frontiers
in Genetics

METHODS
published: 24 September 2020
doi: 10.3389/fgene.2020.518644

GenomeChronicler: The Personal Genome Project UK Genomic Report Generator Pipeline

José Afonso Guerra-Assunção[1,2]*, *Lucia Conde*[2], *Ismail Moghul*[3], *Amy P. Webster*[3], *Simone Ecker*[3], *Olga Chervova*[3], *Christina Chatzipantsiou*[4], *Pablo P. Prieto*[4], *Stephan Beck*[3] *and Javier Herrero*[2]

[1] Infection and Immunity, University College London, London, United Kingdom, [2] Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London, United Kingdom, [3] Medical Genomics, UCL Cancer Institute, University College London, London, United Kingdom, [4] Lifebit, London, United Kingdom

OPEN ACCESS

Edited by:
*Athanasios Alexiou,
Novel Global Community Educational Foundation (NGCEF), Hebersham, Australia*

Reviewed by:
*Chih-hsing Ho,
Academia Sinica, Taiwan
Margaret Woodhouse,
United States Department of Agriculture, Agricultural Research Service, United States*

***Correspondence:**
*José Afonso Guerra-Assunção
a.guerra@ucl.ac.uk*

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 December 2019*
Accepted: *02 September 2020*
Published: *24 September 2020*

Citation:
Guerra-Assunção JA, Conde L, Moghul I, Webster AP, Ecker S, Chervova O, Chatzipantsiou C, Prieto PP, Beck S and Herrero J (2020) GenomeChronicler: The Personal Genome Project UK Genomic Report Generator Pipeline. Front. Genet. 11:518644. doi: 10.3389/fgene.2020.518644

In recent years, there has been a significant increase in whole genome sequencing data of individual genomes produced by research projects as well as direct to consumer service providers. While many of these sources provide their users with an interpretation of the data, there is a lack of free, open tools for generating reports exploring the data in an easy to understand manner. GenomeChronicler was developed as part of the Personal Genome Project UK (PGP-UK) to address this need. PGP-UK provides genomic, transcriptomic, epigenomic and self-reported phenotypic data under an open-access model with full ethical approval. As a result, the reports generated by GenomeChronicler are intended for research purposes only and include information relating to potentially beneficial and potentially harmful variants, but without clinical curation. GenomeChronicler can be used with data from whole genome or whole exome sequencing, producing a genome report containing information on variant statistics, ancestry and known associated phenotypic traits. Example reports are available from the PGP-UK data page (personalgenomes.org.uk/data). The objective of this method is to leverage existing resources to find known phenotypes associated with the genotypes detected in each sample. The provided trait data is based primarily upon information available in SNPedia, but also collates data from ClinVar, GETevidence, and gnomAD to provide additional details on potential health implications, presence of genotype in other PGP participants and population frequency of each genotype. The analysis can be run in a self-contained environment without requiring internet access, making it a good choice for cases where privacy is essential or desired: any third party project can embed GenomeChronicler within their off-line safe-haven environments. GenomeChronicler can be run for one sample at a time, or in parallel making use of the Nextflow workflow manager. The source code is available from GitHub (https://github.com/PGP-UK/GenomeChronicler), container recipes are available for Docker and Singularity, as well as a pre-built container from SingularityHub (https://singularity-hub.org/collections/3664) enabling easy deployment in a variety of settings. Users without access to computational resources to run GenomeChronicler can access the software from the Lifebit CloudOS platform (https://lifebit.ai/cloudos) enabling the production of reports and variant calls from raw sequencing data in a scalable fashion.

Keywords: personal genomics, PGP-UK, genomic report, open consent, participant engagement, open source, cloud computing

INTRODUCTION

The publication of the first draft human genome sequence (International Human Genome Sequencing Consortium, 2001) promised a revolution in knowledge of how we see ourselves as individuals and how future medical care should take our genetic background into account. Almost ten years later, the perspective of widespread personal genomics was still to be achieved (Venter, 2010).

Following the establishment of 23andMe and others from 2007 onward, there is now a wide range of easily accessible clinical and non-clinical genetic tests that are routinely employed to detect individuals' carrier status for certain disease genes or particular mutations of clinical relevance. Many more associations between genotype and phenotype have been highlighted by research, sometimes with uncertain clinical relevance or simply describing personal traits such as eye color (Pontikos et al., 2017; Kuleshov et al., 2019).

Over the past few years, we have seen a dramatic reduction of the cost to sequence the full human genome. This reduction in cost enables many more projects to start using whole genome sequencing (WGS) approaches, as well as the marked rise in the number of personal genomes being sequenced.

Personal genomics is very much a part of the public consciousness as can be seen by the rampant rise in direct to consumer (DTC) genomic analysis offerings on the market. In this context, it is unsurprising that the analysis of one's own genome provides a valuable educational opportunity (Salari et al., 2013; Linderman et al., 2018) as well as increasing participant engagement as part of biomedical trials (Sanderson et al., 2016).

The Personal Genome Project (PGP) set up by George Church in 2005 is the earliest initiative enabled by the increased popularity of whole genome sequencing and its lowering costs. The global PGP network currently consists of 5 projects spread around the world, managed independently but joined by a common goal of providing open access data containing genomic, environmental and trait information[1].

Data analysis within PGP-UK poses important ethical challenges, as all the data and genome reports are intended to become freely and openly available on the World Wide Web. However, until the completion and approval of the reports, the data must be treated as confidential private information. Prior to enrollment, all participants are well informed through an online study guide and tested for their understanding of the potential risks of participating in a project of this nature. Upon receipt of their report, participants have a cool-off period of four weeks to explore their data and reports and to seek all the required clarifications. During that time, they can trigger the release of their report and data themselves by selecting the 'release immediately' option in their personal accounts. To date, 67% of participants have selected this release option. They also have the option to withdraw from the study in which case no release occurs and all data will be deleted. This option has never been selected by any participant. If neither of these options are chosen, the data and reports are released automatically by the end of the cool-off period.

There are several resources aimed at users of DTC genetic testing companies on the internet including Promethease (2019) and Genomelink (2019). There are some other tools with a focus on clinical aspects or particular diseases (Nakken et al., 2018), as well as academic databases containing genotypes of other individuals (Greshake et al., 2014), pharmacogenomic information (Klein and Ritchie, 2018) or genotype to phenotype links (Ramos et al., 2014; Pontikos et al., 2017; Kuleshov et al., 2019) that can be useful for the interpretation of personal genomes. Many of these are linked into resources like SNPedia (Cariaso and Lennon, 2012), allowing a wide range of exploration options for the known associations of each genotype from multiple perspectives.

Surprisingly, we found no pre-existing solution that would allow the annotation and evaluation of variants on the whole genome level, assessment of ancestry and more focused analysis of variants that have been previously associated with specific phenotypes. In particular, one that could be run locally ensuring full control of the data before the results are scrutinized and approved.

GenomeChronicler represents, to the best of our knowledge, the first pipeline that can be run offline or in the cloud, to generate personal genomics reports that are not limited to disease only, from whole genome or whole exome sequencing data.

GenomeChronicler contains a database of positions of interest for ancestry or phenotype. The genotype at each of these positions is inferred from the user provided data that has been mapped to the human genome. These genotypes are then compared to local versions of a series of publicly available resources to infer ancestry and likely phenotypes for each individual participant. These results are then presented as a PDF document containing hyperlinks where more information about each variant and phenotype can be found. A visual representation of the pipeline and its underlying resources is shown in **Figure 1**.

This pipeline will continue to be improved and expanded by PGP-UK, e.g., to include methylome and transcriptome

[1] https://www.personalgenomes.org/

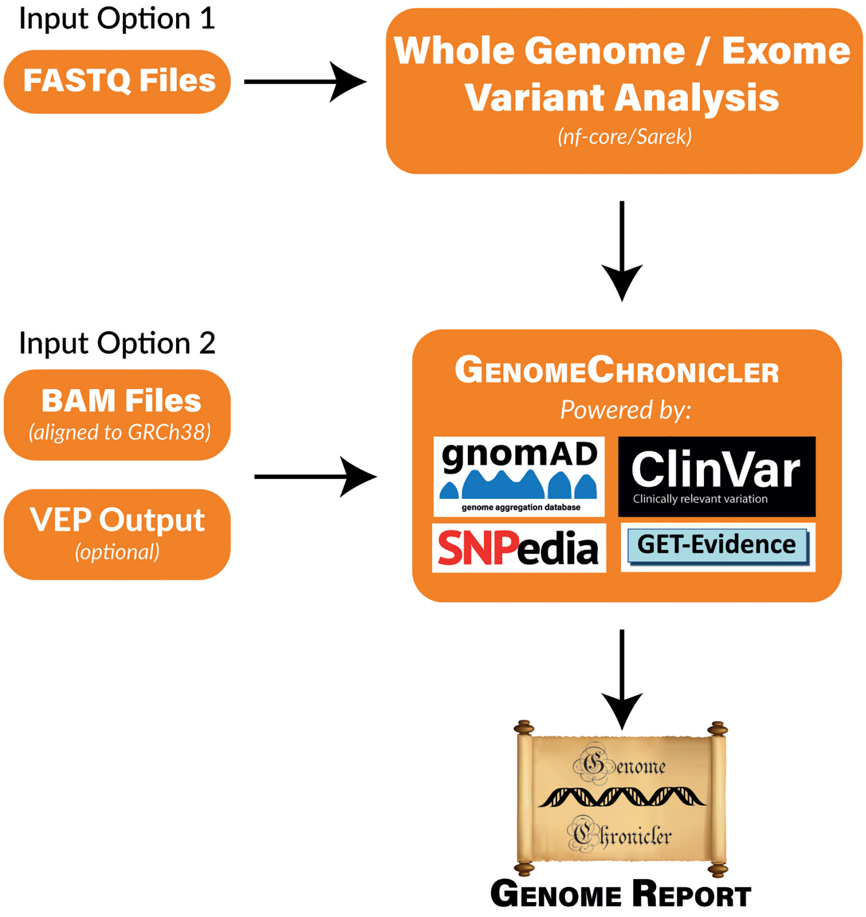

FIGURE 1 | Flow Diagram of GenomeChronicler processing pipeline, illustrating the multiple entry points for the pipeline, resources integrated by default and generated outcomes. Either entry point of the pipeline can be run locally in a single machine, as a Nextflow workflow or in the Cloud. All source code and integrations are freely available in their respective GitHub repositories. The stand-alone GenomeChronicler is available at (https://github.com/PGP-UK/GenomeChronicler), the integration of GenomeChronicler with Nextflow is available at (https://github.com/PGP-UK/GenomeChronicler-nf) and the combined GenomeChronicler with Sarek variant calling is available at (https://github.com/PGP-UK/GenomeChronicler-Sarek-nf). The recipe files for the Docker and Singularity containers are available within the respective GitHub repositories. The resource logos are reproduced from the respective resource websites and remain copyright of their original owner.

reports (Beck et al., 2018). We envision this project will also be useful to other research endeavors that want to provide personal genomes information to their participants to increase engagement; e.g., to altruistic individuals who have obtained their whole genome sequencing data from a DTC or health care provider and are looking for an ethics-approved framework to share their data. PGP-UK already supports this through their Genome Donation program.

MATERIALS AND METHODS

Data Input

The GenomeChronicler pipeline was designed to run downstream of a standardized germline variant calling pipeline. GenomeChronicler requires a pre-processed BAM or CRAM file with deduplication and quality recalibrated alignments against the GRCh38 genome assembly and optionally, the summary HTML report produced by the Ensembl Variant Effect Predictor (McLaren et al., 2016).

GenomeChronicler can be run with any variant caller provided that the reference dataset is matched to the reference genome used (the included GenomeChronicler databases currently use GRCh38). It is also imperative, to obtain good quality results, that the BAM or CRAM files used have had their duplicates removed and quality recalibrated prior to being used for GenomeChronicler.

To simplify this entire process and to make the tool more accessible to users who may not know how to run a germline variant calling pipeline, GenomeChronicler can also be run in a fully automated mode from the raw sequencing data, where the germline variant calling pipeline is also run and the whole process is managed by the Nextflow workflow management system (Di Tommaso et al., 2017). In this scenario, GenomeChronicler uses the Sarek pipeline[2] (Garcia et al., 2020) to process raw FASTQ files

[2] https://github.com/nf-core/sarek

in a manner that follows the GATK variant calling best practices guidelines (Van der Auwera et al., 2013). Manual inspection of the initial quality control steps of Sarek is recommended prior to perusing the final results.

The combined version of Sarek + GenomeChronicler written using the Nextflow workflow manager (Di Tommaso et al., 2017) is available both on Github[3] and on Lifebit CloudOS.

Ancestry Inference

We infer the ancestry of each individual through a Principal Components Analysis (PCA) which is a widely used approach for identifying ancestry similarities among individuals (Novembre et al., 2008).

For each sample of interest, we intersect the genotypes with a reference dataset consisting of genotypes from the 1000 Genomes Project samples (The 1000 Genomes Project Consortium, 2015), containing individuals from 26 different worldwide populations and applying PCA on the resulting genotype matrix.

The reference samples from the 1000 Genome Project are filtered to keep only unrelated individuals. In order to avoid strand issues when merging the datasets, all ambiguous (A/T and C/G) SNPs were removed, as well as non-biallelic SNPs, SNPs with $> 5\%$ of missing data, rare variants ($MAF < 0.05$) and SNPs out of Hardy-Weinberg equilibrium ($pval < 0.0001$). From the remaining SNPs, a subset of unlinked SNPs are selected by pruning those with $r2 > 0.1$ using 100-SNP windows shifted at 5-SNP intervals.

These genotypes are used to run PCA based on the variance-standardized relationship matrix, selecting twenty as the number of PCs to be extracted. We then project the data over the first three principal components to identify clusters of populations and highlight the sample of unknown ancestry on the resulting plot.

Here, we used PLINK (Purcell et al., 2007) to process the genotype data and the R Statistical Computing platform for plotting the final PCA figures to illustrate the ancestry of each sample. An example of the distribution of the reference samples on the PCA is shown in **Figure 2**.

Variant Annotation Databases

SNPedia

SNPedia (Cariaso and Lennon, 2012) is a large public repository of manually added as well as automatically mined genotype to phenotype links sourced from existing literature. SNPedia is the core resource behind the phenotype tables in GenomeChronicler; it provides annotations for both single-gene phenotypes as well as for a few phenotypes involving multiple loci referred to as genosets in the produced reports.

ClinVar

ClinVar (Landrum and Kattman, 2018) is a database hosted by the NCBI that focuses exclusively on variants related to health and has been running since 2013. In comparison to SNPedia, ClinVar is a much smaller database but it is closely linked to the clinical relevance of each variant. ClinVar is curated more strictly with a clinical review – something unique among the data sources used by GenomeChronicler.

GETevidence

GETevidence was developed as part of the Personal Genome Project Harvard (Mao et al., 2016) to showcase the variants present within its participants and to allow manual annotation and interpretations of the results. For some of the genotypes present, it also contains manual annotations that have been added by the users or curation team. GETevidence allows individuals to compare their genotypes with those from other personal genomes available within the PGP-Harvard project.

gnomAD

Spanning several human populations, the Genome Aggregation Database (gnomAD) (Karczewski et al., 2019) aggregates data from multiple sources to produce an atlas of variation across the human genome. Extensively annotated and now covering most of the latest assembly of the human genome, these links enable easy access to information such as allele frequencies for the genotype across different populations around the world, as well as some annotation context for each variant, regarding potential effect on genes if relevant and how selection forces are constraining the genomic locus.

Database Availability, Building and Update

The underlying databases required to run GenomeChronicler are provided within the package. A set of scripts to regenerate these SQLite databases is also provided within the source code. The datasets are limited to positions of interest is compiled so that when genotyping is performed only relevant positions are computed to save computational time.

SNPedia provides an API to query its records in a systematic way. The other linked databases provide regular dumps of the whole dataset, enabling easy assessment for which dbSNP rs identifiers are represented within the full database. The use of rs identifiers and genotypes to link between the different databases enables an unambiguous way to compare information between different resources.

Genotype Assessment and Reporting

Typical germline variant calling pipelines result in a VCF file where positions that match the reference sequence are not reported. Homozygous reference genotypes thus become indistinguishable from positions in the genome where there is no read coverage.

To ensure comparable results between runs, genotype VCFs (gVCFs) instead of VCFs are computed during each run of GenomeChronicler, but only for a subset of genomic positions that informative for ancestry inference or phenotype annotation, saving computational time and storage space.

The Genome Report Template

GenomeChronicler is designed in a modular way where the final report is only compiled at the end, integrating all the results.

[3] https://github.com/PGP-UK/GenomeChronicler-Sarek-nf

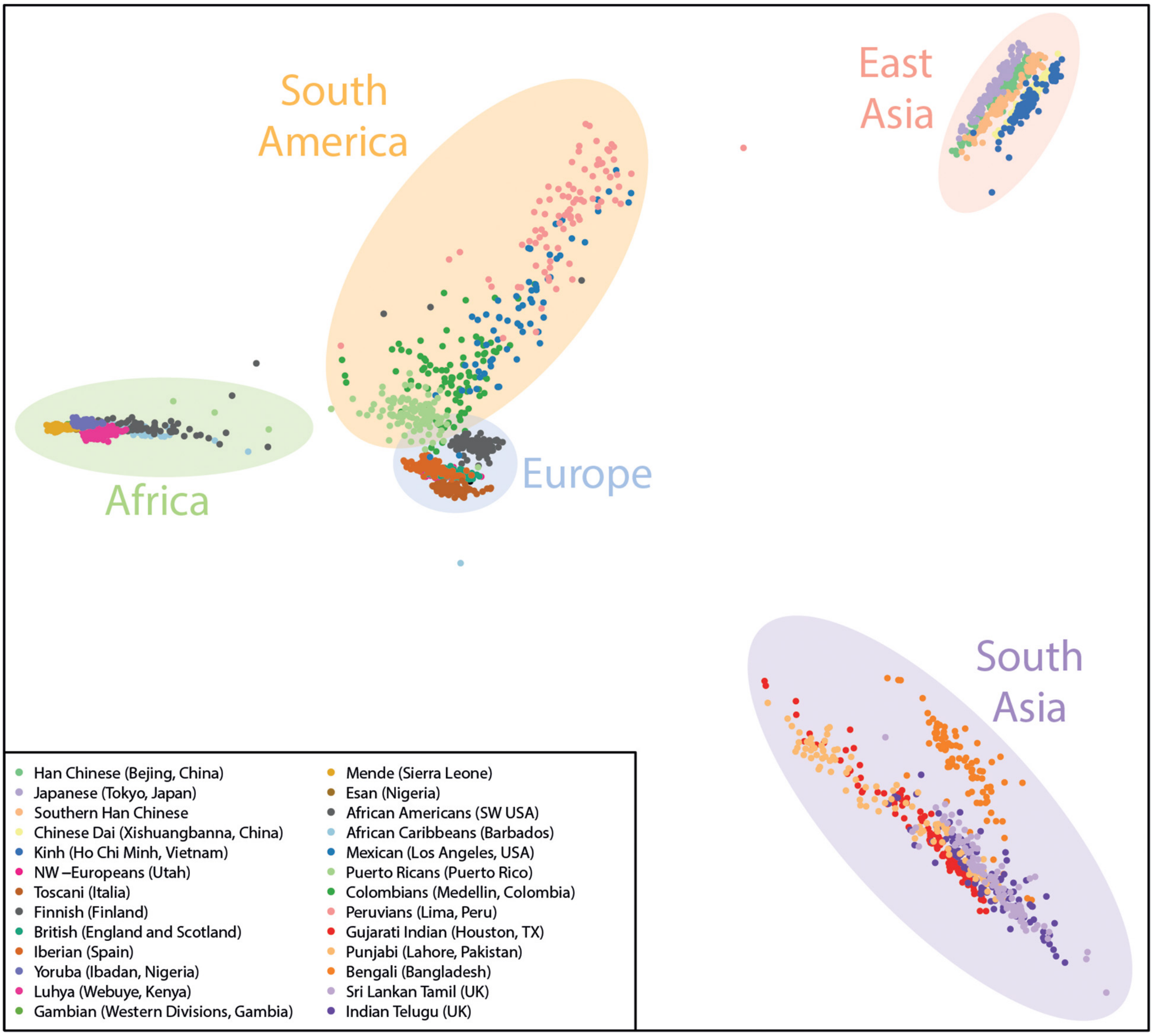

FIGURE 2 | Example Ancestry PCA plot containing the current reference data from the 1000 genomes project used by GenomeChronicler, with shaded areas broadly illustrating the origin of the populations represented.

To customize the report layout, the content and the amount of extra information, GenomeChronicler uses a template file written in LaTeX. For example, one can modify the branding and introductory text of the report, integrate custom or third-party analyses provided the results are in a format that can be typeset using LaTeX, omit certain sections, or even modify the structure of the report produced.

Output Files

The main output of GenomeChronicler is a report in PDF format, containing information from all sections of the pipeline that have run as set by the LaTeX template provided when running the script. Additionally, an Excel file containing the genotype phenotype link information, and all corresponding hyperlinks is also produced, allowing the user to explore the results in a familiar environment. While most intermediate files are automatically removed at the end of the GenomeChronicler run, the original PDF version of the ancestry PCA plot, as well as a file containing the sample name, genotyping results and pipeline log files are retained within the results directory to ease automation.

Pipeline Validation

To further validate the pipeline, 1000 Genome Project generated illumina data for sample NA12878 was used. Genomic data for sample NA12878 mapped to the human reference genome (GRCh38) was retrieved from the 1000 Genome Project[4] and

[4] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/CEU/NA12878/alignment/

converted to BAM file using the SAMtools toolkit. High confidence genotype calls were retrieved from Genome-in-a-Bottle[5]. The GenomeChronicler pipeline was run on the data, and the resulting genotype calls in high confidence regions were compared to the reference calls using BCFtools to assess genotype concordance.

Running GenomeChronicler

In line with the PGP-UK data, all the code for GenomeChronicler is freely available. To make it easier to implement, several options are available to eliminate the need for installing dependencies and underlying packages, or even the need to have access to computer hardware capable of handling the processing of a human genome. The range of options available is detailed below and illustrated in **Figure 1**.

Running GenomeChronicler Locally

From the available source code

The source code for GenomeChronicler is available on GitHub at https://github.com/PGP-UK/GenomeChronicler. A setup script is included to automatically download the pre-compiled accessory databases and other required data. Software dependencies including LaTeX, R and Perl need to be installed independently if not using the Singularity container. The provided Singularity recipe file provides a useful list of required packages, in particular for those installing it on a Debian/Ubuntu based system.

Using a pre-compiled container

GenomeChronicler is also available as a Singularity container (Kurtzer et al., 2017) with all dependencies pre-installed and ready. This can be obtained from SingularityHub (Sochat et al., 2017) by running the command: singularity pull "shub://PGP-UK/GenomeChronicler" on any machine that has Singularity installed.

Once downloaded, the main script (GenomeChronicler_mainDruid.pl) can be run with the desired data and options to produce genome reports.

Running GenomeChronicler on Cloud

To enable reproducible, massively parallel, cloud native analyses, GenomeChronicler has also been implemented as a Nextflow pipeline. The implementation abstracts the installation overhead from the end user, as all the dependencies are already available via pre-built containers, integrated seamlessly in the Nextflow pipeline. The source code for this implementation is available on GitHub at https://github.com/PGP-UK/GenomeChronicler-nf, as a standalone Nextflow process.

To provide an end-to-end FASTQ to PGP-UK reports pipeline, we also implemented an integration of GenomeChronicler, with a curated and widely used by the bioinformatics community pipeline, namely Sarek (Ewels et al., 2019; Garcia et al., 2020). This PGP-UK implementation of Sarek is available on GitHub at https://github.com/PGP-UK/GenomeChronicler-Sarek-nf.

The aforementioned pipeline is available in the collection of curated pipelines on the Lifebit CloudOS platform[6]. Lifebit CloudOS enables users without any prior cloud computing knowledge to deploy analysis in the cloud. In order to run the pipeline, the user only needs to specify input files, desired parameters and select resources from an intuitive graphical user interface. After the completion of the analysis on Lifebit CloudOS, the user has a permanent shareable live link that includes performance and file metadata, the associated GitHub repository revision and also links to the generated results. The relevant analysis page can be used to repeat the exact same analysis. The analysis page for the PGP-UK user with id uk35C650 can be accessed in the following permalink https://cloudos.lifebit.ai/public/jobs/5e74d60babdee600f94df39b. Each analysis can have different privacy settings allowing the user to choose if the results are publicly visible, making it easier for sharing or private use, thus maintaining data confidentiality.

RESULTS

The main resulting document is a PDF file which contains sections related to variants of unknown significance, ancestry estimation (as exemplified in **Figure 2**) and variants with associated phenotypes, separated by either potentially beneficial or potentially harmful phenotypes as well as phenotypes affected by multiple variants, referred to as genosets (Cariaso and Lennon, 2012).

Initial versions of the GenomeChronicler pipeline were validated by comparing its results to those provided by DTC company 23andMe for participant PGP-UK1, as well as phenotype feedback from the pilot participants (Beck et al., 2018).

Further validations was done using sample NA12878, which is an often-analyzed as a benchmark reference for personal genomics.

The GATK genotype calls produced as part of GenomeChronicler were directly compared to the high confidence variant calling for the sample as part of the Genome-in-a-Bottle consortium (Zook et al., 2014). The concordance rate was 99.97% at the genotype level, resulting in no phenotype changes.

Sample NA12878 is part of pedigree 1463 from the HapMap project and is known to correspond to a female individual of CEPH ancestry. These are correctly reflected in the ancestry and genoset sections of the GenomeChronicler report.

To date, more than one hundred such reports have been produced and made available as part of the PGP-UK (Beck et al., 2018). They are publicly available in the PGP-UK open access data page[7]. This collection contributes to the educational potential of the project as a whole. On one hand, it allows participants of PGP-UK and other users of the GenomeChronicler tool to compare their genome report results to those of other individuals. On the other hand, it allows

[5] ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/

[6] https://cloudos.lifebit.ai/

[7] https://www.personalgenomes.org.uk/data/

individuals that are interested in the subject but did not have their genome sequenced to explore the kind of information that one can learn from a personal genome.

While the method presented here focuses on the analysis of the genomic data (whole genome and whole exome), PGP-UK also contains multi-omics data, including RNAseq and methylation data, as well as genotype data sourced elsewhere (e.g., 23andMe) and deposited by the participants.

Methods such as GenomeChronicler allow other research projects in possession of personal genome data to easily produce genome reports, customize them with static text providing information about the project that can differ from the default template file, or even add links to other relevant databases.

CONCLUSION

Here we present GenomeChronicler, a computational pipeline to produce genome reports including variant calling summary data, ancestry inference, and phenotype annotation from genotype data for personal genomics data obtained through whole genome or whole exome sequencing.

The pipeline is modular, fully open source, and available as containers and on the Lifebit CloudOS computing platform, enabling easy integration with other projects, regardless of available computational resources and bioinformatics expertise.

The pipeline presented here incorporates a range of well-established open source resources, which have been validated independently in different scenarios (Garcia et al., 2020). We have also cross-referenced the data produced by this pipeline to ensure it is providing a coherent output (Chervova et al., 2019).

While we follow the GATK best practices, as implemented in Sarek, to produce an accurate and reliable variant call set, unforeseen sources of error can be introduced at the sequencing stage, resulting in the pipeline potentially calling an erroneous genotype at a certain genomic position.

Finally, the interpretation of genotype to phenotype links is heavily context-dependent and fraught with its own challenges. Recognizing that this task requires experience and/or cognitive abilities that cannot be imparted on an automated computer system, we instead opted for providing a report that focuses on the biomedical and phenotypic associations obtained through SNPedia (Cariaso and Lennon, 2012), supplemented with hyperlinks to a wide range of other databases. This allows the user to explore the results and the supporting research data in more depth if desired. Some of the reported links between genotypes and phenotypes have been strongly validated by multiple research groups over the years, while others are not as well supported, and as such, require careful interpretation by the user.

This work was developed as part of PGP-UK and incorporates feedback from early participants to improve the usefulness of the reports produced, and of participant engagement. It is designed to be easily expandable, adaptable to other contexts and most of all, suitable for projects with a wide range of ethical requirements, from those that need the data to be processed inside a safe-haven environment to those that process all the data in the public domain. It can also be of interest to educational groups such as Open Humans (Greshake et al., 2019). Open Humans[8] is a vibrant community of researchers, patients, data and citizen scientists who want to learn more about themselves.

For PGP-UK participants, there is a well-established ethical framework that ensures that participants are aware of the limitations of the information they receive. It also makes provision for the project to refrain from issuing reports if the quality of the input data fails the quality control stage.

Personal genomics has become a public commodity and individuals can access their own or even someone else's genome. It is important to note that GenomeChronicler is essentially a tool that collates information from different sources but is not suitable for the clinical interpretation of the results. Indeed, inaccurate interpretation might result from poor quality genomic data or unreliable annotations. However, the potential for negative consequences should be minimal provided the users heed the stated recommendations of not relying on this tool for clinical decision making.

Future directions for this work will include the integration of other omics data types that are produced within PGP-UK, as well as potentially expanding the databases that are linked by default when running the pipeline.

We hope that GenomeChronicler will be useful to other projects and interested individuals. As it is open source, the pipeline can easily adapt custom templates to satisfy any curiosity-driven analyses and increase the level of genomic understanding in general.

DATA AVAILABILITY STATEMENT

The datasets analyzed and used for the development of the approach here described are deposited at the European Nucleotide Archive (ENA) hosted by the EMBL-EBI under the umbrella accession PRJEB24961 [https://www.ebi.ac.uk/ena/data/view/PRJEB24961]. The PGP-UK pilot data was described in a data descriptor published in Scientific Data (Chervova et al., 2019). The source code for the software is deposited and maintained in GitHub and available at [https://github.com/PGP-UK/GenomeChronicler]. The Nextflow integrated version is available at [https://github.com/PGP-UK/GenomeChronicler-nf] and finally, the version also containing the Sarek variant calling pipeline is available at [https://github.com/PGP-UK/GenomeChronicler-Sarek-nf]. Reports generated using this approach for PGP-UK samples are archived in the PGP-UK data page https://www.personalgenomes.org.uk/data.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UCL Research Ethics Committee (ID number 4700/001). The patients/participants provided their written informed consent to participate in this study.

[8] https://www.openhumans.org/

AUTHOR CONTRIBUTIONS

JAG-A led the development and implementation of the method and wrote the manuscript with input from all authors. JAG-A and LC contributed computer code. CC contributed the Nextflow and Lifebit CloudOS integrations with support from PP. JAG-A, LC, IM, APW, SE, JH, OC, and SB contributed to the conceptual development of the method and usability. All authors read and approved the manuscript.

FUNDING

PGP-UK gratefully acknowledges support from the Frances and Augustus Newman Foundation, Dangoor Education and the National Institute for Health Research (NIHR) UCLH Biomedical Research Centre (BRC369/CN/SB/101310). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

ACKNOWLEDGMENTS

The authors acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL) and associated support services. The authors thank all PGP-UK participants for their contributions to the project and feedback on the items that feature in the reports produced by GenomeChronicler. This manuscript has also been submitted as a preprint to the bioRxiv preprint server, available at https://www.biorxiv.org/content/10.1101/2020.01.06.873026v3.

REFERENCES

Beck, S., Alison, M. B., Graham, B., Maggie, B., Martin, J. C., Olga, C., et al. (2018). Personal genome project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. *BMC Med. Genom.* 11:108. doi: 10.1186/s12920-018-0423-1

Cariaso, M., and Lennon, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* 40, D1308–D1312. doi: 10.1093/nar/gkr798

Chervova, O., Lucia, C., José, A. G.-A., Ismail, M., Amy, P. W., Alison, B., et al. (2019). The personal genome project-UK, an open access resource of human multi-omics data. *Sci. Data* 6, 1–10. doi: 10.1038/s41597-019-0205-4

Di Tommaso, P., Maria, C., Evan, W. F., Pablo, P. B., Emilio, P., and Cedric, N. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Ewels, P. A., Alexander, P., Sven, F., Johannes, A., Harshil, P., Andreas, W., et al. (2019). Nf-Core: community curated bioinformatics pipelines. *BioRxiv*. doi: 10.1101/610741

Garcia, M., Szilveszter, J., Malin, L. P. I., Olason, M. M., Jesper, E., Sebastian, D. L., et al. (2020). Sarek: a portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* 9:63. doi: 10.12688/f1000research.16665.1

Genomelink (2019). *Upload Raw DNA Data for Free Analysis On 25 Traits*. Available online at: https://genomelink.io/ (accessed November 19, 2019).

Greshake, B., Bayer, P. E., Rausch, H., and Reda, J. (2014). OpenSNP–a crowdsourced web resource for personal genomics. *PLoS One* 9:e89204. doi: 10.1371/journal.pone.0089204

Greshake, T., Bastian, M. A., Kevin, A., Mairi, D., Vero, E.-G., Beau, G., et al. (2019). Open humans: a platform for participant-centered research and personal data exploration. *GigaScience* 8:giz076. doi: 10.1093/gigascience/giz076

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Karczewski, K. J., Laurent, C. F., Grace, T., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* . doi: 10.1101/531210

Klein, T. E., and Ritchie, D. M. (2018). PharmCAT: a pharmacogenomics clinical annotation tool. *Clin. Pharmacol. Therapeut.* 104, 19–22. doi: 10.1002/cpt.928

Kuleshov, V., Jialin, D., Christopher, V., Braden, H., Alexander, R., Yang, L., et al. (2019). A machine-compiled database of genome-wide association studies. *Nat. Commun.* 10, 1–8. doi: 10.1038/s41467-019-11026-x

Kurtzer, G. M., Vanessa, S., and Michael, W. B. (2017). Singularity: scientific containers for mobility of compute. *PLoS One* 12:e0177459. doi: 10.1371/journal.pone.0177459

Landrum, M. J., and Kattman, L. B. (2018). ClinVar at five years: delivering on the promise. *Hum. Mutat.* 39, 1623–1630. doi: 10.1002/humu.23641

Linderman, M. D., Saskia, C. S., Ali, B., George, A. D., Andrew, K., Randi, Z., et al. (2018). Impacts of incorporating personal genome sequencing into graduate genomics education: a longitudinal study over three course years. *BMC Med. Genom.* 11:5. doi: 10.1186/s12920-018-0319-0

Mao, Q., Serban, C., Zhang, R. Y., Ball, M. P., Chin, R., Carnevali, P., et al. (2016). The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *GigaScience* 5:42. doi: 10.1186/s13742-016-0148-z

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

Nakken, S., Fournous, G., Vodák, D., Aasheim, L. B., Myklebost, O., and Hovig, E. (2018). Personal cancer genome reporter: variant interpretation report for precision oncology. *Bioinformatics (Oxf. Engl.)* 34, 1778–1780. doi: 10.1093/bioinformatics/btx817

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., et al. (2008). Genes mirror geography within europe. *Nature* 456, 98–101. doi: 10.1038/nature07331

Pontikos, N., Yu, J., Moghul, I., Withington, L., Blanco-Kelly, F., Vulliamy, T., et al. (2017). Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics* 33, 2421–2423. doi: 10.1093/bioinformatics/btx147

Promethease(2019) Available online at: https://www.promethease.com/ (accessed November 19, 2019).

Purcell, S., Benjamin, N., Kathe, T.-B., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ramos, E. M., Hoffman, D., Junkins, H. A., Maglott, D., Phan, L., Sherry, S. T., et al. (2014). Phenotype–genotype integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* 22, 144–147. doi: 10.1038/ejhg.2013.96

Salari, K., Karczewski, K. J., Hudgins, L., and Ormond, K. E. (2013). Evidence that personal genome testing enhances student learning in a course on genomics and personalized medicine. *PLoS One* 8:e68853. doi: 10.1371/journal.pone.0068853

Sanderson, S. C., Linderman, M. D., Suckiel, S. A., Diaz, G. A., Zinberg, R. E., Ferryman, K., et al. (2016). Motivations, concerns and preferences of personal genome sequencing research participants: baseline findings from the healthseq project. *Eur. J. Hum. Genet.* 24, 14–20. doi: 10.1038/ejhg.2015.118

Sochat, V. V., Prybol, C. J., and Kurtzer, G. M. (2017). Enhancing reproducibility in scientific computing: metrics and registry for singularity containers. *PLoS One* 12:e0188511. doi: 10.1371/journal.pone.0188511

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Ami, L.-M., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Prot. Bioinform.* 11, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Venter, J. C. (2010). Multiple personal genomes await. *Nature* 464, 676–677. doi: 10.1038/464676a

Zook, J. M., Brad, C., Wang, J., Mittelman, D., Hofmann, O., Hide, W., et al. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32, 246–251. doi: 10.1038/nbt.2835

Conflict of Interest: PP is CTO of Lifebit and CC is an employee of Lifebit.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.