping Pace with Modern Genetics? CBE-Life Sciences Education 10: 318–327. doi: 10.1187/cbe.l0-09-0122 [PubMed: 21885828]

Erlich Y, Shor T, Pe'er I, Carmi S (2018) Identity inference of genomic data using long-range familial searches. Science 362: 690. doi: 10.1126/science.aau4832 [PubMed: 30309907]

Facio FM, Eidem H, Fisher T, Brooks S, Linn A, Kaphingst KA, Biesecker LG, Biesecker BB (2012) Intentions to receive individual results from whole-genome sequencing among participants in the ClinSeq study. European Journal Of Human Genetics 21: 261. doi: 10.1038/ejhg.2012.179 [PubMed: 22892536]

Finney Rutten LJ, Gollust SE, Naveed S, Moser RP (2012) Increasing Public Awareness of Direct-to-Consumer Genetic Tests: Health Care Access, Internet Use, and Population Density Correlates. Journal of Cancer Epidemiology 2012: 309109. doi: 10.1155/2012/309109 [PubMed: 22899921]

Gailly J-I, Adler M (1992) GNU Gzip. GNU. https://www.gnu.org/software/gzip/. Accessed 01/08/2019 2019

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. BioRxiv.

Greshake B, Bayer PE, Rausch H, Reda J (2014) openSNP--a crowdsourced web resource for personal genomics. PLoS One 9: e89204. doi: 10.1371/journal.pone.0089204 [PubMed: 24647222]

Guerrini G, Robinson JO, Petersen D, McGuire AL (2018) Should police have access to genetic genealogy databases? Capturing the Golden State Killer and other criminals using a controversial new forensic technique. PLOS Biology 16: e2006906. doi: 10.1371/journal.pbio.2006906 [PubMed: 30278047]

Haga SB, Barry WT, Mills R, Ginsburg GS, Svetkey L, Sullivan J, Willard HF (2013) Public knowledge of and attitudes toward genetics and genetic testing. Genet Test Mol Biomarkers 17: 327–335. doi: 10.1089/gtmb.2012.0350 [PubMed: 23406207]

Hansen NF (2016) Variant Calling From Next Generation Sequence Data In: Mathé E, Davis S (eds) Statistical Genomics: Methods and Protocols. Springer New York, New York, NY, pp 209–224

Hansen T, Keeler J (2018) The NIH is bypassing tribal sovereignty to harvest genetic data from Native Americans. Motherboard. https://motherboard.vice.com/en_us/article/8xp33a/the-nih-is-bypassing-tribal-sovereignty-to-harvest-genetic-data-from-native-americans. Accessed 01/10/2019 2019

Hart SN, Duffy P, Quest DJ, Hossain A, Meiners MA, Kocher JP (2016) VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. Brief Bioinform 17: 346–351. doi: 10.1093/bib/bbv051 [PubMed: 26210358]

Hazel JW, Clayton EW, Malin BA, Slobogin C (2018) Is it time for a universal genetic forensic database? Science 362: 898–900. doi: 10.1126/science.aav5475 [PubMed: 30467160]

Health Nlo (2018a) All of Us Research Program Genome Centers (OT2). OT-PM-18–002

Health Nlo (2018b) NIH Strategic Plan for Data Science. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf. Accessed 04/04/2019 2019

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Research 21: 734–740. doi: 10.1101/gr.114819.110 [PubMed: 21245279]

Illumina (2018) BaseSpace Sequence Hub In: Illumina (ed), pp 1–6

Jarvik GP, Amendola LM, Berg JS, Brothers K, Clayton EW, Chung W, Evans BJ, Evans JP, Fullerton SM, Gallego CJ, Garrison NA, Gray SW, Holm IA, Kullo IJ, Lehmann LS, McCarty C, Prows CA, Rehm HL, Sharp RR, Salama J, Sanderson S, Van-Driest SL, Williams MS, Wolf SM, Wolf WA, Burke W (2014) Return of Genomic Results to Research Participants: The Floor, the Ceiling, and the Choices In Between. The American Journal of Human Genetics 94: 818–826. [PubMed: 24814192]

Johnson JA, Caudle KE, Gong L, Whirl-Carrillo M, Stein CM, Scott SA, Lee MT, Gage BF, Kimmel SE, Perera MA, Anderson JL, Pirmohamed M, Klein TE, Limdi NA, Cavallari LH, Wadelius M (2017) Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. Clinical Pharmacology & Therapeutics 102: 397–404. doi: 10.1002/cpt.668 [PubMed: 28198005]

Jolie A (2013) My Medical Choice. The New York Times, pp A25

Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT (2016) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genetics In Medicine 19: 249. doi: 10.1038/gim.2016.190 [PubMed: 27854360]

Karow J (2018) All of Us Program plans to return disease variants, PGx results, primary genomic data. GenomeWeb. https://allofus.nih.gov/sites/default/files/all_of_us_program_plans_to_return_disease_variants_pgx_results_primary_genomic_data.pdf. Accessed 01/02/2019 2019

Kaufman D, Murphy J, Erby L, Hudson K, Scott J (2009) Veterans' attitudes regarding a database for genomic research. Genetics In Medicine 11: 329. doi: 10.1097/GIM.0b013e31819994f8 [PubMed: 19346960]

Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL (2016) A Survey of U.S Adults' Opinions about Conduct of a Nationwide Precision Medicine Initiative® Cohort Study of Genes and Environment. PLoS ONE 11: e0160461. doi: 10.1371/journal.pone.0160461 [PubMed: 27532667]

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649. doi: 10.1093/bioinformatics/bts199 [PubMed: 22543367]

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucl Acids Res 42: D980–D985. doi: 10.1093/nar/gkt1113 [PubMed: 24234437]

Lewis KL, Han PKJ, Hooker GW, Klein WMP, Biesecker LG, Biesecker BB (2015) Characterizing Participants in the ClinSeq Genome Sequencing Cohort as Early Adopters of a New Health Technology. PLOS ONE 10: e0132690. doi: 10.1371/journal.pone.0132690 [PubMed: 26186621]

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993. doi: 10.1093/bioinformatics/btr509 [PubMed: 21903627]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi: 10.1093/bioinformatics/btp352 [PubMed: 19505943]

Liede A, Cai M, Crouter TF, Niepel D, Callaghan F, Evans DG (2018) Risk-reducing mastectomy rates in the US: a closer examination of the Angelina Jolie effect. Breast Cancer Research and Treatment 171: 435–442. doi: 10.1007/s10549-018-4824-9 [PubMed: 29808287]

Lontok KS, Zhang H, Dougherty MJ (2015) Assessing the Genetics Content in the Next Generation Science Standards. PLOS ONE 10: e0132742. doi: 10.1371/journal.pone.0132742 [PubMed: 26222583]

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45: D896–D901. doi: 10.1093/nar/gkw1133 [PubMed: 27899670]

Mamedova S, Pawlowski E (2018) A description of U.S. adults who are not digitally literate In: Hudson L (ed) Stats in Brief. US Department of Education, National Center for Education Statistics

Marcon AR, Bieber M, Caulfield T (2018) Representing a “revolution”: how the popular press has portrayed personalized medicine. Genetics in Medicine 20: 950–956. doi: 10.1038/gim.2017.217 [PubMed: 29300377]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303. doi: 10.1101/gr.107524.110 [PubMed: 20644199]

Middleton A, Wright CF, Morley KI, Bragin E, Firth HV, Hurles ME, Parker M (2015) Potential research participants support the return of raw sequence data. Journal of Medical Genetics 52: 571–574. doi: 10.1136/jmedgenet-2015-103119 [PubMed: 25995218]

Milicchio F, Rose R, Bian J, Min J, Prosperi M (2016) Visual programming for next-generation sequencing data analytics. BioData Mining 9: 16. doi: 10.1186/sl3040-016-0095-3 [PubMed: 27127540]

Moscarello T, Murray B, Reuter CM, Demo E (2019) Direct-to-consumer raw genetic data and third-party interpretation services: more burden than bargain? Genetics in Medicine 21: 539–541. doi: 10.1038/S41436-018-0097-2 [PubMed: 29997392]

Müller A, Dalzotto A (2018) GATTACA and genetic determinism. Nurse Education Today 70: 94–95. doi: 10.1016/j.nedt.2018.08.004 [PubMed: 30172229]

Muller H, Jimenez-Heredia R, Krolo A, Hirschmugl T, Dmytrus J, Boztug K, Bock C (2017) VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. Nucleic Acids Res 45: W567–W572. doi: 10.1093/nar/gkx425 [PubMed: 28520890]

Navale V, Bourne PE (2018) Cloud computing applications for biomedical science: A perspective. PLOS Computational Biology 14: el006144. doi: 10.1371/journal.pcbi.1006144

Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H, Wright MW, Zou J, Gignoux CR, Bonham VL, Plon SE, Bustamante CD, Clinical Genome Resource A, Diversity Working G (2018) The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum Mutat 39: 1713–1720. doi: 10.1002/humu.23644 [PubMed: 30311373]

Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, Ramirez AH, Delaney JT, Bowton E, Brothers K, Johnson K, Crawford DC, Schildcrout J, Masys DR, Dilks HH, Wilke RA, Clayton EW, Shultz E, Laposata M, McPherson J, Jirjis JN, Roden DM (2012) Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. Clin Pharmacol Ther 92: 87–95. doi: 10.1038/clpt.2011.371 [PubMed: 22588608]

Ram N, Guerrini CJ, McGuire AL (2018) Genealogy databases and the future of criminal investigation. Science 360: 1078–1079. doi: 10.1126/science.aau1083 [PubMed: 29880677]

Ray T (2018) Public awareness of personlized medicine not growing in step with industry, survey shows. GenomeWeb, www.genomeweb.com

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS (2015) ClinGen — The Clinical Genome Resource. New England Journal of Medicine 372: 2235–2242. doi: 10.1056/NEJMsr1406261 [PubMed: 26014595]

Relling M, Krauss R, Roden D, Klein T, Fowler D, Terada N, Lin L, Riel-Mehan M, Do T, Kubo M, Yee S, Johnson G, Giacomini K (2017) New Pharmacogenomics Research Network: An Open Community Catalyzing Research and Translation in Precision Medicine. Clinical Pharmacology & Therapeutics 102: 897–902. doi: doi:10.1002/cpt.755 [PubMed: 28795399]

Report TTCWG (2018) Considerations for meaningful collaboration with tribal pouplations. All of Us Research Program Advisory Panel https://allofus.nih.gov/sites/default/files/tribal_collab_work_group_rept.pdf. Accessed 01/10/2019 2019

Ryan C, Lewis J,M. (2017) Computer and Internet Use in the United States: 2015. https://www.census.gov/content/dam/Census/library/publications/2017/acs/acs-37.pdf. Accessed 01/02/2019 2019

Salloum RG, George TJ, Silver N, Markham MJ, Hall JM, Guo Y, Bian J, Shenkman EA (2018) Rural-urban and racial-ethnic differences in awareness of direct-to-consumer genetic testing. BMC Public Health 18: 277. doi: 10.1186/s12889-018-5190-6 [PubMed: 29471813]

Schwartz MLB, McCormick CZ, Lazzeri AL, Lindbuchler DAM, Hallquist MLG, Manickam K, Buchanan AH, Rahm AK, Giovanni MA, Frisbie L, Flansburg CN, Davis FD, Sturm AC, Nicastro C, Lebo MS, Mason-Suares H, Mahanta LM, Carey DJ, Williams JL, Williams MS, Ledbetter DH, Faucett WA, Murray MF (2018) A Model for Genome-First Care: Returning Secondary Genomic Findings to Participants and Their Healthcare Providers in a Large Research Cohort. The American Journal of Human Genetics 103: 328–337. doi: 10.1016/j.ajhg.2018.07.009 [PubMed: 30100086]

Schweitzer NJ, Saks MJ (2007) The CSI effect: popular fiction about forensic science affects the public's expectations about real forensic science. Jurimetrics J 47: 357–364.

Seward J (1996) bzip2. GitLab. https://gitlab.com/bzip/bzip2. Accessed 01/08/2019 2019

Sirugo G, Williams SM, Tishkoff SA (2019) The Missing Diversity in Human Genetic Studies. Cell 177: 26–31. doi: 10.1016/j.cell.2019.02.048 [PubMed: 30901543]

Smith DR (2015) Buying in to bioinformatics: an introduction to commercial sequence analysis software. Briefings in Bioinformatics 16: 700–709. doi: 10.1093/bib/bbu030 [PubMed: 25183247]

Tabor HK, Jamal SM, Yu JH, Crouch JM, Shankar AG, Dent KM, Anderson N, Miller DA, Futral BT, Bamshad MJ (2017) My46: a Web-based tool for self-guided management of genomic test results in research and clinical settings. Genet Med 19: 467–475. doi: 10.1038/gim.2016.133 [PubMed: 27632689]

Tandy-Connor S, Guiltinan J, Krempely K, LaDuca H, Reineke P, Gutierrez S, Gray P, Tippin Davis B (2018) False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. Genet Med 20: 1515–1521. doi: 10.1038/gim.2018.38 [PubMed: 29565420]

Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Program NCS, Margulies EH, Green ED, Collins FS, Mullikin JC, Biesecker LG (2010) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Research 20: 1420–1431. doi: 10.1101/gr.106716.110 [PubMed: 20810667]

Thieme N (2018) After Hurrican Maria, Puerto Rico's Internet Problems Go from Bad to Worse. Public Broadcasting Service https://www.pbs.org/wgbh/nova/article/puerto-rico-hurricane-maria-internet/. Accessed 01/02/2019 2019

Thorogood A, Bobe J, Prainsack B, Middleton A, Scott E, Nelson S, Corpas M, Bonhomme N, Rodriguez LL, Murtagh M, Kleiderman E, Genomics obotPVTTotGAf, Health (2018) APPLaUD: access for patients and participants to individual level uninterpreted genomic data. Human Genomics 12: 7. doi: 10.1186/s40246-018-0139-5 [PubMed: 29454384]

University JH (2016) Command Line Tools for Genomic Data Science. Coursera https://www.coursera.org/learn/genomic-tools. Accessed 01/08/2019 2019

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43: 11.10.1–11.10.33. doi: doi:10.1002/0471250953.bi1110s43 [PubMed: 25431634]

Woolley JP, McGowan ML, Teare HJA, Coathup V, Fishman JR, Settersten RA, Sterckx S, Kaye J, Juengst ET (2016) Citizen science or scientific citizenship? Disentangling the uses of public engagement rhetoric in national research initiatives. BMC Medical Ethics 17: 33. doi: 10.1186/s12910-016-0117-1 [PubMed: 27260081]

Yu J-H, Crouch J, Jamal SM, Bamshad MJ, Tabor HK (2014) Attitudes of non-African American focus group participants toward return of results from exome and whole genome sequencing. American Journal of Medical Genetics Part A 164: 2153–2160. doi: doi:10.1002/ajmg.a.36610

Zalunin V (2012) CRAM. European Nucleotide Archive. https://www.ebi.ac.uk/ena/software/cram-toolkit. Accessed 01/08/2019 2019

Zhu Z, Zhang Y, Ji Z, He S, Yang X (2015) High-throughput DNA sequence data compression. Brief Bioinform 16: 1–15. doi: 10.1093/bib/bbt087 [PubMed: 24300111]

Table 1.

Libraries and software associated with generating and manipulating mapped and annotated sequencing data files.

Website links associated with the tools and example tutorials are given, where available. CRAMtools is no longer being supported (https://github.com/enasequence/cramtools), and users are being directed to SAMtools.

Library, Software, or Platform	Example Tutorials
Bam2mpg	-
https://github.com/nhansen/bam2mpg	
BCFtools	-
https://samtools.github.io/bcftools/bcftools.html	
CRAMtools	http://www.htslib.org/workflow/
	https://www.ebi.ac.uk/ena/software/cram-usage
	https://www.ebi.ac.uk/ena/support/cram-tutorial
	https://software.broadinstitute.org/gatk/best-practices/
Galaxy	https://galaxyproject.org/tutorials/ngs/
https://usegalaxy.org/	https://galaxyproject.org/learn/
Genome Analysis Toolkit (GATK)	https://software.broadinstitute.org/gatk/documentation/topic?name=tutorials
https://software.broadinstitute.org/gatk/	
Freebayes	http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html
https://github.com/ekg/freebayes	
Picard	-
https://broadinstitute.github.io/picard/	
SAMtools	http://biobits.org/samtools_primer.html
http://samtools.sourceforge.net/	https://samtools.github.io/hts-specs/SAMv1.pdf
http://www.htslib.org/	http://quinlanlab.org/tutorials/samtools/samtools.html
	http://www.htslib.org/doc/samtools.html
Scramble	-
https://sourceforge.net/projects/staden/files/io_lib/	
VarAFT	https://varaft.eu/
https://varaft.eu/	
VCFtools	https://faculty.washington.edu/browning/intro-to-vcf.html
https://vcftools.github.io/index.html	https://github.com/opencb/hpg-variant/wiki/VCF-Tools-tutorial
VCF-Miner	-
http://bioinformaticstools.mayo.edu/research/vcf-miner/	
VCF.Filter	-
https://biomedical-sequencing.at/VCFFilter/	