PERSONAL GENOMES: ACCESSING, SHARING, AND INTERPRETATION

EDITED BY: Manuel Corpas, Stephan Beck, Gustavo Glusman and Mahsa Shabani

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88971-127-7
DOI 10.3389/978-2-88971-127-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

PERSONAL GENOMES: ACCESSING, SHARING, AND INTERPRETATION

Topic Editors:
Manuel Corpas, Cambridge Precision Medicine, United Kingdom
Stephan Beck, University College London, United Kingdom
Gustavo Glusman, Institute for Systems Biology (ISB), United States
Mahsa Shabani, Ghent University, Belgium

Citation: Corpas, M., Beck, S., Glusman, G., Shabani M., eds. (2021). Personal Genomes: Accessing, Sharing, and Interpretation. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-127-7

Table of Contents

frontiers in Genetics

EDITORIAL
published: 04 June 2021
doi: 10.3389/fgene.2021.687584

Editorial: Personal Genomes: Accessing, Sharing, and Interpretation

Manuel Corpas[1,2,3]*, *Stephan Beck*[4], *Gustavo Glusman*[5] *and Mahsa Shabani*[6]

[1] Cambridge Precision Medicine Limited, ideaSpace, University of Cambridge Biomedical Innovation Hub, Cambridge, United Kingdom, [2] Department of Madingley Hall, Institute of Continuing Education, University of Cambridge, Cambridge, United Kingdom, [3] Facultad de Ciencias de la Salud, Universidad Internacional de La Rioja, Madrid, Spain, [4] Department of Cancer Biology, University College London (UCL) Cancer Institute, University College London, London, United Kingdom, [5] Institute for Systems Biology (ISB), Seattle, WA, United States, [6] Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium

Keywords: personal genomics, direct-to-consumer genetic testing, citizen science, data sharing, ethical, legal, social implications

Editorial on the Research Topic

Personal Genomes: Accessing, Sharing, and Interpretation

Over the past few years we have witnessed a number of advances in the personal genomics space including (a) more affordable sequencing technology, (b) mainstreaming of genomics in healthcare systems, (c) augmented sharing of genomic data, and (d) increased demand for direct-to-consumer genetic testing. All of these developments have brought us closer to the long-awaited genomics revolution. This genomics revolution is not exempt from challenges, in part amplified by lack of standards (ethical, legal, and technological), slow translation of knowledge to the clinic, and unequal access of personal genome benefits for all.

As the vast majority of reference data in public databases continue to be of European ancestries, existing health disparities between rich and poor are likely to continue. In parallel, access to direct-to-consumer personal genetic testing continues to increase the public's appetite for genotyping and ancestry testing, resulting in greater number of enquiries with clinicians and public healthcare systems. This has left many medical professionals unprepared and unable to harness the wave of patient-focused healthcare and demand for data sharing—data sharing which is crucial for establishing better, more precise tools for diagnosis and treatment. However, sharing also opens the door for privacy concerns as secure access of genome data and metadata cannot always be guaranteed. Increased legal protection and institutional support are likely to keep promoting positive impact for diverse participation.

On April 11–12, 2019, we helped organize the First Personal Genomes Conference at the Wellcome Genome Campus, Hinxton, UK, to discuss issues around personal genomic data access, sharing, and interpretation (Rubin and Glusman, 2019). In this Research Topic in *Frontiers in Genetics,* we collected a selection of representative research related to the Conference's topics revolving around the themes of (a) personal genetic testing, (b) interpretation of personal genomes, (c) personal genomics citizen science, (d) return of data to genome research participants, (e) data protection, privacy and the ethics of data sharing, and (f) clinical perspective—from patients to the public.

In our Research Topic, Du and Wang describe how the direct-to-consumer market in China has been particularly buoyant in recent years, with many providers offering multiple channels for purchasing genetic testing products. They argue, however, that a regulatory vacuum exists in how to obtain valid informed consent, and protect customers' genetic data from access by third parties.

OPEN ACCESS

Edited and reviewed by:
Dov Greenbaum,
Yale University, United States

Correspondence:
Manuel Corpas
m.corpas@cpm.onl

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *29 March 2021*
Accepted: *04 May 2021*
Published: *04 June 2021*

Citation:
Corpas M, Beck S, Glusman G and Shabani M (2021) Editorial: Personal Genomes: Accessing, Sharing, and Interpretation. Front. Genet. 12:687584. doi: 10.3389/fgene.2021.687584

In India, Pemmasani et al. stress the need for existing data reference repositories to expand their variant data information, offering non-European personal genomes equitable access to resources and tools for their interpretation. Folkersen et al. present *Impute.me*, an open source tool for analyzing direct-to-consumer genetic data. With tools such as *Impute.me* anyone in the world can calculate and interpret polygenic risk scores free of charge. Guerra-Assunção et al. present another tool, *GenomeChronicler*, that uses open access Personal Genome Project (PGP-UK Consortium, 2018) data to generate reports (for research use only) that include information relating to possibly beneficial and harmful variants as well as ancestry. Mehandziska et al. show an analytical pipeline to effectively report variants of unknown significance, which to date remain among the most challenging types of variants to interpret. Corpas et al. illustrate how existing tools and resources can be leveraged for whole genome analysis when combined. Their extensive battery of analyses for a single family provides a use case for clinicians on how to develop healthcare plans for the individual, based on genetic and other healthcare data.

Access to raw personal genomic data in clinical settings is becoming commonplace in many European nations. Narayanasamy et al. explored personal genome access policies and practices from a pool of European sequencing institutions engaged in generating massive amounts of sequencing data. They report a generalized lack of clear policies and processes for raw genomic data retention and access among large sequencing facilities. Wallace et al. argue that even when raw data are available, enabling genomic and biomedical data to be accessed and shared for secondary research purposes is not always straightforward for existing "legacy" datasets. A filter used by researchers could help determine the extent to which a given dataset can be shared. Ahmed and Shabani suggest that data sharing promoted by DNA marketplaces raises concerns about consent and privacy, and may have implications for public-funded research that does not offer incentives to share. Yet, for parents of children suffering from rare and common diseases, there are powerful incentives to share whole genome sequencing data. Beauvais et al. provide recommendations for healthcare professionals in the clinical and research contexts when faced with sharing genomic data on parental request for a child's raw genomic data. Wöhlke et al. contribute to this debate by suggesting that lay people's sharing perceptions are important because they affect both their interest in undertaking genetic testing as well as their interpretations of test results. Their survey on personal sharing preferences in both Germany and Italy shows a relatively high willingness among participants to share information with their social circle, but an overall strong reluctance to share data with certain institutions (such as employers, health insurance) due to fear of genetic discrimination. In Korea, however, Kim et al. found that public interest toward establishing a citizen participation cohort is very high.

In conclusion, we observe that although general access to personal genome data is becoming more widespread, the benefits of such advances are being deployed unevenly. Tools are being implemented that help facilitate the interpretation of personal genomic data and their increased, more secure sharing. We see these advances being undertaken both by academic and industry sectors. But a number of ethical challenges persist, including how to return data to participants in different regions of the world, or how to access direct-to-consumer services and raw data for personal genome analysis, which still remain biased depending on the individual's local jurisdiction. It is our wish to raise awareness about these hurdles and to bring all stakeholders involved into fruitful discussions to promote greater access to the benefits of personal genomics for all.

AUTHOR CONTRIBUTIONS

MC wrote the paper with contributions from all authors.

ACKNOWLEDGMENTS

We are grateful to Wellcome Connecting Science Courses and Conferences colleagues Treasa Creavin and Lucy Criddle for their support with the organization of the Personal Genomes Conference, the source of many articles in this Research Topic.

REFERENCES

PGP-UK Consortium (2018). Personal genome project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. *BMC Med. Genomics* 11:108. doi: 10.1186/s12920-018-0423-1

Rubin, I. R., and Glusman, G. (2019). Opportunities and challenges in interpreting and sharing personal genomes. *Genes* 10. doi: 10.3390/genes10090643

Conflict of Interest: At the time of writing, MC is associated with Cambridge Precision Medicine Limited.

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.

Copyright © 2021 Corpas, Beck, Glusman and Shabani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

PERSPECTIVE
published: 05 November 2019
doi: 10.3389/fgene.2019.01107

DNA Data Marketplace: An Analysis of the Ethical Concerns Regarding the Participation of the Individuals

Eman Ahmed[1,2] *and Mahsa Shabani*[1,3]*

[1] *Center for Biomedical Ethics and Law, Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium,* [2] *Clinical Pharmacology Department, Faculty of Medicine, Suez Canal University, Ismailia, Egypt,* [3] *Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium*

Personal genomic data and the related health data are valuable resources for both public-funded research, and for-profit entities in development of new drugs, therapies, and diagnostic tests. In order to access to large datasets, pharmaceutical and biotech companies have developed partnerships with public and private entities such as direct-to-consumer genetic testing companies to buy genomic and health related databases collected from research participants and customers. Although individuals mainly support data sharing for research purposes, the for-profit nature of such data sharing raises some questions regarding the rights of the data subjects and fairness in sharing benefits. In response, a new generation of sequencing and data sharing startups such as Nebula Genomics, LunaDNA, and EncrypGen are emerging which aim for leaving the data control in the hands of each individual customer. In particular, such so-called "DNA data marketplaces" allow individuals to receive various types of monetary incentives to sequence their genome and share it with interested commercial parties. This paper aims to provide an exploratory and critical review of the ethical challenges related to establishing such marketplaces for genomic and health data sharing. In the view of the growing number of startups developing such marketplaces, a thorough analysis of the relevant ethical concerns is timely and needed.

Keywords: genomics, data sharing, incentives, research ethics, privacy

OPEN ACCESS

Edited by:
Go Yoshizawa,
OsloMet – Oslo Metropolitan University, Norway

Reviewed by:
Gabriele Werner-Felmayer,
Innsbruck Medical University, Austria
Chih-hsing Ho,
Academia Sinica, Taiwan

Correspondence:
Mahsa Shabani,
mahsa.shabani@kuleuven.be

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *30 July 2019*
Accepted: *15 October 2019*
Published: *05 November 2019*

Citation:
Ahmed E and Shabani M (2019) DNA Data Marketplace: An Analysis of the Ethical Concerns Regarding the Participation of the Individuals. Front. Genet. 10:1107. doi: 10.3389/fgene.2019.01107

INTRODUCTION

Personal genomic data and the related health data are valuable resources for both public-funded research, and for-profit entities in development of new drugs, therapies, and diagnostic tests. In order to access to large datasets, pharmaceutical, and biotech companies have developed partnerships with public and private entities such as direct-to-consumer (DTC) genetic testing companies to buy genomic and health related databases collected from research participants and customers.

Most of the customers of DTC companies such as 23andMe opt-in to participate in research activities of the service providers and the downstream data sharing by the companies for research purposes (Hirschler, 2018). The existing studies with customers have revealed that the underlying reasons are mainly out of altruistic motivation to participate in research and help advancement of science (Trinidad et al., 2010; Goodman et al., 2017). However, the for-profit nature of sharing customers' data by DTC companies has been perceived objectionable by some customers (Skloot, 2015). Notably, by giving consent to research, customers should accept that they acquire no rights to research, products, or profits that are made and may link to their DNA (Ducharme, 2018). This is viewed as unfair where a clear asymmetry in sharing benefits and interests is witnessed.

Moreover, the active participation of the individuals in managing sharing and access to their own genomic and health data in the framework of the current data sharing models is not fully supported. The importance of this matter is recently pronounced by the European Data Protection Supervisor in their statement: "In principle, individuals should be able to decide whether and with whom to share their personal information, for what purposes, for how long, and to keep track of them and decide to take them back when so wished" (European Data Protection Supervisor, 2016).

In response, a new generation of startups are emerging which propose, among others, to leave data control in the hands of each individual customer (Rosenbaum, 2018). These so-called "DNA data marketplaces" propose that people can share their data with companies that are interested to have access to their data for various research leading to product development (Harris, 2018) and receive monetary compensation or incentives (Jones, 2018). Although offering direct incentives to individuals to engage them in genomic data sharing may seem beneficial, this has seen as a sensitive issue drawing a lot of attention in the area of research ethics.

In an effort to address the associated concerns with DNA data marketplaces, this paper provides an exploratory and critical review of the associated ethical challenges related to participation of the individuals through analysis of different arguments discussed in academic papers and gray literature.

DNA DATA MARKETPLACE: THREE EXAMPLES

In order to illustrate our discussion, we reviewed the information provided in the websites of three startups namely Nebula Genomics, LunaDNA, and EncrypGen, which enable individuals to share their genomic data and related health information and receive various monetary incentives. We also consulted the information published in other websites in relation to the visions, policies, and strategies of these startups.

NEBULA GENOMICS

Nebula Genomics is a startup established by George Church, plans to "upend the usual way genomic information is owned," claiming that the current system applied by DTC companies is "very paternalistic" (Harris, 2018). Nebula Genomics is aiming for establishing a "Nebula marketplace," where those consenting to share their genetic information can earn the cryptocurrency called "Nebula tokens" (Buhr, 2018). In Nebula marketplace, individuals are meant to acquire and store their own genomic sequencing directly from Nebula Genomics (in partnership with Veritas) instead of obtaining the service from a personal genomics company. The Nebula's business model anticipates that companies and research organizations would be willing to pay for the cost of sequencing in exchange to get access to key medical information of the individuals involved. To this end, a blockchain platform is designed to enable customers to choose how and with whom they want their data to be shared, and to be compensated for it (Morris, 2018).

Moreover, Nebula aims for assisting pharmaceutical companies in recruiting research participants with conditions that are interesting for their current studies, by launching an anonymized search for such patients. Once contacted by the companies, the patients can decide if they will grant access to the companies to their genomic and other medical data (Harris, 2018).

ENCRYPGEN

EncrypGen is a startup aiming to "bring together genomic data sellers and buyers in one platform" (Wilson, 2019) and "looks forward to solving the problem of retaining customers' DNA data by DTC companies to be resold to research and development companies" (Matthews, 2018). EncrypGen "Gene-Chain" DNA Data Marketplace connects individual DNA data owners with data buyers and providers of other health related services. The Gene-Chain's aim is to empower users to store and monetize their genetic data by sharing it with third parties looking to obtain genetic data such as research scientists and pharmaceutical companies (Home–EncrypGen | The DNA Data Marketplace–EncrypGen., 2018).

According to the EncrypGen's website, the individuals are invited to contribute data: "If you have had your DNA tested you may upload your raw DNA data file and create a Gene-Chain profile now. EncrypGen de-identifies the raw DNA data file by stripping it away from name, email, and other sensitive information. DNA data buyers search Gene-Chain profiles suitable for their projects and purchase de-identified genomic data with DNA tokens" (Buy DNA Tokens–EncrypGen., 2018). In addition, EncrypGen has announced the plans for developing partnerships "with testing companies, analytics software developers, and various parties, like employee health benefits services," in an attempt to drive more users to the platform and monetize data (Levy, 2018).

LUNADNA

LunaDNA is a community-owned platform that is created by the Public Benefit Corporation, LunaPBC. LunaDNA offers company shares to individuals for contributing their DNA data as well as uploading their medical reports and lifestyle health activities. Those shares entitle members to a share in the profits from medical research and development. Users are supposed to get different portions of shares depending on the data they provide. For example, if a user donates DNA-targeted genes they will receive 10 shares, but if they submit their whole DNA genome, they will receive 300 shares (Lovett, 2018).

LunaDNA platform is powered by blockchain technology and provides aggregated data to researchers with the consent of the involved individuals (Lovett, 2018). In addition, LunaDNA has announced plans for collaboration with pharmaceutical companies in the future.

ETHICAL CONCERNS

Participation of the individuals in for-profit data contribution startups raises a number of ethical concerns for the rights and interests of the individuals and society in general. While some of these concerns are related to the impact of incentives strategies that such startups utilize on consent and participation in research, the other concerns are related to potential privacy concerns that may arise from use of emerging technologies such as Blockchain.

Consent-Related Concerns

In the context of DNA data marketplace, the impact of monetary incentives on validity of consent should be thoroughly investigated. We will discuss the consent-related issues here under two major concerns of undue influence and withdrawal of consent.

Undue Influence: Informed consent must be obtained from participants under circumstances that minimize the possibility of coercion or undue influence. It is important to evaluate whether or under what research circumstances financial incentives might affect a subject's judgment, and to what degree the payments induce people to participate while having deep objections (Grady, 2005). For instance, according to the official Institutional Review Board guidebook published by the US Office for Human Research Protections, "an offer is troublesome if it is so attractive [that it] may blind prospective subjects to the risks or impair their ability to exercise proper judgment" (U. S. Department of Health and Human Services, 1993).

The question here is under what circumstances offering financial incentives in exchange for individuals personal and health-related data may threat the validity of the consent and compromise the participant's ability to respond reasonably, resulting in undue induction the participation. In particular, it is crucial to investigate how both patients and healthy participants, with various socio-economic backgrounds respond to the financial incentives in personal data sharing. In traditional research settings, it is expected that the research ethics committees assess the risks of undue influence that may arise from use of monetary and other incentives to recruit research participants. However, in the context of DNA data marketplaces it is not clear if such ethics oversight is present to assess the ethical underpinnings of offering financial incentives in exchange for individuals' genomic and health-related data.

Other consent-related concerns are based on the nature of genetic data. Given the commonly shared genetic information among relatives, the involvement of the family members in the process of personal genomic data sharing and consent is a matter of discussion. Should all family members approve sharing and selling of common genetic and health information? And should they all benefit from the shares of the same individual account?

Moreover, provision of monetary incentives may have broader impact on biomedical research and data sharing, by undermining altruistic participation in research. One can argue that public-funded research that does not offer monetary incentives can be negatively impacted as a result of recruitment strategies of DNA data marketplaces.

Consent Withdrawal: Research participants should be aware that they have the right to freely withdraw their consent at any time during the research (Edwards, 2005), and voluntary terminate their participation in research (Gabriel and Mercado, 2011), without necessarily providing reasons. Notably, offering financial incentives to individuals for sharing their genomic data could be a barrier to consent withdrawal. In particular, the questions arise about whether individuals can withdraw their consent after receiving various types of financial incentive, such as tokens, shares, or free sequencing (Roberts et al., 2017). The procedure of withdrawal could be much more complex when individuals have already allowed access to their data in return for free sequencing of their genome by interested companies.

For instance, the LunaDNA consent policy informs patients that: "Your continued consent to LunaDNA' s use of your Shared Data is required for your continued ownership of any shares in LunaDNA issued to you in connection with the contribution of that Shared Data. If you elect to purge Shared Data for which you were issued ownership shares in LunaDNA, LunaDNA will redeem (i.e. cancel) those shares, and may also elect to cancel certain other shares that may have been issued to you. […] If you revoke your consent or delete your account, LunaDNA will redeem all shares issued to you." (LunaDNA, 2018). The other two startups however have not provided any information on this matter on their website. It is highly recommended that these emergent startups establish clear policies regarding consent withdrawal and communicate that to the participants.

Blockchain-Based Platforms and Privacy Concerns

Sharing personal genomic data raises considerable privacy and security concerns, due to unique nature of genomic data that contains identifiers which makes the complete de-identification of the data hard if not impossible (Wang et al., 2017). In addition, genomic data can reveal a wide range of sensitive health and non-health related data about the individuals and their family members (Genomeweb, 2018). For example, in a study analyzing Y-chromosome haplotypes together with combining data from genealogical registries, researchers were able to predict the surnames of a number of anonymized participants in the dataset (Gitschier, 2009).

As it is reported above, some of the startups aim for implementing blockchain technology as an approach to better protect genomic and health data, while allowing participatory control on access to the data. Blockchain is an emerging technology of a decentralized, digitized database medium and a public ledger of all transactions in the network (Ozercan et al., 2018). The key feature of a blockchain is the distributed database where the database is present in many copies across several computer systems creating a peer-to-peer network indicating that there is no longer a centralized body controlling access to data (Han et al., 2014; Duan et al., 2016). Arguably, blockchain-based platforms can help to solve the governance problems in sharing genomic data by using technical solutions. These platforms promise their customers to provide distributed data stewardship and control together with provision of effective ways for strengthening data access and ownership agreements (Shabani, 2019). In terms of the security of the networks, although blockchain use is expected to improve data encryption (Weintraub, 2018), no technology is infallible and concerns about possible hacking and breaching of the blockchain system have been noticed by the experts and the platform developers (Erickson, 2019).

Nebula Genomics privacy policy includes that they take a number of organizational, technical, and physical measures to protect the personal information they collect, both during transmission and once received. However they note that, "no security safeguards are 100% secure and we cannot guarantee the security of your information"(Privacy Policy, 2018). Moreover, the questions remain about the compatibility of using such technologies with applicable data protection regulations in different jurisdictions. (Price, 2018).

Finally, the possibility of access by third parties such as for law enforcement purposes should be investigated (Weintraub, 2018). The Nebula Genomics privacy policy includes the possibility of providing such access when required by law or believed to be necessary or appropriate to comply with applicable laws and lawful requests and legal process (Privacy Policy, 2018). In principle, this could be seen at odds with the rationale behind blockchain technology, which restricts access to data for those who are not part of the network.

Education and Awareness of the Potential Risks

Individuals should be encouraged to carefully weigh the benefits and risks of getting engaged in a DNA data marketplace. Moreover, raising awareness regarding the implications and possible consequences of personal genomic data sharing for the individuals and their family members is essential (Shabani and Borry, 2015). Currently, the potential concerns regarding genomic data sharing in the conventional research settings are being investigated (Middleton et al., 2018). However, the similar studies and educational materials in the context of data sharing in DNA data marketplace are missing.

Previously, in the context of Personal Genome Project (PGP), following educational videos have been required for those who agreed to share their genome publicly. In addition, the requirements such as higher level of education has been expected from volunteers of PGP (Reuter et al., 2018). Although this can be seen as one way to mitigate the concerns regarding awareness about the associated risks with such data sharing, but it may lead to biasing the sample of participants and work against diversity.

Moreover, the associated risks with sharing data through DNA data marketplace are not fully known yet. It is expected that some of the concerns such as those related to risks for privacy emerge only in the future and due to technological advances. The participants therefore should be aware of unknown risks.

CONCLUDING REMARKS

The emerging DNA Data marketplaces are promising to introduce a fair model of data sharing among individuals and the interested parties such as pharmaceutical and biotechnology companies. They encourage the individuals to directly take part in sharing their data and practice their ownership rights regarding their DNA information. However, our analysis showed that developing DNA data marketplace raises concerns about consent and privacy and may have externalities for public-funded research that do not offer incentives.

One of the main arguments of developing DNA data marketplace is to empower individuals to directly share their data and control who can have access to data. In essence, empowerment of the individuals by enabling them to actively be involved in management of their personal health information has recently received an increasing attention. For example, The European Data Protection Supervisor published in October 2016 an opinion on this subject and recognized the potential of Personal Information Management Systems (PIMS) as one approach for effectively implementing citizens' rights on their personal data at the practical level. PIMS "allow individuals to manage their personal data in secure, local or online storage systems and share them when and with whom they choose." (European Data Protection Supervisor, 2016).

DNA data marketplace could be seen as an example of such approach, aiming for involving individuals in managing how to share their health data and with whom. However, in order to truly empower patients and individuals, it is crucial to ensure that they are adequately informed about the limitations on controlling their data once have been shared and accessed by companies and interested parties. In addition, the companies should develop fair and transparent policies on issues related to consent withdrawal in the view of offering tokens, shares, etc. in exchange for data.

Moreover, in discussions related to DNA marketplace, the attentions should be paid to the fact that human beings are relational beings sharing a lot of genetic details with others, and in particular family members. In particular, since genetic data carry family connections, the implications of data donation and receiving financial incentives for family members should be taken into considerations. Currently, the discussions related to consent and withdrawal mainly limited to the concerns related to individual rights in such data donation, and do not sufficiently address the pertinent interests of the family members. On practical level, it is also crucial to investigate how far family members should/can be involved in the process of personal genomic data sharing, including giving informed consent. Notably, in the context of genetic data, the applicable legal frameworks for personal data protection are predominantly limited to recognizing individuals as "data subjects" and do not extend to the family members.

In addition, in promoting the notions of self-interest and individual empowerment, values such as altruism and solidarity in the society should not be undermined (Prainsack, 2018). This is particularly may appear concerning to traditional biomedical research which relies on altruistic participation of the individuals to advance research as a public good. Moreover, offering monetary incentives may be considered as commodifying human resources, which has been extensively debated to date, as it may lead to undermining individuals' dignity.

Notably, the success of data collection through such marketplaces is hinged on attracting a large number of participants; otherwise it would be hard to foresee a significant impact on the current way the medical research has been performed. It should be noted that currently some of other non-profit data sharing platforms such as DNA.Land that enables individuals to share their own genome- succeeded in collecting more than 150,000 genomes (Check Hayden, 2015). Therefore, the scalability of DNA data marketplaces may be seen as an achievable goal. Moreover, developing DNA Data Marketplaces and recruiting individuals directly may be considered as a solution to the problem of lack of diversity among study groups in biomedical sciences. The future studies are needed to survey the participants in such marketplaces and examine the level of diversity in terms of nationality, ethnicity, gender, and the like.

Finally, the use of the terms such as data ownership, buying and selling data, and data control in the context of personal genomic and health data should be thoroughly scrutinized, as such claims are surrounded by legal and practical uncertainties (Blasimme et al., 2018). One pertinent question is how the monetary value of DNA data can be estimated, and how this can be ethically and legally enforced (McNulty, 2018). EncrypGen declared that the price of access to data would be decided by the open market, while LunaDNA proposes different pricing for non-profits and corporations. In a recently published paper, LunaDNA presented a new model for research in which participants are issued US Securities and Exchange Commission (SEC)-qualified shares in whatever database holds their data. Thereby, "as shareholders, the participants would be eligible to receive commercial proceeds generated by mining their datasets, effectively transforming them from research subjects to partners."(Curtis and Hereward, 2018; Kain et al., 2019).

This calls attention to the necessity of developing adequate guidelines, policies (soft-governance tools), and regulations in order to ensure both ethical and legal underpinnings of DNA data marketplaces as well as transparency and fairness of the procedure. That said, the existing national and European regulations regarding personal data protection and consumer protection provide general framework for some aspects of data collection and processing by such data marketplaces, including in relation with consent, data portability, and transparency of data processing.

AUTHOR CONTRIBUTIONS

EA and MS both contributed to the structuring, drafting, and revising the manuscript.

ACKNOWLEDGMENT

MS is supported by a Research Foundation-Flanders (FWO) postdoctoral fellowship.

REFERENCES

Blasimme, A., Vayena, E., and Hafen, E. (2018). Democratizing health research through data cooperatives. *Philos. Technol.* 31, 473–479. doi: 10.1007/s13347-018-0320-8

Buhr, S. (2018). George Church's genetics on the blockchain startup just raised $4.3 million from Khosla, *TechCrunch*. Available at: http://social.techcrunch.com/2018/08/29/george-churchs-genetics-on-the-blockchain-startup-just-raised-4-3-million-from-khosla/ [Accessed May 4, 2019].

Buy DNA Tokens – EncrypGen. (2018). Available at: https://encrypgen.com/buy-dna-tokens/.

Check Hayden, E. (2015). Scientists hope to attract millions to "DNA.LAND." *Nat. News*. doi: 10.1038/nature.2015.18514

Curtis, C., and Hereward, J. (2018). New cryptocurrencies could let you control and sell access to your DNA data. *The Conversation*. Available at: http://theconversation.com/new-cryptocurrencies-could-let-you-control-and-sell-access-to-your-dna-data-89499 [Accessed May 13, 2019].

Duan, Z., Yan, M., Cai, Z., Wang, X., Han, M., and Li, Y. (2016). Truthful incentive mechanisms for social cost minimization in mobile crowdsourcing systems. *Sensors* 16, 481. doi: 10.3390/s16040481

Ducharme, J. (2018). A major drug company now has access to 23andMe's genetic data. should you be concerned? *Time*. Available at: http://time.com/5349896/23andme-glaxo-smith-kline/ [Accessed December 6, 2018].

Edwards, S. J. L. (2005). Research participation and the right to withdraw. *Bioethics* 19, 112–130. doi: 10.1111/j.1467-8519.2005.00429.x

Erickson, S. (2019). Wha's your DNA worth? LunaDNA will help you find out -. *Motley Fool*. Available at: https://www.fool.com/investing/2019/03/18/whats-your-dna-worth-lunadna-will-help-you-find-ou.aspx [Accessed May 4, 2019].

European Data Protection Supervisor. (2016). EDPS opinion on personal information management systems. Available at: https://edps.europa.eu/sites/edp/files/publication/16-10-20_pims_opinion_en.pdf.

Gabriel, A. P., and Mercado, C. P. (2011). Data retention after a patient withdraws consent in clinical trials. *Open Access J. Clin. Trials* 3, 15–19. doi: 10.2147/OAJCT.S13960

Genomeweb (2018). Family ties can compromise genomic data privacy, new studies suggest. *GenomeWeb*. Available at: https://www.genomeweb.com/genetic-research/family-ties-can-compromise-genomic-data-privacy-new-studies-suggest [Accessed May 6, 2019].

Gitschier, J. (2009). Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.* 84, 251–258. doi: 10.1016/j.ajhg.2009.01.018

Goodman, D., Johnson, C. O., Bowen, D., Smith, M., Wenzel, L., and Edwards, K. (2017). De-identified genomic data sharing: the research participant perspective. *J. Community Genet.* 8, 173–181. doi: 10.1007/s12687-017-0300-1

Grady, C. (2005). Payment of clinical research subjects. *J. Clin. Invest.* 115, 1681–1687. doi: 10.1172/JCI25694

Han, M., Yan, M., Li, J., Ji, S., and Li, Y. (2014). Neighborhood-based uncertainty generation in social networks. *J. Comb. Optim.* 28, 561–576. doi: 10.1007/s10878-013-9684-y

Harris, R. (2018). Startup offers to sequence your genome free of charge, then let you profit from it. *NPR.org*. Available at: https://www.npr.org/sections/health-shots/2018/11/15/667946213/startup-offers-to-sequence-your-genome-free-of-charge-then-let-you-profit-from-i [Accessed December 7, 2018].

Hirschler, B. (2018). Cashing in on DNA: race on to unlock value in genetic data. *Reuters*. Available at: https://uk.reuters.com/article/uk-health-dna-idUKKBN1KO0WW [Accessed December 6, 2018].

Home – EncrypGen | The DNA data marketplace – EncrypGen. (2018). Available at: https://encrypgen.com/.

Jones, B. (2018). Nebula Genomics will let you rent out your genetic information. *Futurism*. Available at: https://futurism.com/nebula-genomics-rent-genetic-information [Accessed December 7, 2018].

Kain, R., Kahn, S., Thompson, D., Lewis, D., Barker, D., Bustamante, C., et al. (2019). Database shares that transform research subjects into partners. *Nat. Biotechnol.* 37, 1112–1115. doi: 10.1038/s41587-019-0278-9

Levy, G. (2018). On gene-chain's beta launch: interview with encrypGen's Dr. David Koepsell. *Bitsonline*. Available at: https://bitsonline.com/genomic-data-blockchain-encrypgen/ [Accessed May 5, 2019].

Lovett, L. (2018). LunaDNA offers company shares in exchange for health data. *MobiHealthNews*. Available at: https://www.mobihealthnews.com/content/lunadna-offers-company-shares-exchange-health-data [Accessed April 16, 2019].

LunaDNA. (2018). LunaDNA consent: help center. Available at: https://support.lunadna.com/support/solutions/articles/43000076335-lunadna-consent [Accessed May 6, 2019].

Matthews, K. (2018). Can encrypgen (DNA) generate a 30x return in 2019? *Hacked Hacking Finance*. Available at: https://hacked.com/can-encrypgen-dna-generate-a-30x-return-in-2019/ [Accessed May 5, 2019].

McNulty, E. (2018). Ethics to ecotech: 5 unmissable talks at data natives 2018. *Dataconomy*. Available at: https://dataconomy.com/2018/10/data-natives-2018-best-talks/ [Accessed May 13, 2019].

Middleton, A., Niemiec, E., Prainsack, B., Bobe, J., Farley, L., Steed, C., et al (2018). "Your DNA, your say": global survey gathering attitudes toward genomics: design, delivery and methods. *Pers. Med.* 15, 311–318. doi: 10.2217/pme-2018-0032

Morris, N. (2018). Nebula launches "free" DNA sequencing blockchain platform. *Ledger Insights*. Available at: https://www.ledgerinsights.com/nebula-launches-dna-blockchain-platform/ [Accessed April 3, 2019].

Ozercan, H. I., Ileri, A. M., Ayday, E., and Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome Res.* 28, 1255–1263. doi: 10.1101/gr.207464.116

Prainsack, B. (2018). The "we" in the "me": solidarity and health care in the era of personalized medicine. *Sci. Technol. Hum. Values* 43, 21–44. doi: 10.1177/0162243917736139

Price, D. (2018). 5 Blockchain problems: security, privacy, legal, regulatory, and ethical issues - blocks decoded. Available at: https://blocksdecoded.com/blockchain-issues-security-privacy-legal-regulatory-ethical/[Accessed May 6, 2019].

Privacy Policy(2018) *Nebula Genomics*. Available at: http://nebulagenomics.zendesk.com/hc/en-us/articles/360024597131-Privacy-Policy- [Accessed May 14, 2019].

Reuter, M. S., Walker, S., Thiruvahindrapuram, B., Whitney, J., Cohn, I., Sondheimer, N., et al (2018). The personal genome project canada: findings from whole genome sequences of the inaugural 56 participants. *Can. Med. Assoc. J.* 190, E126–E136. doi: 10.1503/cmaj.171151

Roberts, J. L., Pereira, S., and McGuire, A. L. (2017). Should you profit from your genome? *Nat. Biotechnol.* 35, 18–20. doi:10.1038/nbt.3757.

Rosenbaum, E. (2018). Harvard genetics pioneer will monetize DNA with digital currency. Available at: https://www.cnbc.com/2018/02/08/harvard-genetics-pioneer-will-monetize-dna-with-digital-currency.html[Accessed December 7, 2018].

Shabani, M. (2019). Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems? *J. Am. Med. Inform. Assoc.* 26, 76–80. doi: 10.1093/jamia/ocy149

Shabani, M., and Borry, P. (2015). Challenges of web-based personal genomic data sharing. *Life Sci. Soc. Policy* 11, 1–13. doi: 10.1186/s40504-014-0022-7

Skloot, R. (2015). Opinion | your cells. their research. your permission? - the new york times. Available at: https://www.nytimes.com/2015/12/30/opinion/your-cells-their-research-your-permission.html [Accessed December 7, 2018].

Trinidad, S. B., Fullerton, S. M., Bares, J. M., Jarvik, G. P., Larson, E. B., and Burke, W. (2010). Genomic research and wide data sharing: views of prospective participants. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 12, 486–495. doi: 10.1097/GIM.0b013e3181e38f9e

U. S. Department of Health and Human Services (1993). *Institutional Review Board Guidebook*. HHS.gov. Available at: https://www.hhs.gov/ohrp/ [Accessed May 21, 2019].

Wang, S., Jiang, X., Tang, H., Wang, X., Bu, D., Carey, K., et al (2017). A community effort to protect genomic data sharing, collaboration and outsourcing. *NPJ Genomic Med.* 2, 33. doi: 10.1038/s41525-017-0036-1

Weintraub, K. (2018). Genetics start-up wants to sequence people's genomes for free. *Sci. Am.* Available at: https://www.scientificamerican.com/article/genetics-start-up-wants-to-sequence-peoples-genomes-for-free/[Accessed May 6, 2019].

Wilson, P. (2019). Encrypgen's gene-chain will soon allow for dna token trading. Available at: https://globalcoinreport.com/encrypgens-gene-chain-dna-token-trading/ [Accessed April 16, 2019].

Conflict of Interest: The 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.

Copyright © 2019 Ahmed and Shabani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 March 2020
doi: 10.3389/fgene.2020.00102

German and Italian Users of Web-Accessed Genetic Data: Attitudes on Personal Utility and Personal Sharing Preferences. Results of a Comparative Survey (n=192)

Sabine Wöhlke[1]*, Manuel Schaper[1], Serena Oliveri[2,3], Ilaria Cutica[2,3], Francesca Spinella[4], Gabriella Pravettoni[2,3], Daniela Steinberger[5] and Silke Schicktanz[1]

[1] Department of Medical Ethics and History of Medicine, University Medical Center Göttingen, Göttingen, Germany,
[2] Department of Oncology and Hematology Oncology, Faculty of Medicine and Surgery, University of Milan, Milan, Italy,
[3] Applied Research Division for Cognitive and Psychological Science, European Institute of Oncology, Milan, Italy,
[4] Laboratory GENOMA, Rome, Italy, [5] bio.logis Genetic Information Management GmbH, Frankfurt, Germany

OPEN ACCESS

Edited by:
Mahsa Shabani,
Ghent University, Belgium

Reviewed by:
Davit Chokoshvili,
CISTIM, Belgium
Danya Vears,
KU Leuven, Belgium

***Correspondence:**
Sabine Wöhlke
sabine.woehlke@medizin.uni-goettingen.de

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: 11 June 2019
Accepted: 29 January 2020
Published: 18 March 2020

Citation:
Wöhlke S, Schaper M, Oliveri S, Cutica I, Spinella F, Pravettoni G, Steinberger D and Schicktanz S (2020) German and Italian Users of Web-Accessed Genetic Data: Attitudes on Personal Utility and Personal Sharing Preferences. Results of a Comparative Survey (n=192). Front. Genet. 11:102. doi: 10.3389/fgene.2020.00102

Genetic information is increasingly provided outside of the traditional clinical setting, allowing users to access it directly *via* specialized online platforms. This development is possibly resulting in changing ethical and social challenges for users of predictive genetic tests. Little is known about the attitudes and experiences of users of web-accessed genetic information. This survey analyzes data from two European countries with regard to the utility of genetic information, the users' ways of making use of and dealing with information, and their sharing behavior. Particular focus is given to ethical and social questions regarding the motivation to share personal genetic results with others. Social factors tested for are national background, gender, and marital, parental, and educational status. This study will contribute to public discourse and offer ethical recommendations. The study will also serve to validate the developed questionnaire for use in population representative surveys.

Keywords: genomics, health information, attitudes, experiences, survey, lay people, utility

INTRODUCTION

Lay people are increasingly able to access digitized data regarding their personal health, ranging from information provided by self-tracking and fitness apps to electronic patient records (Lupton, 2014; Rexhepi et al., 2018). Within this trend, genetic information has become widely available, presenting lay people in the role of patients and consumers of health services with a variety of implications and possibilities regarding application, utility and information sharing. Research has shown that the public's interest in genetic information is high (Townsend et al., 2012), and there are different plausible reasons for that interest: Genetic tests can confirm or rule out genetic traits or a suspected genetic condition, or they can help to determine a person's chance of developing or passing on a genetic disorder, and in some cases they provide relevant information that can be used to the patient's benefit (Burke, 2014). In such cases, genetic testing has "clinical utility". Recently, scholars have discussed the possibility that genetic testing also has "personal utility", e.g., that it plays a role in shaping individual understandings of disease or personal identities of their carriers (Bunnik et al., 2015; Kohler et al., 2017a; Kohler et al., 2017b; Urban and Schweda, 2018). However, lay understandings of genetic information and its implications diverge from those of experts, and may be shaped by specific life situations, such as experience of disease, personal attitudes and beliefs, and psycho-social circumstances (Oliveri et al., 2015; Oliveri et al., 2016a; Oliveri and Pravettoni, 2018; Oliveri et al., 2018), as well as by cultural background (Raz and Schicktanz, 2016). Lay people's perceptions are important because they affect both their interest in undergoing genetic testing as well as their interpretations of test results.

Some institutions offering genetic testing provide direct access to own genetic test results *via* specialized online-platforms. This article focuses on users of such direct access to personal genetic information (i.e., lay people in regard to understanding genetic information) and their specific attitudes and behaviors regarding information sharing and the exercise of responsibility within families (e.g., decisions regarding whether or not to inform relatives about their genetic risks or regarding reproductive behavior) (Welch and Burke, 1998; Anderson and Wasson, 2015; Baars et al., 2016). This perspective is relevant because lay people often consider the decision to undergo genetic testing to be an individual choice rather than a socially embedded decision (Corpas, 2012; Schaper et al., 2018). Receiving genetic risk information can potentially cause psychological harm because some conditions are currently untreatable and being affected may lead to stigmatization and discrimination (Slaughter, 2006; Kollek and Lemke, 2008; Ross et al., 2015). Furthermore, ethical conflicts may arise when the needs of the client/patient do not accord with those of other family members or society at large, and genetic counselors are increasingly faced with conflicting obligations, e.g. when there is critical information available that applies to multiple persons with different information preferences (Muthuswamy, 2011). While a moral duty may exist to share genetic information in order to prevent harm to others, the nature of a specific condition and the predicted harm associated with it need to be considered as well (D'Agincourt-Canning, 2001; Parens and Appelbaum, 2019). However, privacy and confidentiality are central issues in genetic testing and making use of and dealing with genetic information, and while there is consensus that individuals are entitled to knowing about existing genetic information, a right not to know has become the central moral norm, especially regarding genetic risk of contracting a disease (Chadwick et al., 2014; Domaradzki, 2015; Lupton and Michael, 2017).

A German study showed that lay people perceive risk information as highly normatively charged, and often as an emotionally significant threat (Wöhlke et al., 2019). It would therefore seem to be necessary to provide lay people with a deeper understanding of risk information and of the limitations of genetic knowledge with respect to one's own health responsibility (Wöhlke and Perry, 2019). Similar results were found for lay people in Italy, who perceived genetic testing to be very helpful for disease prevention but were simultaneously afraid that a positive result, the detection of a genetic variant, could affect their life planning and leave them without the ability to act to address the risk (Oliveri et al., 2016a). Across Europe, the regulation of genetic testing is focused on the prevention of harm to the individual—therefore, public opinion should be taken into account in the creation of policy and legislation regarding the communication of genetic risk (Oliveri et al., 2016b).

Questions regarding the implications of personal access to genetic information are becoming increasingly important in the eHealth era, where health information is becoming more accessible to lay people in the role of patients and consumers as well as to various other actors in the healthcare sector. Currently, there are significant differences between countries in terms of the political will to implement eHealth, available infrastructure, and actual use of these possibilities. Here, Germany and Italy offer contrasting examples within Europe, with Germany being less advanced than Italy in eHealth implementation efforts (Poss-Doering et al., 2018; Thiel et al., 2019). Italian eHealth initiatives have mainly been in the areas of improving access to health services and availability of patients' clinical histories, innovating primary care, and redesigning the healthcare services network through Telemedicine (Domenichiello, 2015). For this study, we conducted a survey of Germans and Italians who have access to their personal genetic information in order to gain a deeper insight into the practical and ethical questions associated with accessing and sharing such information.

Sharing of health information for more efficiency in health care and research is a central notion in the eHealth paradigm. Privacy and confidentiality are therefore important issues in relation to personal data that are acknowledged by political decision makers in both Germany and Italy (Thiel et al., 2019). The present notion of sovereignty over one's own genetic information touches a number of ethical aspects related to both the self-determination and the privacy of patients. However, it is unclear how the autonomy and right to know of individuals can be reconciled with the self-determination and right not to know of their family members. The holder of genetic information has a special responsibility because of its relevance for other biologically related persons (Leefmann et al., 2017). With the introduction of the General Data Protection Regulation (GDPR), a uniform legal requirement for the handling of personal data was adopted in the European Union, aimed at guaranteeing data security and data sovereignty. However, there is great variety in how genetic testing is legally defined and regulated internationally (Borry et al., 2012; Soini, 2012; Varga et al., 2012). In both Germany and Italy, genetic testing for medical purposes is subject to legislation that requires specialized physicians and the provision of genetic counseling (Kalokairinou et al., 2018).

AIM

Given the topic outlined above, the overall aim of this study was to gather information about the personal experiences and moral and social attitudes of lay people as well as their ways of making use of and coping with genetic information and examine the similarities and differences between German and Italian users (lay people) of direct access to personal genetic information, and the way these similarities and differences are related to age, sex, and social and educational background.

METHODS

In 2018, we conducted an online survey of persons with direct access to their own genetic information, provided *via* centers for human genetics in Germany and Italy.

The survey consisted of 13 questions in three thematic blocks (see **Supplementary 1**):

a. Experience with genetic testing: questions concerning the level of understanding of own genetic test results and perceived controllability of their implications for health.
b. Personal opinion on genetic testing in general: questions mainly concerning the utility of genetic testing, who should undergo genetic testing, the right to know or not to know, and regulation.
c. Making use of and dealing with test results: questions mainly concerning preferences and reasons in sharing genetic own genetic information.

Further, the survey included a set of sociodemographic questions to contextualize the answers. The survey was initially developed in German by the Göttingen research group. Its content was developed based on the research question and tailored to the target population based on previous experience in studying lay perspectives on genetic testing with qualitative methodology. The survey was adapted and improved in close cooperation with the heads of GenomaLab and bio.logis Zentrum für Humangenetik (ZfH) to meet the practicalities of conducting the survey based on those institutions' technical infrastructure. Critical feedback from all co-authors was included at an early stage of development. The survey was successfully tested with academic staff of the German and Italian research groups' affiliation before application in the study. The survey was translated into English by the Göttingen research group, and thence from English to Italian for application by the research groups in Italy. The Italian translation was checked by translating it back into English.

In order to participate in the survey, participants had to read and acknowledge the study information telling them that by proceeding to the questionnaire and submitting it they gave consent to participate.

RECRUITMENT

We recruited participants who had undergone genetic testing and had online access to their personal genetic information. In the following sections we provide links to a sample account for each website.

German Sample

Participants were recruited *via* bio.logis (ZfH) in Frankfurt (Main). bio.logis (ZfH) is a clinical institute for pre- and postnatal genetic diagnostics and counselling which provides a web-portal designed to give patients direct access to selected categories of genetic information. Online access to genetic information is offered only for selected categories, such as pharmacogenetics, carrier status for recessively inherited diseases, and preventive targets. Non-treatable conditions or those whose diagnosis would lead to relatively invasive treatments, such as pronounced surgical or chemotherapeutical interventions, were excluded. Patients may log in to their personal account and see the current status of genetic analyses and results as well as news and updates provided by bio.logis (ZfH).[1] The User ID for access to the portal is provided directly to patients and to their doctors, who in the majority of cases were responsible for the referrals of patient's samples. For the purpose of recruitment users were contacted *via* an internal e-mail system of the bio.logis (ZfH) portal. The survey data was then collected online using the survey tool EvaSys. As an incentive, participants were given the option to enter a raffle for four Amazon vouchers of 50 Euros each. The recruitment mail started on May 9th, 2018. A reminder was sent out on June 22nd, 2018 and the survey was closed on August 31st.

Italian sample

Participants were recruited *via* GenomaLab - Molecular Genetics Laboratory in Rome. GenomaLab (MGL) offers a variety of genetic testing services, including screening tests for predisposition to breast and colon cancer, cardiovascular disease, and nutrigenetic and noninvasive prenatal testing.[2] The survey was advertised on GenomaLab's website, and clients who had received their genetic results in the previous two weeks were invited to participate. The link to the questionnaire was sent to other clients two weeks after they had received their genetic results. Data were collected using Survey Monkey, an open source online survey application which enables users to develop and publish surveys and register responses (www.surveymonkey.com). Recruitment started in April 2018 and ended in October 2018.

STATISTICAL ANALYSIS

The analysis was performed using *SPSS statistics (version 25)*. Descriptive statistics were calculated on raw data to depict the socio-demographic characteristics of both German and Italian samples. Frequencies were performed on the total distribution of our sample, whereas contingency tables and Chi-Square tests were performed to make comparisons based on country of origin, gender, age range, educational level, and parental status

[1] A guest login to the genetic information services is available via https://my.pgsbox.de, username: SurveyGER-I, password: GeneticInformation2019.

[2] https://www.laboratoriogenoma.eu – a guest login to the services is available via https://www.genomagroup.com/LoginRefertazione.aspx?ln=EN, username: SurveyITA-G, password: GeneticInformation2019a! (Access 25. September 2019).

for each question. Expected values and residuals in every box were calculated. Contingency tables allowed us to verify whether a specific group (German vs. Italian participants) gave a significantly higher or lower rate of response (observed values) to certain items compared to the percentage expected and calculated according to the number of subjects recruited (expected values). The analysis focused on which groups agreed to certain positions and the comparison of national, gender and age differences.

RESULTS

A sample of 192 participants was enrolled. The response rate for Germany was 7% (n=103 of 1,517 persons contacted). Of the 1,860 Italian clients who underwent genetic testing in the period of recruitment, n = 89 completed the questionnaire, a response rate of 5%. The gender distribution reflected the overall membership distribution here.

Overall, respondents were 28% men and 69% women, with 2% not defined, and 1% not responding. 52% had previous experience with genetic testing (41% participants had no experience). The sample comprised Christians (62%), Agnostics (6%), and nonreligious people (27%). The socio-demographic characteristics of the German and Italian samples are described in **Table 1**.

Experience With Predictive Genetic Testing

When asked about their experience with genetic testing and genetic information, 89% of German participants and 87% of Italian participants answered that they understood the reports on their genetic data, while 6% of German participants and 10% of Italian participants answered that they did not understand the reports on their genetic data (**Figure 1**).

When asked if they were able to explain the results to others (e.g., family members), 77% of German participants and 79% of Italian participants answered affirmatively, while 12% of German participants and 13% of Italian participants answered negatively (**Figure 1**). No significant differences were found based on sociodemographic variables, such as gender, parental status, education, etc.

Apart from these similarities, there were significant differences among German and Italian participants: more Italian participants answered that they felt they could act in some way against a genetic predisposition (73% versus 55% of German participants), and more German participants answered that they felt they could not act in any way (24% versus 11% of Italians) (X^2(1, N = 168) = 4.676, p < 0.01) (**Figure 1**).

Attitudes Toward Predictive Genetic Testing

German participants answered more frequently that genetic testing was useful to "understand myself" (60% vs. 21% of Italian participants, with a significant difference, X^2(1, N = 192) = 29.540, p < 0.01), and that genetic information had entertainment value to them (15% compared to 0% of Italians, X^2(1, N = 192) = 14.060, p < 0.01). By contrast, Italian participants answered more frequently that genetic results are useful for other people, such as their family members (40% vs. 12% of German participants) (X^2(1, N = 192 21.119, p < 0.01) (**Figure 2**).

TABLE 1 | Sociodemographic data.

Variables	Germany		Italy	
	N	%	N	%
Number of participants	103	53.6	89	46.4
Male	41	21.4	13	6.8
Female	60	31.2	72	37.5
Not defined	2	1.0	4	1.0
Age (years)				
18–25	–	–	5	2.6
26–35	23	12	28	14.6
36–50	25	13	33	17.2
51–70	49	25.5	19	10
70+	4	2.1	–	–
Missing	2		4	
Marital status				
Single	21	11	15	7.8
Married	68	35.4	42	21.9
Life-partnership	7	3.6	24	12.5
Widowed	6	3.1	2	1.0
Missing	1		6	
Number of children				
None	39	20.3	35	18.2
One	30	15.6	21	10.9
Two	17	8.9	17	8.9
Three or more	14	7.3	2	1.0
Missing	3		14	
Level of education	64	36.3	45	24.7
Academic degree	7	6.9	3	3.8
Vocational school	13	12.7	28	35
High school	14	13.7	1	1.3
year 10year 9	2	2.0	3	3.8
No education	0	0	0	0
Missing			9	

Interestingly, only 5% of women answered that genetic information had entertainment value to them; compared to 15% of the men, with a significant difference (X^2 (1, N = 186) = 4.676, p < 0.01).

55% of participants without children agreed that genetic testing "is helping me to understand myself" compared to 35% of participants with children, with a significant difference (X^2(1, N = 176) = 7.049, p < 0.01). In particular, German participants without children were significantly more likely to state that genetic information "is helping me to understand myself" (79%) than Italian participants with (12%) and without (29%) children (X^2(3, N = 176) = 41.344, p < 0.01). A similar result emerged for the question whether results have entertainment value, with 23% of German participants without children opting for this answer compared to 0% of Italian participants regardless of their parental status (X^2(1, N = 176) = 17.812, p < 0.01). Italian participants with children considered genetic test results as "important for others (e.g., family, kids)" more frequently (54%) than German participants with (12%) and without (10%) children (X^2(1, N = 176) = 30.335, p < 0.01) (**Figure 3**).

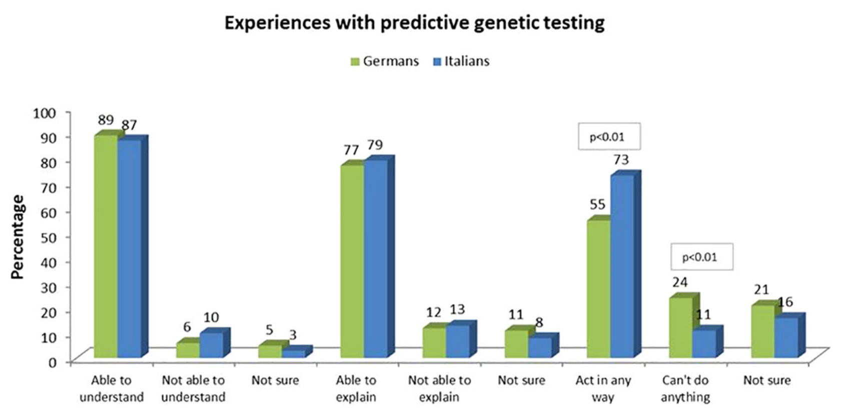


FIGURE 1 | Experiences with predictive genetic testing.

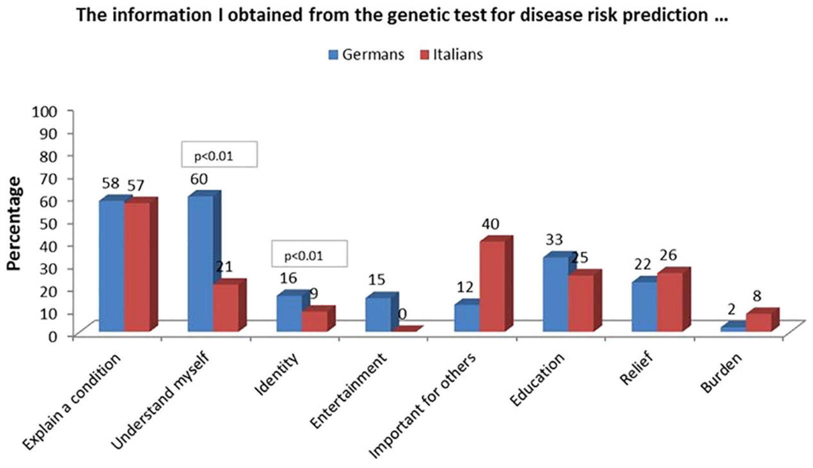


FIGURE 2 | Attitudes regarding utility towards predictive genetic testing.

Across the whole sample, more participants (47%) answered that everybody should undergo genetic testing for disease risk prediction to get information about personal disease risks, against 33% of participants who were against this option, and 20% who were unsure.

German participants responded more often that patients/clients have a right not to know about disease predisposition regardless of the circumstances (84%), compared to Italian participants (38%), who answered more frequently that such a right exists “in no case” and “do not know”, (X^2(3, N = 187) = 53.186, p < 0.01). Women tended to answer more frequently that patient/clients have a right not to know about disease predisposition (16%) than men (0%), this difference was significant (X^2(3, N = 184) = 11.439, p < 0.01) (**Figure 4**).

Italian participants responding “in no case” were all women and were mostly aged between 26–35 (33%) or 36–50 (33.3%). They were predominantly married (57%) and had an academic level of education (57%). 52% had children, whereas 38% did not. 76% already had previous experience with genetic testing. They wanted to share genetic results mainly with their partner (81%), parents (67%), and children (57%), and they actually shared results with the partner (81%) and parents (76%) at roughly the same frequency as they wanted, but not with children (33%).

Interestingly, only 23% of this group of women answered that the main reasons for sharing genetic results with family members would be “the right to share”. Other responses included: 19% “have trust in others”, 10% “share the burden”, 14% “receive comfort”, 5% “feel responsible for their life”, and 38% “It is important for reproductive planning”. Most answered “They have a right to know” (47%).

German participants answered more often that for them genetic information means certainty (59% vs. 30% Italians, (X^2 (1, N = 192) = 16.047, p < 0.01), and claimed that genetic testing includes preventive possibilities (93% vs. 83% Italians, X^2(1, N = 192) = 4.761, p < 0.05). Significant differences were also evident regarding the perceived possibility of life planning with a view to one’s own professional life (43% German participants vs. 17% Italian participants, X^2(1, N = 192) = 15.005, p < 0.01), the

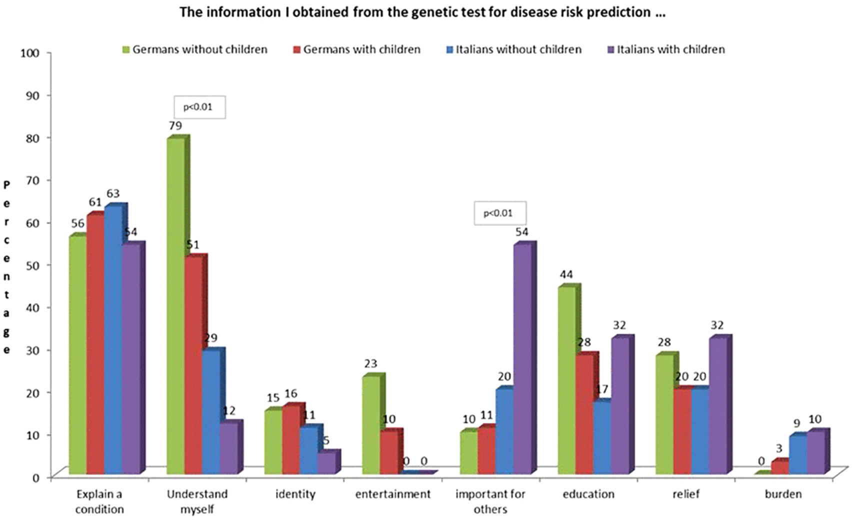


FIGURE 3 | Attitudes regarding utility towards predictive genetic testing selected in with and without children.

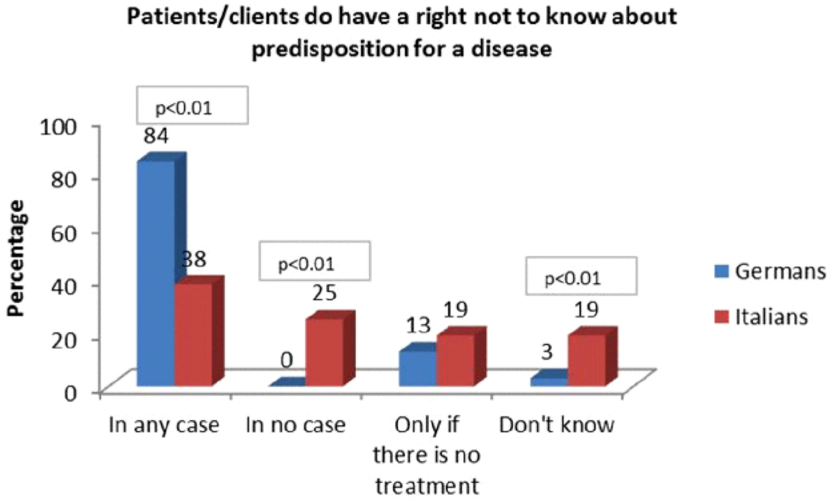


FIGURE 4 | Patients/Clients do have a right not to know about predisposition for a disease.

possibility of life planning with a view to family (63% German participants vs. 43% Italian participants, $X^2(1, N = 192) = 7.998$, $p < 0.01$). German participants were also more likely to state that genetic testing involves the risk of discrimination in health insurance (32% German participants vs. 5% Italian participants, $X^2(1, N = 192) = 23.284$, $p < 0.01$) (**Figure 5**).

Women declared more often than men that genetic testing for disease risk prediction means preventive possibilities (95% of women vs. 80% of men) ($X^2(1, N = 186) = 9.953$, $p < 0.01$). More participants with a vocational school education (60%) or academic degree (39%) answered that genetic testing means a possibility of life planning with a view to one's professional life compared to the other groups, and particularly participants with high school education (15%) ($X^2(4, N = 182) = 14,364$, $p < 0.01$).

Specifically, German participants without children answered "certainty" significantly more often (74%) compared to Italian participants without (37%) and with children (24%) ($X^2(3, N = 176) = 21.846$, $p < 0.01$). German participants with children more often stated that genetic testing allowed the possibility of life planning with a view to profession (46%) than Italian participants without children (9%) ($X^2(3, N = 176) = 16.680$, $p < 0.01$). Italian participants without children answered less frequently that genetic testing means the possibility of life planning with a view to family compared to the other groups (31% ($X^2(3, N = 176) = 12.573$, $p < 0.01$).

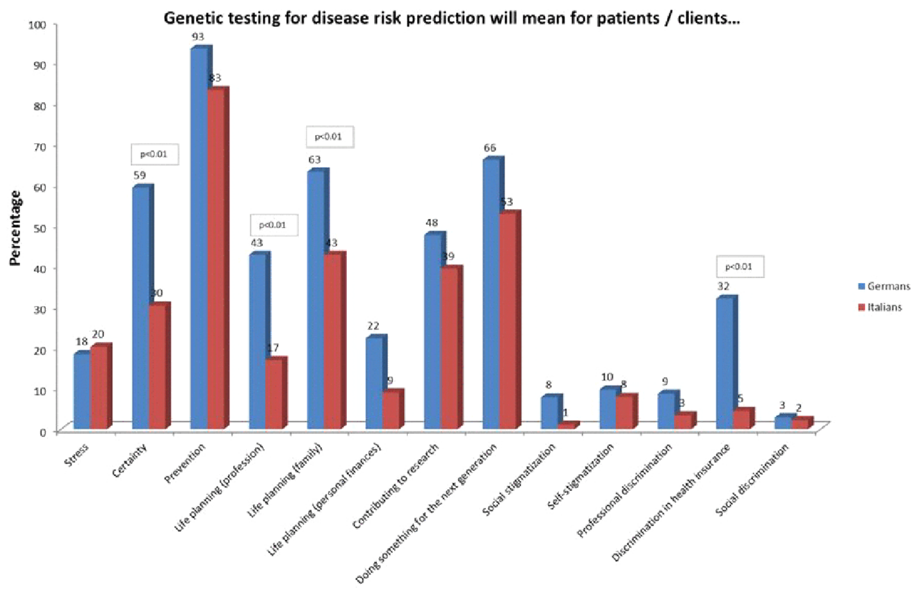


FIGURE 5 | Attitudes towards opportunities and risks towards predictive genetic testing.

German participants without children were significantly more likely to state that there is utility in the possibility of life planning with a view to personal finances than Italians without children (45% vs. 3%, $X^2(3, N = 176) = 13.316$, $p < 0.01$). German participants without (36%) and with children (30%) more often saw a risk of discrimination in health insurance than Italian participants with (2%) and without children (9%) ($X^2(3, N = 176) = 20.005$, $p < 0.01$) (**Figure 6**).

Italian participants stated more often (74%) than German participants (47%) that predictive genetic testing is generally useful, and significantly less often that predictive genetic testing is useful in the case that an effective treatment is available (24% compared to 38% of Germans, $X^2(3, N = 188) = 17.557$, $p < 0.05$). Italian participants without children answered significantly more often that genetic testing is generally useful (80%) compared to German participants with children (42%) ($X^2(4, N = 174) = 25.742$, $p < 0.01$). German participants with children answered more frequently that genetic testing is useful in case an effective treatment is available (47%) compared to the other groups ($X^2(4, N = 174) = 25.742$, $p < 0.01$).

When asked about regulations needed to offer genetic testing, German participants answered more frequently that genetic testing needs a standardization of test methods and limits (i.e. reliable and comparable test procedures with comparable properties) (69% vs. 51% Italian participants, $X^2(1, N = 192) = 6.737$), medical guidelines (85% vs. 62% Italians, $X^2(1, N = 192) = 14.037$, $p < 0.01$), directives for data protection (72% vs. 35%, $X^2(1, N = 192) = 26.396$, $p < 0.01$), and the possibility of effective treatment (49% vs. 29% Italian participants, $X^2(1, N = 192) = 7.460$, $p < 0.01$).

Men answered more frequently that a standardization of test methods and limits (76% vs. 57% women, $X^2(1, N = 186) = 5.961$) and directives for data protection are important (70% vs. 50% women, $X^2(1, N = 186) = 6.451$, $p < 0.01$). A standardization of test methods and limits was also important to participants with an academic degree (74% vs. 29% compared to the other groups. High school 42% vs. 59%; ten years of education 47% vs. 53%; < 9 years of education 40% vs. 60% $X^2(4, N = 182) = 17.132$, $p < 0.01$).

Dealing With Genetic Test Results

Italian participants preferred to involve parents more than Germans (64% vs. 42% of Germans, $X^2(1, N = 192) = 9.511$, $p < 0.01$). Italian participants without children stated that they would share their test results with their parents (71%) more than German participants without children (54%) and people with children (63% Italians and 34% Germans) in general ($X^2(3, N = 176) = 15.009$, $p < 0.01$).

More women than men stated they would share results with their parents (60% of the women vs. 37% of the men, $X^2(1, N = 186) = 8.010$, $p < 0.01$). Participants with an academic degree answered (97%) they would share results with the partner more frequently than the other groups (high school diploma 83% yes; 10 years of education 67% yes, < 9 years of education 100% yes) $X^2(4, N = 182) = 20.407$, $p < 0.01$).

Married participants wanted to share results with the partner (98%) more than any other group ($X^2(3, N = 178) = 24.694$, $p < 0.01$). Significantly more participants in a life-partnership stated the intention to share results with their parents (75%) than married (44%) and single participants (64%), $X^2(3, N = 178) = 11.110$, $p < 0.01$).

Italian participants shared results with parents more frequently than German participants (65% Italian participants vs. 27% German participants, $X^2(1, N = 192) = 27.857$, $p < 0.01$) (**Figure 7**). German and Italian participants shared results with their children equally (34% German participants and 42% Italian

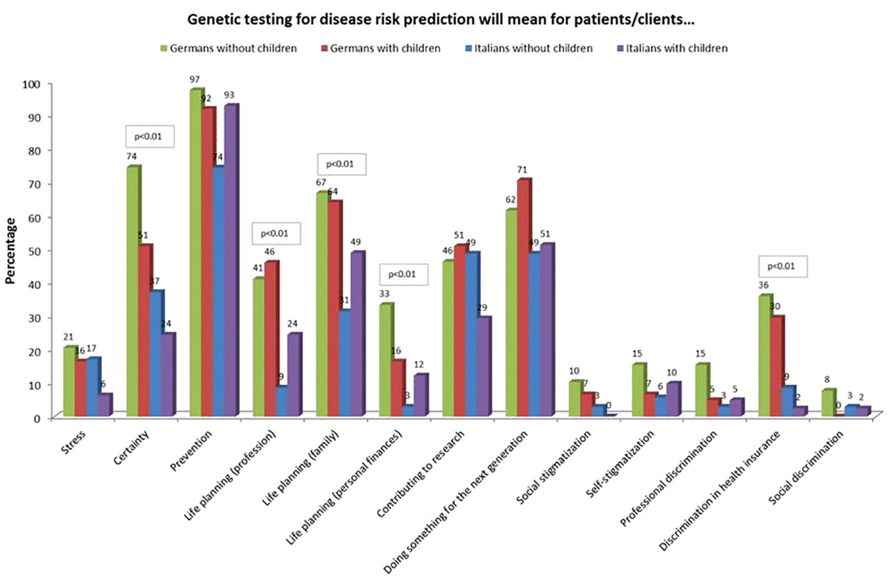


FIGURE 6 | Attitudes towards opportunities and risks towards predictive genetic testing selected in with and without children.

participants). Women tended to share their results with their parents more than men (55% of the women vs. 24% of the men, $X^2(1, N = 186) = 15.035$, $p < 0.01$). Overall, there was a relatively high willingness to share results within the social circle (89% with the partner, 52% with parents, 52% with their children, 16% with friends, of the whole sample of participants) while most of the participants reported reluctance to share results with employers (1%) and other institutions like health insurance (6%) (**Figure 7**).

Considering the whole sample, 80% of participants had actually shared results with the partner, 45% with the parents, 22% with their children, 26% with friends, whereas only 3 participants each actually shared results with health insurance or with the employer. Only 1 participant reported to have shared information with authorities (unspecified).

Differences were evident regarding the reasons for sharing genetic information: German participants chose the answer option "my right to test means that I can share the information" more often (54%) than Italian participants (36%) ($X^2(1, N = 192) = 6.521$, $p < 0.05$). Similarly, German participants answered more often "I have trust in others" (44% vs. 19% of Italian participants, $X^2(1, N = 192) = 13.202$, $p < 0.01$), and "I feel responsible for their [family members] life" (34% vs. 19% for Italians, $X^2(1, N = 192) = 5.353$, $p < 0.05$).

German participants without children answered "…means that I can share the information" (69%, $X^2(1,N = 176) = 11.851$, $p < 0.01$), and "I have trust in others" (51%, $X^2(1, N = 176) = 11.851$) more frequently than Italian participants without children, and both German and Italian participants with children.

Italian participants wanted to share genetic results mainly with their partner (81%), parents (67%) and children (57%), and they actually shared results with the partner (81%), parents (76%) but not with children the same frequency they wanted (33.3%).

Interestingly, only 23% of Italian participants answered that the main reasons for sharing genetic results with family members would be "the right to share", 19% "have trust in others", 10% "share the burden", 14% "receive comfort", 5% "feel responsible for their life", 38% "It is important for reproductive planning". Most answered "They have a right to know" (47%).

Men report more trust in others than women (50% men vs. 26% women, $X^2(1, N = 186) = 10.219$, $p < 0.01$), and feel more responsible for their family members' lives (40% men vs. 22% women, $X^2(1, N = 186) = 6.785$, $p < 0.01$). Participants with an academic degree answered more frequently "I have trust in others", particularly compared to participants with high school level (39% academic degree vs. 10% high school, $X^2(4, N = 182) = 15.465$, $p < 0.01$).

Men answered more frequently that "…persons or institutions can control me with the information" (9% men vs. 0% women, $X^2 (1,N = 186) = 12.560$, $p < 0.01$).

DISCUSSION

Our results provide empirical insights to the notions of "personal utility" and "data sharing", which are often used as umbrella terms in discussions of the usability of genomic information. The results show a relatively high willingness among participants to share information with their social circle but an overall strong reluctance to share data with official institutions (employers, health insurance) due to fear of genetic discrimination. Several studies showed that, while there are limits in regard to people's willingness to share genetic information, there is a significant interest in sharing it for research purposes (e.g., in health data

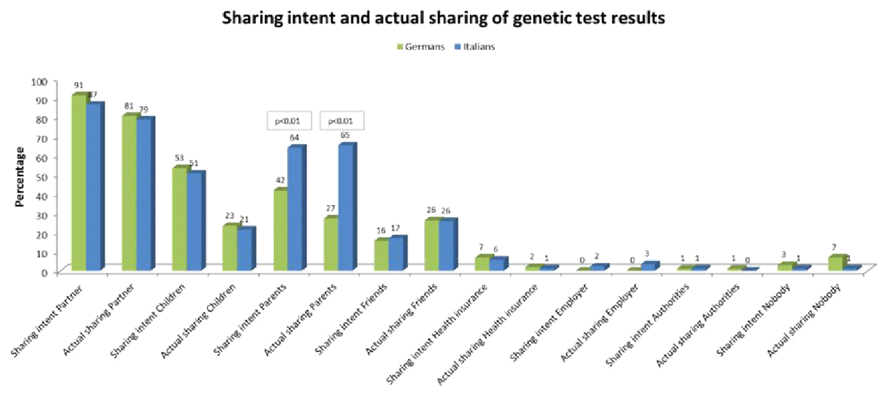


FIGURE 7 | Dealing with genetic test results.

cooperatives) medical progress (Wicks et al., 2010; Haga and O'Daniel, 2011; Hafen et al., 2014; Aitken et al., 2016; Thorogood et al., 2018). This can be interpreted as openness to the shared exchange of genetic information when societal benefits are expected. The perspective may be different when it comes to sharing information with other people and institutions, such as insurance companies or employers, that have an interest other than research.

Our data also supports the idea of Wöhlke et al. (2019) that using genetic information can lead to stronger beliefs in self-efficacy. The fact that patients are willing to share their data within social groups shows that social objectives play an important role, e.g., the comparison of health data with other patients, or the exchange of information on dealing with the disease and its treatment.

The danger of stigmatization and discrimination based on genetic information is often cited as ethically problematic (DiMillio et al., 2015; German Ethics Council, 2018). However, only the German participants saw a significant danger of discrimination in health insurance, and our study showed overall little indication of fear of such negative consequences. Genetic knowledge is therefore less often perceived as a risk of individualization of health risks and loss of social solidarity, as feared by some experts (Lemke et al., 2010; Wöhlke et al., 2015). Instead, there is an apparent optimism regarding the possibilities of sharing genetic information to everyone's benefit—a notion that also drives the development of new genetic data sharing cooperatives (see Prainsack, 2017).

It is also interesting that many participants were unsure whether it was advisable for everyone to undergo a genetic test. A clear cultural difference is evident between German respondents, who support the right not to know and find aspects of personal utility of genetic information very important, and Italian respondents, who saw the value of genetic information more in terms of one's own and family prevention, i.e. in its potential to aid in exercising genetic responsibility (Leefmann et al., 2017). Comparing the two countries, it becomes clear that responsibility for the family was more important among the Italian respondents and that moral values are strongly influenced by this. In our view, these findings indicate a plurality of lay moralities regarding duties and rights related to genetic testing. They are in line with previous studies of affected people, which found national differences regarding the moral duty to undergo genetic testing (Raz and Schicktanz, 2009), or moral conflicts regarding whether or not one should know, and tell, in the context of Huntington's Disease (Konrad, 2005). Our findings suggest that a possible moral obligation to share genetic information does not necessarily depend on specific conditions or predicted negative outcomes (D'Agincourt-Canning, 2001). Rather, moral obligation is closely related to family responsibility (Leefmann et al., 2017). As our results show, the Italian respondents associated a significantly higher level of family responsibility with genetic information. In contrast, German users appeared to place much more importance on individual interest and benefit.

The vast majority of our participants claimed that they understood their genetic reports. However, there are differences in the assessment of the benefits of such data: German participants were much more skeptical than Italian participants that they could counter-balance a genetic predisposition with preventive measures. In line with other empirical studies (Paton et al., 2012; Sommer et al., 2013; Lupton and Michael, 2017), we found that German participants use genetic information to learn more about themselves. In contrast, the motivation of Italian respondents in dealing with genetic information is more focused on the benefit to others, such as the family. This could be connected to the fact that in Germany there is a tendency to discuss individual genetic testing and genetic carrier screening separately (German Ethics Council, 2013). Similarly, it is striking that 1 in 4 Italians disagreed that one has a right not to know about predisposition to a disease—this right is rarely contested by experts and also exists as a legal right in both countries (German Genetic Diagnostics Act §9, (2009) Oviedo Convention 1997, Ar. 10, co 2). This could also be explained by cultural differences regarding the value of family and responsibility for others, which appears to be more significant in Italy than Germany (Rodotà, 2006). In Italy the right not to know is regulated by article 10, co. 2, of the Italian Oviedo Convention: "Everyone has the right to know all the information collected on their own health. However, the will of the person not to be informed must be respected". Nevertheless, despite the current regulatory framework, the very nature of genetic information limits individual choice in this field, since various private law regulations affecting the family, community or society become relevant and must be adapted to the peculiar characteristics of genetic information. Therefore, with regard to genetic information, the "right for personal health" prevails (Rodotà, 2006).

Our results suggest that practices of dealing with digital health-related data vary depending on the different legal frameworks in which they are embedded. Also relevant are the respective social and cultural frameworks, which refer to standards of handling health-related data as well as the demands and acceptance of the relevant actors (Lupton, 2014). Moreover, technological progress often challenges legal frameworks with new implications. Since 2015, there is a "Law for Secure Digital Communication and Applications in Health Care" ("eHealth Law") in Germany (German Federal digital Law, 2015). This law provides for the establishment of an electronic patient record, in which patients can store the self-collected health data and make it available to their attending physician (Federal Ministry of Health, 2015).

In addition to technical and political aspects, the resulting legal and ethical consequences must also be considered (Frizzo-Barker et al., 2016). In our view, more comparative studies on data ownership involving lay people are necessary in order to better understand cultural differences such as attitudes towards the "right not to know" in the handling of genetic digital data.

The question of benefit primarily addresses different forms of individual interest or benefit provided by genetic information that go beyond improved health outcomes, and our findings indicate that the information is used for "potential" prevention for the benefit of others (e.g., future generations, one's own children). Cultural differences are evident in the value given to genetic information for preventing financial, family or professional problems. Those aspects were much more important to the German participants than to the Italian ones. In addition, there seems to be a cultural difference regarding the perception of genetic information as providing certainty, which was supported by about three-quarters of German participants but by far fewer Italian participants. In line with other studies, this could be an indication that in Germany genetic information is perceived to be very useful since it is a product of scientific insights and progress (Urban and Schweda, 2018).

Finally, some interesting differences emerged in our sample based on educational level and gender. It seems that people with an academic education tend to consider genetic risk information as something useful for the professional life planning and that a standardization of methods and limits for genetic analysis is paramount for them. Further, they are more interested in sharing results with their partner and have trust in other people when deciding to share their personal information such as a genetic risk predisposition. Moreover, among the Italian population, people with an academic degree also believed more in the notion that there is no "right not to know". Our results show that people with higher education show greater openness to share this type of personal information, especially if they are generated with reliable methods, and in particular in the Italian context, excluding the right not to know". Other studies have been conducted in the past on the attitudes toward genetic testing and their perceived utility, that have revealed differences based on the level of education, too (Haga et al., 2013; Roberts et al., 2017; Flatau et al., 2018; Schaper and Schicktanz, 2018). Our study also showed gender differences regarding the perceived utility of genetic testing and attitudes towards data sharing, such as the fact that women consider undergoing genetic testing as a preventive possibility and they want to share (and actually share) the results with parents more than men. Men on the other hand have higher privacy concerns and appear to be more interested in standardization of test methods and limits and directives for data protection, since they are worried about the possibility that persons or institutions can control them using genetic risk information. While gender differences regarding attitudes toward genetic testing have been observed in several other studies in different countries, there seems to be no clear recurring pattern this finding relates to, probably because of different studied populations and varying study designs and methods (Aro et al., 1997; Henneman et al., 2013).

LIMITATIONS

This work is explorative in nature and subject to several limitations in regard to representativeness. Given the narrow field of research and the research question, the total target population is very small. The difference in response rates between the countries may be attributed to the use of incentives in the German setting. However, in both countries the response rate was very low, which might lead to sample bias. We cannot generalize our findings to the broader population; however, we may assume that it is somewhat representative of the smaller target population. The invitation mail in the German data collection technically allowed participants to share the link or participate in the survey multiple times. In in an unknown number of cases, doctors keep patients' User IDs, making it impossible for the latter to respond. A limitation of the survey and related statistical analysis is the lack of continuous variables, which did not allow analysis of variance in investigating group differences.

CONCLUSIONS

Our survey demonstrates the importance of cross-cultural comparisons (Raz and Schicktanz, 2016) to better understand national differences and similarities in lay perspectives in regard to using und sharing genetic information to indicate responsibilities and reservations. Our findings contribute to the discussion about the personal utility of genetic information. Above all, the broad spectrum of different attitudes shows that lay people see a great potential for prevention, and that predictive genetic tests will in future increase lay people's perceived responsibility for their own health.

This raises the question of how individual autonomy and the right to know can be reconciled with the self-determination of family members and their right not to know. Predictive genetic tests can lead to an overestimation of the predictive ability of genetic information. At the same time, neglecting social risk factors for certain diseases could be both physically and psychologically detrimental for those affected. As we become increasingly exposed to genetic information in our lives, it is all the more important that we, as citizens, patients or consumers, are sensitized to, or "socialized" with, ethical questions arising from such information (Parry and Middleton, 2017; Roberts and Middleton, 2017). However, more information and educational work is needed while genetic information is combined with prevention measures aimed purely at medical interventions or family planning. In the private, family or professional spheres alike, there is a lack of information about which preventive measures can be affected by genetic knowledge. Communication challenges also arise beyond the handling of predictive information. For example, it is important not only to educate lay people about the opportunities and risks of using their genetic information, but also to avoid raising unrealistic expectations by, for example, making a factual distinction between individual therapeutic and future benefits for patients.

OUTLOOK

The sharing of genetic information *via* digitized patient records promises a more transparent, efficient and secure flow of information between patients, physicians and other groups in the healthcare system (Lupton, 2014). Therefore, in addition to technical and political solutions, the resulting legal and ethical consequences must also be considered (Frizzo-Barker et al., 2016). In our view, more comprehensive studies on data ownership involving lay people are necessary in order to do justice to cultural differences such as the "right not to know" in the digital age. Further, there seem to be interesting correlations between sociodemographic factors and willingness to share genetic information worth investigating. In order to evaluate future ethical problems that may arise through the integration of genetic information into eHealth and to guarantee informational self-determination, the perspectives of lay people (as users) should be taken into account along with those of experts during the development of these new digital technologies (Hartzler et al., 2013).

Since most users only partially comprehend the complex mutual relationship between data generation and use and their consequences, ethical aspects of dealing with digital health-related data, e.g. with regard to data protection and data autonomy, should be prioritized (Rothstein, 2015).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was carried out in Germany in accordance with the recommendations of 'University of Göttingen Human Research Review Committees Ref. Nr. 16/10/14' and in Italy the research protocol was approved by the Centro Interdipartimentale di Ricerca e Intervento sui Processi Decisionali (IRIDE) and the ethical committee of the University of Milan 08/14. All subjects gave consent to participation after reviewing the study information online. The protocol was approved by the 'University of Göttingen Human Research Review Committee'. All people enrolled in Italy gave consent to participation after reviewing the study information online.

AUTHOR CONTRIBUTIONS

SW, MS, and SS contributed to the design of the study, developed the survey, organized data collection in Germany and interpretation of data, and drafted the manuscript. DS (bio.logis ZfH) assisted in the design of survey and the data collection at bio.logis and worked on the manuscript. SO, IC, and FS organized data collection at GenomaLab in Italy, performed statistical analysis, including creating the figures and the table for this manuscript, and worked on the interpretation of the data. GP supported data collection and revised the contents of the *Results* and *Introduction*. All authors agreed to be accountable for all aspects of the work.

FUNDING

This research is developed within the framework of Mind the Risk [funded by the Swedish Riksbankens Jubileumsfond, 2015–2019 (grant no. 1351730)].

ACKNOWLEDGMENTS

We acknowledge the support by the Open Access Publication Funds of the Göttingen University

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00102/full#supplementary-material

REFERENCES

Aitken, M., de St. Jorre, J., Pagliari, C., Jepson, R., et al. (2016). Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med. Ethics.* 17, 73. doi: 10.1186/s12910-016-0153-x

Anderson, E., and Wasson, K. (2015). Personal narratives of genetic testing: expectations, emotions, and impact on self and family. *Narrat. Inq Bioeth.* 5, 229–235. doi: 10.1353/nib.2015.0062

Aro, A. R., Hakonen, A., Hietala, M., Lönnqvist, J., Niemelä, P., Peltonen, L., et al. (1997). Acceptance of genetic testing in a general population: age, education and gender differences. *Patient Educ. Couns.* 32 (1-2), 41–49. doi: 10.1016/S0738-3991(97)00061-X

Baars, J. E., Ausems, M. G., van Riel, E., Kars, M. C., and Bleiker, E. M. (2016). Communication between breast cancer patients who received inconclusive genetic test results and their daughters and sisters years after testing. *J. Genet. Couns.* 25 (3), 461–471. doi: 10.1007/s10897-015-9889-6

Borry, P., van Hellemondt, R. E., Sprumont, D., Jales, C. F., Rial-Sebbag, E., Spranger, T. M., et al. (2012). Legislation on direct-to-consumer genetic testing in seven European countries. *Eur. J. Hum. Genet.* 20, 715–721. doi: 10.1038/ejhg.2011.278

Bunnik, E. M., Janssens, A. C., and Schermer, M. H. (2015). Personal utility in genomic testing: is there such a thing? *J. Med. Ethics* 41, 322–326. doi: 10.1136/medethics-2013-101887

Burke, W. (2014). Genetic tests: clinical validity and clinical utility. *Curr. Protoc. Hum. Genet.* 81, 9.15.1–9.15.8. doi: 10.1002/0471142905.hg0915s81

Chadwick, R., Levitt, M., and Shickle, D. (Eds.). (2014). *The Right to Know and the Right Not to Know: Genetic Privacy and Responsibility* (Cambridge Bioethics and Law). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139875981

Corpas, M. (2012). A family experience of personal genomics. *J. Genet. Couns.* 21, 386–391. doi: 10.1007/s10897-011-9473-7

D'Agincourt-Canning, L. (2001). Experiences of genetic risk: disclosure and the gendering of responsibility. *Bioethics* 15, 231–247. doi: 10.1111/1467-8519.00234

DiMillio, J., Samson, A., Thériault, A., Lowry, S., Corsini, L., Shailendra, V., et al. (2015). Genetic testing: when prediction creates stigmatization. *J. Health Psychol.* 20 (4), 393–400. doi: 10.1177/1359105313502566

Domaradzki, J. (2015). Patient rights, risk, and responsibilities in the genetic era – a right to know, a right not to know, or a duty to know? *Ann. Agric. Environ. Med.* 22 (1), 156–162. doi: 10.5604/12321966.1141387

Domenichiello, M. (2015). State of the art in adoption of e-Health services in Italy in the context of European Union E-Government strategies. *Proc. Econ. Finance* 23, 1110–1118. doi: 10.1016/S2212-5671(15)00364-0

Flatau, L., Reitt, M., Duttge, G., Lenk, C., Zoll, B., Poser, W., et al. (2018). Genomic information and a person's right not to know: a closer look at variations in hypothetical informational preferences in a German sample. *PloS One* 13 (6), e0198249. doi: 10.1371/journal.pone.0198249

Frizzo-Barker, J., Chow-White, P. A., Charters, A., and Ha, D. (2016). Genomic big data and privacy: challenges and opportunities for precision medicine. *Comput. Supported Cooperative Work* 25 (2–3), 115–136. doi: 10.1007/s10606-016-9248-7

German Ethics Council. (2013). *Die Zukunft der genetischen Diagnostik – von der Forschung in die klinische Anwendung* (Berlin). Available at: https://www.ethikrat.org/en/publications/publication-details/?tx_wwt3shop_detail%5Bproduct%5D=11&tx_wwt3shop_detail%5Baction%5D=index&tx_wwt3shop_detail%5Bcontroller%5D=Products&cHash=a1ed4855ba228593e1996c9e66939f96 (Accessed February 12, 2020).

German Ethics Council. (2018). *Big Data and Health: Data Sovereignty as the Shaping of Informational Freedom* (Berlin). Available at: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/englisch/opinion-big-data-and-health-summary.pdf (Accessed February 12, 2020).

German Federal digital Law. (2015). https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl115s2408.pdf#:bgbl:%2F%2F*%5B%40attr_id%3D%27bgbl115s2408.pdf%27%5D:1559903184708 [accessed 07.07.2019].

German Genetic Diagnostics Act. (2009). (Gendiagnostikgesetz). https://www.gesetze-im-internet.de/gendg/GenDG.pdf (accessed May 13, 2019).

Hafen, E., Kossmann, D., and Brand, A. (2014). Health data cooperatives - citizen empowerment. *Methods Inf Med.* 53, 82–86. doi: 10.3414/ME13-02-0051

Haga, S. B., and O'Daniel, J. (2011). Public perspectives regarding data-sharing practices in genomics research. *Public Health Genomics* 14, 319–324. doi: 10.1159/000324705

Haga, S. B., Barry, W. T., Mills, R., Ginsburg, G. S., Svetkey, L., Sullivan, J., et al. (2013). Public knowledge of and attitudes toward genetics and genetic testing. *Genet. testing Mol. biomarkers.* 17 (4), 327–335. doi: 10.1089/gtmb.2012.0350

Hartzler, A., McCarty, C. A., Rasmussen, L. V., Williams, M. S., Brilliant, M., Bowton, E. A., et al. (2013). Stakeholder engagement: a key component of integrating genomic information into electronic health records. *Genet. Med.* 15, 792–801. doi: 10.1038/gim.2013.127

Henneman, L., Vermeulen, E., Van El, C. G., Claassen, L., Timmermans, D. R., and Cornel, M. C. (2013). Public attitudes towards genetic testing revisited: comparing opinions between 2002 and 2010. *Eur. J. Hum. Genet.* 21 (8), 793–799. doi: 10.1038/ejhg.2012.271

Kalokairinou, L., Howard, H. C., Slokenberga, S., Fisher, E., Flatscher-Thöni, M., Hartlev, M., et al. (2018). Legislation of direct-to-consumer genetic testing in Europe: a fragmented regulatory landscape. *J. Community Genet.* 9 (2), 117–132. doi: 10.1007/s12687-017-0344-2

Kohler, J. N., Turbitt, E., and Biesecker, B. B. (2017a). Personal utility in genomic testing: a systematic literature review. *Eur. J. Hum. Genet.* 25, 662–668. doi: 10.1038/ejhg.2017.10

Kohler, J. N., Turbitt, E., Lewis, K. L., Wilfond, B. S., Jamal, L., Peay, H. L., et al. (2017b). Defining personal utility in genomics: a delphi study. *Clin. Genet.* 92, 290–297. doi: 10.1111/cge.12998

Kollek, R., and Lemke, T. (2008). *Der medizinische Blick in die Zukunft* (Gesellschaftliche Implikationen prädiktiver Gentests. Frankfurt am Main/New York: Campus).

Konrad, M. (2005). *Narrating the New Predictive Genetics: Ethics, Ethnography and Science.* Cambridge: Cambridge University Press.

Leefmann, J., Schaper, M., and Schicktanz, S. (2017). The concept of "genetic responsibility" and its meanings: a systematic review of qualitative medical sociology literature. *Front. In Sociology/ELSI In Sci. Genet.* doi: 10.3389/fsoc.2016.00018

Lemke, A. A., Wolf, W. A., Herbert-Beirne, J., and Smith, M. E. (2010). Public and biobank participant attitudes toward genetic research participation and data sharing. *Public Health Genomics* 13, 368–377. doi: 10.1159/000276767

Lupton, D., and Michael, M. (2017). Depends on who's got the data: public understandings of personal digital dataveillance. *Surveillance Soc.* 15, 254–268. doi: 10.24908/ss.v15i2.6332

Lupton, D. (2014). Critical perspectives on digital health technologies, sociology compass. *Sociol. Compass* 8, 1344–1359. doi: 10.1111/soc4.12226

Muthuswamy, V. (2011). Ethical issues in genetic counselling with special reference to haemoglobinopathies. *Indian J. Med. Res.* 134, 4.

Oliveri, S., and Pravettoni, G. (2018). Capturing how individuals perceive genetic risk information: a phenomenological perspective. *J. Risk Res.* 21, 259–267. doi: 10.1080/13669877.2017.1281333

Oliveri, S., Renzi, C., and Pravettoni, G. (2015). Toward an in-depth profiling of DTC users. *Clin. Genet.* 88, 505–506. doi: 10.1111/cge.12599

Oliveri, S., Masiero, M., Arnaboldi, P., Cutica, I., Fioretti, C., and Pravettoni, G. (2016a). Health orientation, knowledge, and attitudes toward genetic testing and personalized genomic services: preliminary data from an Italian sample. *BioMed. Res. Int.* 2016, 6824581. doi: 10.1155/2016/6824581

Oliveri, S., Pravettoni, G., Fioretti, C., and Hansson, M. G. (2016b). Let the individuals directly concerned decide: a solution to tragic choices in genetic risk information. *Public Health Genomics* 19, 307–313. doi: 10.1159/000448913

Oliveri, S., Ferrari, F., Manfrinati, A., and Pravettoni, G. (2018). A systematic review of the psychological implications of genetic testing: a comparative analysis among cardiovascular, neurodegenerative and cancer diseases. *Front. In Genet.* 9, 624. doi: 10.3389/fgene.2018.00624

Oviedo Convention (1997). Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: convention on human rights and biomedicine. ETS No.164. https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164 (accessed May 13, 2019).

Parens, E., and Appelbaum, P. S. (2019). Bioethicists worried patients couldn't handle their own genetic testing results. They were (mostly) wrong. *STAT.*

https://www.statnews.com/2019/07/30/genetic-information-disclosure/ (Accessed February 20, 2020).

Parry, V., and Middleton, A. (2017). Socialising the genome. *Lancet* 389, 1603–1604. doi: 10.1016/S0140-6736(17)31011-5

Paton, C., Hansen, M. M., Fernandez-Luque, L., and Lau, A. Y. S. (2012). Self-tracking, social media and personal health records for patient empowered self-care. *Yearbook Med. Inf.* 7 (1), 16–24.

Poss-Doering, R., Kunz, A., Pohlmann, S., Hofmann, H., Kiel, M., Winkler, E. C., et al. (2018). Utilizing a prototype patient-controlled electronic health record in germany: qualitative analysis of user-reported perceptions and perspectives. *JMIR Formativ Res.* 2 (2), e10411. doi: 10.2196/10411

Prainsack, B. (2017). *Personalized Medicine. Empowered Patients in the 21st Century?* (New York: New York University Press).

Raz, A. E., and Schicktanz, S. (2009). Diversity and uniformity in genetic responsibility: moral attitudes of patients, relatives and laypeople in Germany and Israel. *Med. Health Care Philos.* 12, 433–442. doi: 10.1007/s11019-009-9215-x

Raz, A. E., and Schicktanz, S. (2016). *Comparative Empirical Bioethics: Dilemmas of Genetic Testing and Euthanasia in Israel and Germany* (Springer: Springer International Publishing AG Switzerland).

Rexhepi, H., Åhlfeldt, R.-M., Cajander, Å, and Huvila, I. (2018). Cancer patients' attitudes and experiences of online access to their electronic medical records. *A Qual. Study Health Inf. J.* 24 (2), 115–124. doi: 10.1177/1460458216658778

Roberts, J., and Middleton, A. (2018). Genetics in the 21st Century: Implications for patients, consumers and citizens. *F1000Res* 6, 2020. doi: 10.12688/f1000research.12850.2

Roberts, J. S., Gornick, M. C., Carere, D. A., Uhlmann, W. R., Ruffin, M. T., and Green, R. C. (2017). Direct-to-consumer genetic testing: user motivations, decision making, and perceived utility of results. *Public Health Genomics* 20 (1), 36–45. doi: 10.1159/000455006

Rodotà, S. (2006). La vita e le regole: tra diritto e non diritto. Milano: Giangiacomo Feltrinelli Editore.

Ross, J., Stevenson, F., Lau, R., and Murray, E. (2015). Exploring the challenges of implementing e-health: a protocol for an update of a systematic review of reviews. *BMJ Open* 5, e006773. doi: 10.1136/bmjopen-2014006773

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *J. Law Med. Ethics* 43 (2), 425–429. doi: 10.1111/jlme.12258

Schaper, M., and Schicktanz, S. (2018). Medicine, market and communication: ethical considerations in regard to persuasive communication in direct-to-consumer genetic testing services. *BMC Med. Ethics.* 19, 56. doi: 10.1186/s12910-018-0292-3

Schaper, M., Wöhlke, S., and Schicktanz, S. (2018). "I would rather have it done by a doctor" - Laypeople's perceptions of direct-to-Consumer Genetic Testing (DTC GT) and its ethical implications. *Med. Health Care Philos.* 22, 31–40. doi: 10.1007/s11019-018-9837-y

Slaughter, M. L. (2006). Genetic Testing and Discrimination: How Private is Your Information? *Stanford Law Pol. Rev.* 17, 67–81.

Soini, S. (2012). Genetic testing legislation in Western Europe—a fluctuating regulatory target. *J. Community Genet.* 3, 143–153. doi: 10.1007/s12687-012-0078-0

Sommer, J. E., Sii, F., Bourne, R., Cross, V., and Shah, P. (2013). What do patients with glaucoma think about personal health records? *Ophthalmic Physiol. Opt.:* 33, 627–633. doi: 10.1111/opo.12084

Thiel, R., Deimel, L., Schmidtmann, D., Piesche, K., Hüsing, T., and Rennoch, J. (Eds.). (2019). #SmartHealthSystems: international comparison of digital strategies. Gütersloh: Bertelsmann-Stiftung.

Thorogood, A., Bobe, J., Prainsack, B., Middleton, A., Scott, E., Nelson, S., et al. (2018). APPLaUD: access for patients and participants to individual level uninterpreted genomic data, human genomics. *Hum. Genomics* 12, 7. doi: 10.1186/s40246-018-0139-5

Townsend, A., Adam, S., Birch, P. H., Lohn, Z., et al. (2012). "I want to know what's in pandora's box": comparing stakeholder perspectives on incidental findings in clinical whole genome sequencing. *Am. J. Med. Genet.* A 158A, 2519–2525. doi: 10.1002/ajmg.a.35554

Urban, A., and Schweda, M. (2018). Clinical and personal utility of genomic high-throughput technologies: perspectives of medical professionals and affected persons. *New Genet. Soc.* 37, 153–173. doi: 10.1080/14636778.2018.1469976

Varga, O., Soini, S., Kääriäinen, H., Cassiman, J. J., Nippert, I., Rogowski, W., et al. (2012). Definitions of genetic testing in European legal documents. *J. Community Genet.* 3, 125–141. doi: 10.1007/s12687-012-0077-1

Wöhlke, S., and Perry, J. (2019). Responsibility in dealing with genetic risk information. *Soc. Theor. Health* 1–22. doi: 10.1057/s41285-019-00127-8

Wöhlke, S., Perry, J., and Schicktanz, S. (2015). "Taking it Personally: Patients' Perspectives on Personalised Medicine and its Ethical Relevance," in *Personalised Medicine: Ethical, Medical, Economic and Legal Critical Perspectives*. Eds. J. Vollmann, V. Sandow, S. Wäscher and J. Schildmann (Ashgate: Farnham).

Wöhlke, S., Schaper, M., and Schicktanz, S. (2019). How uncertainty influences lay people's attitudes and risk perceptions concerning predictive genetic testing and risk communication. *Front. Genet.* 10, 380. doi: 10.3389/fgene.2019.00380

Welch, H. G., and Burke, W. (1998). Uncertainties in genetic testing for chronic disease. *JAMA* 280, 1525–1527. doi: 10.1001/jama.280.17.1525

Wicks, P., Massagli, M., Frost, J., Brownstein, C., et al. (2010). Sharing health data for better outcomes on patientslikeme. *J. Med. Internet Res.* 12 (2), e19. doi: 10.2196/jmir.1549

Conflict of Interest: Author DS is founder of the company bio.logis genetic information management.

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.

Copyright © 2020 Wöhlke, Schaper, Oliveri, Cutica, Spinella, Pravettoni, Steinberger and Schicktanz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 28 April 2020
doi: 10.3389/fgene.2020.00416

Genetic Privacy and Data Protection: A Review of Chinese Direct-to-Consumer Genetic Test Services

Li Du and Meng Wang*

Faculty of Law, University of Macau, Macau, China

Background: The existing literature has not examined how Chinese direct-to-consumer (DTC) genetic testing providers navigate the issues of informed consent, privacy, and data protection associated with testing services. This research aims to explore these questions by examining the relevant documents and messages published on websites of the Chinese DTC genetic test providers.

Methods: Using Baidu.com, the most popular Chinese search engine, we compiled the websites of providers who offer genetic testing services and analyzed available documents related to informed consent, the terms of services, and the privacy policy. The analyses were guided by the following inquiries as they applied to each DTC provider: the methods available for purchasing testing products; the methods providers used to obtain informed consent; privacy issues and measures for protecting consumers' health information; the policy for third-party data sharing; consumers right to their data; and the liabilities in the event of a data breach.

Results: 68.7% of providers offer multiple channels for purchasing genetic testing products, and that social media has become a popular platform to promote testing services. Informed consent forms are not available on 94% of providers' websites and a privacy policy is only offered by 45.8% of DTC genetic testing providers. Thirty-nine providers stated that they used measures to protect consumers' information, of which, 29 providers have distinguished consumers' general personal information from their genetic information. In 33.7% of the cases examined, providers stated that with consumers' explicit permission, they could reuse and share the clients' information for non-commercial purposes. Twenty-three providers granted consumer rights to their health information, with the most frequently mentioned right being the consumers' right to decide how their data can be used by providers. Lastly, 21.7% of providers clearly stated their liabilities in the event of a data breach, placing more emphasis on the providers' exemption from any liability.

Conclusions: Currently, the Chinese DTC genetic testing business is running in a regulatory vacuum, governed by self-regulation. The government should develop a

OPEN ACCESS

Edited by:
Manuel Corpas,
Cambridge Precision Medicine,
United Kingdom

Reviewed by:
Mahsa Shabani,
Ghent University, Belgium
Susan E. Wallace,
University of Leicester,
United Kingdom

****Correspondence:***
Li Du
stephendu@um.edu.mo

Specialty section:
This article was submitted to
ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

Received: *08 January 2020*
Accepted: *01 April 2020*
Published: *28 April 2020*

Citation:
Du L and Wang M (2020) Genetic Privacy and Data Protection: A Review of Chinese Direct-to-Consumer Genetic Test Services. Front. Genet. 11:416. doi: 10.3389/fgene.2020.00416

comprehensive legal framework to regulate DTC genetic testing offerings. Regulatory improvements should be made based on periodical reviews of the supervisory strategy to meet the rapid development of the DTC genetic testing industry.

Keywords: direct-to-consumer genetic testing, Chinese genetic testing providers, genetic information protection, data privacy, consumers' right to their genetic information, data sharing, data protection law

INTRODUCTION

Direct-to-consumer (DTC) genetic testing has gained increasing popularity internationally. The market for DTC genetic testing is estimated to reach 20 billion by 2024 (Global Market Insights, 2018). In recent years, many test providers in China have started to advertise and sell testing products directly to consumers. Similar with providers in the United States such as Ancestry.com and 23andMe, Chinese DTC companies offers genetic testing services for both illness risk determination and lifestyle guidance purposes (Zhao et al., 2013). For example, WeGene, a Shenzhen-based company provides DTC genetic tests for ancestral analysis, personalized sports and weight loss suggestions, nutritional genomics, and genomic medicine, etc. With the increasing influence of popular culture on the public perception of genetics, Chinese consumers' interest in genetic testing is also estimated to gain a considerable increase in the coming years (Luo et al., 2020). According to a 2018 report developed by Yi Ou, an independent commercial consultant company, the number of consumers for DTC genetic testing will increase exponentially in the next 5 years, from 1.52 million in 2018 to 56.8 million in 2022 (YiOu ZhiKu, 2018).

The scientific community and regulatory authorities have consistently questioned the reliability and clinical validity of DTC genetic testing results as the products have become more widely available to the mass market (Caulfield and McGuire, 2012; Covolo et al., 2015; Webborn et al., 2015). Many studies indicate that the offering of DTC genetic testing may pose risks to privacy and data protection, which may result in potential societal harm to consumers (Hall et al., 2017; Niemiec et al., 2017; Hazel and Slobogin, 2018). Confronted with these controversies and concerns of protecting human genetic resources and biosafety, both the China General Administration of Food and Drug Administration (CFDA) and the State Health Planning Commission (now the State Health Commission) jointly issued the *Notice on Strengthening the Management of Products and Technologies Related to Clinical Use of Gene Sequencing* in February 2014, with the *Notice* suspending all genetic testing services in China (China General Administration of Food and Drug Administration, and State Health Planning Commission, 2014). According to the *Notice*, the technology and products related to clinical genetic testing shall be approved and registered by the CFDA and State Health Planning Commission before entering into the market (Lenore et al., 2016). In July 2014, the CFDA approved the second-generation gene sequencing diagnosis, which can be used for non-invasive prenatal examination for fetal chromosomal aneuploidy disease (Jin et al., 2018). Since then, the CFDA has not approved any other genetic sequencing technology. The *Notice* has played an important role in the clinical application of gene detection technology, but it does not address DTC genetic offerings, thus the supervision of DTC has been operating in an irrefutable gray area.

As a business operator, a DTC genetic testing company should follow the requirements stipulated in the *Law of the People's Republic of China on the Protection of Consumer* Rights and Interests 2013 (the *Consumer Protection Law*) when collecting and using consumers' personal information. Companies should inform and obtain the consent of consumers regarding the purposes and scope of collection and use of personal information (The Standing Committee of the National People's Congress, 2013). The testing company and its employees must keep the consumers' personal information confidential and should not illegally provide such information to others. To date, there has been no special legislation on DTC genetic testing services. Relevant laws may be applicable to regulate genetic testing offerings, but the main purposes of the current laws are to protect human genetic resources rather than patients' or consumers' rights. For example, the National State issued the *Regulation of the People's Republic of China on Human Genetic Resources Management* (the *Regulation on Human Genetic Resources*) on May 28, 2019 (China State Council, 2019). The new *Regulation* is developed based on a previous National State administrative regulation, the 1998 *Interim Measures for the Management of Human Genetic Resources* (National Intellectual Property Administration, 2019). Compared with the old version, the new *Regulation* places more emphasis on the protection of the privacy and rights of data subjects, including the rights to voluntarily participate and withdraw from the data collection (XinHuaNet, 2019). According to the new *Regulation*, genetic testing providers shall respect consumers' privacy and cannot collect and use consumers' genetic data without their informed consent (China State Council, 2019). However, the *Regulation* does not provide detailed requirements for informed consent and privacy protection, as the main goal for the new *Regulation* is to effectively protect and rationally utilize human genetic resources in China. As such, it is focused more on safeguarding public health, national security, and social public interests (China State Council, 2019).

In terms of privacy and personal data protection, China does not have special legislation for the protection of personal data – including genetic data – and privacy at the national law level. In September of 2018, the Standing Committee of the National People's Congress of China (the SCNPCC) launched a legislative agenda for a comprehensive data protection law, a few months after the European Union General Data Protection Regulation entered into force on May 25, 2018. The plan shows the direction of China's data protection scheme, and the law is planned to be enacted in 2022 (Feng, 2019). However, relevant laws and standards are applicable to protect consumers' personal information including genetic data in the context of DTC genetic testing services. For example, the *Cybersecurity Law of the People's Republic of China* (the *Cybersecurity Law*), which was promulgated by the SCNPCC in 2016, requests that DTC genetic testing providers must not steal or use other illegal means to obtain consumers' personal information including biometric information, nor illegally sell or provide consumers' personal information to others (Huang, 2019). Moreover, in 2017, China's National Information Security Standardization Technical Committee issued the *Personal Information Security Specification*, a national standard that covers the collection, storage, use, transfer, and disclosure of personal information. Personal genetic information is clearly defined as a type of biometric information and categorized as personal sensitive information (China's National Information Security Standardization Technical Committee, 2017). Different from the *Cybersecurity Law*, which focuses on the regulation of network security and only provides general principles for personal data protection, the *Specification* targeted the protection of personal information and established detailed guidelines for data compliance (Chen and Song, 2018). For example, the *Specification* specifies the details of the content that should be included in the privacy policy and provides a privacy policy template. A genetic testing provider can use the *Specification* as a guideline to set up their specific privacy policy and standards for collecting, storing, using, and processing personal information when dealing with consumer genetic data.

The lack of effective supervision in DTC genetic testing offerings has gained increasing attention from the Chinese news media (Ha, 2019). Many news reports have criticized that the regulatory gap in the industry may result in poor quality testing results and damage to consumer's health information and privacy (Sui and Sleeboom-Faulkner, 2015). Previous research on US and EU-based DTC genetic testing services indicated that informed consent and privacy protection had been poorly implemented by DTC genetic testing providers. For example, a 2008 study led by Hogarth et al. (2008) highlighted the potential danger of discrimination due to consumer privacy breaches in the implementation of DTC genetic testing. In a 2012 review article, Caulfield and McGuire again recognized the potential privacy issues, revealing rather poor management among DTC genetic companies of addressing consumers' privacy protection (Caulfield and McGuire, 2012). More recently, in 2018, Overmaat et al. (2018) investigated five leading DTC genetic testing providers in China by ordering products and comparing the different testing results between companies. Their study revealed that, other than technical defects, there were prominent problems in the communication of the test results, with inadequate informed consent being one of the points of concern. However, few studies have been devoted to examining the nuanced perspectives of the Chinese DTC genetic testing offerings. For example, it is unclear what channels providers offered to consumers who are considering the purchase of genetic testing products, as well as what measures the DTC genetic testing companies use to protect consumers' personal health information, how consumers' data are shared with a third party, and what rights the consumers have to their data. Aiming to explore these important questions, this study reviews the websites of the Chinese companies and organizations that offer DTC genetic testing services, with a focus on examining all available Terms of Service (ToS), Privacy Policy (PP), and Informed Consent Forms (ICF).

MATERIALS AND METHODS

First Round

From January 17, 2019 to February 27, 2019, we used Baidu.com, the most popular search engine in China, and search keywords: "genetic testing" (in Chinese: "基因检测") to identify and collect providers that use the DTC model to market genetic testing services and products. Based on this search, we collected 90 DTC genetic testing offerings. We then visited the websites and captured the webpages of the providers and downloaded all available documents related to the terms of service, privacy protection, and informed consent. It is essential to clarify that our study focuses on examining organizations that mainly offer genetic testing services and products.

From May 7, 2019 to June 18, 2019, one of our authors analyzed the websites using a coding framework focused on the following perspectives: (1) channels provided for purchasing genetic testing products and services; (2) informed consent; (3) privacy issues; (4) strategies used to protect consumers' personal health information; (5) data sharing with a third party; (6) consumers' rights to their health information; (7) responsibility for data breach; and (8) specific laws or legal protections mentioned. These eight perspectives were established based on previous studies on legal and ethical issues associated with DTC genetic testing and personal data protection in the big data era in healthcare (Hogarth et al., 2008; Mostert et al., 2016). The coding framework included 20 items which were developed based on an exploratory content analysis of 30% of the dataset. In this round of content analysis, we found that three websites that were initially identified as genetic testing providers were no longer accessible, and thus were removed from our dataset.

Second Round

Later, on July 1, 2019, a new Decree of the State Council of China took effect, the *Regulations on the Management of Human Genetic Resources of the People's Republic of China*. Since the collection of website information and the content analysis of consumer-related legal documents were carried out before the implementation of the new *Regulation*, we took advantage of the opportunity to examine how Chinese DTC genetic testing providers reacted to the new *Regulation*. For example, we attempted to identify if there were updates to their ToS, PP, and ICF to meet the requirements of the updated requirements 3 months after the *Regulation* took effect. Consequently, we revisited the websites of the companies collected in our dataset and again reviewed the content of the websites and the legal documents to determine if any changes were made to comply with the new *Regulation*. In this second round of collection and analysis, we found that the original links of nine websites were broken. Five of these nine problematic websites changed to new domain names, and four of the nine websites were completely invalid. As a result, our dataset in the second round consists of 83 accessible websites. We analyzed these 83 websites using the same coding framework from September 2, 2019 to September 20, 2019. We used the 83 websites as the final dataset for this research and compared the results of the two rounds of analyses.

For both rounds of content analysis, 30% of the websites were randomly selected to compare the consistency of coding results in order to verify the reliability of the coding. After obtaining the 30% of websites, an independent coder searched the URLs of the websites, reviewed the content and available privacy policy, informed consent forms, and the terms of service. We calculated the agreement between the two codes, using Cohen's Kappa evaluation. The agreement was between 0.85 and 1.00 for all coding frame items, which indicates substantial to perfect agreement (Landis and Koch, 1977).

RESULTS

Methods Provided for Purchasing Genetic Testing Services and Products

Fifty-seven providers of DTC genetic testing services and products offered online purchase options via their websites. Twenty-six providers only accepted traditional banking transfers as payments after counseling with consultants via telephone. For providers that offered online purchases, 21 providers required consumers to register on the websites before they could order services and products online, and five websites integrated with other e-commerce platforms[1,2], which allowed consumers to order genetic testing products or services through a third party. In particular, we found that WeChat, the most widely used social media platform in China, became a popular vehicle for providers to promote their products and services. Thirty-two providers used WeChat to introduce their products and to offer follow-up services to consumers. Consumers could order products and make payments via providers' WeChat stores or by transferring money after adding the providers as WeChat friends. In contrast, very few providers ($n = 3$) used QQ, an instant messenger that at one time was the most widely used in China, as a promotion platform. The case for the relatively few providers using QQ to reach consumers might arguably be due to the increase in WeChat users.

Informed Consent

We only identified 6% of providers ($n = 5$) that provided informed consent forms on their websites. The promulgation of the new *Regulation* did not make a big difference in the availability of informed consent forms. Only one provider added an informed consent form to its website after the new *Regulation* took effect.

Every informed consent form exceeded 500 words. Two informed consent forms were between 1,000 and 2,000 words, and two had more than 2,000 words. In terms of content, all informed consent forms mentioned the protection of consumer privacy and the risk of information leakage. One provider enumerated all possible risks of implementing genetic testing, including: (1) consumers or their families may feel uncomfortable because of survey questions or genetic data results; (2) information leakage due to security breach; (3) information leakage caused by consumers sharing accounts and passwords with others; and (4) other currently unforeseeable risks. Three providers mentioned that their genetic testing reports were predictive, and which can only be used as a health consulting reference not as a clinical diagnostic basis.

With regard to the remaining websites ($n = 78$), although they failed in providing specific informed consent forms, 13 providers did mention the informed consent procedure on their websites. For example, 11 websites indicated that informed consent procedures would be implemented by providers during or after the purchases of the genetic testing services and products.

Privacy Issues

In our first-round examination, we found that 38 websites had provided accessible links to privacy policies. During the second round of collection, one company within the 38 was removed from the dataset because its website was no longer accessible.

Consequently, we identified 37 websites in total that offered a privacy policy. This means, however, that more than half of the websites (55.4%) did not provide consumers with a privacy policy ($n = 46$). Four companies mentioned privacy issues, but they did not offer a privacy policy on their websites. For example, two websites stated in their FAQ section that they had addressed privacy issues in their privacy policies, but offered no links to the privacy policies on their websites. Similarly, two websites had privacy policy tags at the bottom of the webpages, but clicking on these tags did not lead to valid links to the privacy policies. Moreover, in our second-round investigation, we did not identify an obvious difference in the provision of the privacy policy. In fact, except for one provider who updated the privacy policy by adding one sentence addressing the liability distributions in the event of a privacy breach, the rest of the providers did not make any changes to their privacy policies.

Privacy policies with less than 1,000 words were generally not written in an agreement format, instead functioning more like privacy statements, where short sentences were used to indicate service providers' attempts for protecting the consumer's privacy (see **Table 1**). In contrast, privacy policies with 1,000 to 2,000 words or more were generally written in an agreement format, including definitions of the terms involved in the agreements, detailed explanations of the rights and obligations of users and providers, and applicable laws. However, only privacy policies with more than 2,000 words meet the requirements of the *Personal Information Security Specification* on the content of privacy policy.

[1] JD.com

[2] Taobao.com

TABLE 1 | Examples of different lengthy privacy policy used by DTC genetic testing companies.

Word counts	Number of providers (number of providers whose privacy policy covers the content required by the *Personal Information Security Specification*)	Summary of the main content
Less than 200 words	9 (0)	Service providers will protect the genetic privacy of consumers, but no specific protection measures are mentioned.
200–500 words	12 (0)	Privacy clauses focus on privacy protection in the collection and process of personal data, the use and disclosure of personal data, and privacy security. However, the specific contents and measures are not mentioned and explained.
500 – 100 words	4 (0)	Privacy clauses clarify the rights and obligations of users and providers. The exemption clause is included.
1,000–2,000 words	7 (0)	The privacy policy is in the format of an agreement. In general, it includes: definitions of terms used in the agreement, the rights and obligations of users and providers, and applicable and governed laws.
More than 2,000 words	5 (5)	The privacy policy is comprehensive and meets all the requirements of the *Personal Information Security Specification* on the content of privacy policy, including the collection and use of users' personal information, the use of cookies and similar technologies, the sharing, transfer, and disclosure of users' personal information, the measures for protecting user's personal information, the rights of users, and the methods for dealing with children's information, etc.

Strategies Used to Protect Clients' Personal Health Information

Thirty-nine websites published statements either on the front page of their website, their ToS, PP, or ICF declaring that they use measures to safeguard the security of consumers' genetic information. Among these 39 providers, three were identified in the second round of investigation after the new *Regulation* was issued. It is worth noting that 29 privacy policies had distinguished general personal information (GPI) of consumers, such as website registration ID, social security ID, and health information, from their genetic information. Seventeen providers proposed concrete measures for protecting consumer GPI (see **Figure 1**). The most frequently mentioned measures include: using technical methods to keep the confidentiality of GPI and maintain it regularly (mentioned by 13 providers); storing consumer GPI separately so that staff analyzing the genetic information are not able to identify the subject of the genetic data (mentioned by 10 providers); establishing an ethics committee for supervising the protection of GPI (mentioned by seven providers). In terms of consumer genetic data, the measure that discloses how consumer genetic information is stored in laboratories, encrypted, backed up, and maintained regularly is the most frequently stated method (mentioned by 21 providers) (see **Figure 2**). However, none of the providers had clearly stated how long the consumers' data would be kept, and whether the data will be eventually destroyed.

Data Sharing With a Third Party

While 62.7% of providers (n = 52) did not address the reuse, selling, or sharing of information gathered from consumers in their informed consent forms, privacy policies, or terms of service, 28 providers stated that with additional client permissions, the providers can reuse and share the clients' information for non-commercial purposes. Two providers mentioned that they would not sell a client's information unless having obtained the client's additional permission. Only one company mentioned that it would reuse and share a client's information for non-commercial purposes without the client's further permission. Specifically, the company stated that: in a case where the third party agrees to assume the same responsibility of protecting users' privacy as the company does, the company can provide the third party with users' registration and other information without consumers' further permissions. Moreover, 22 websites mentioned the compelled disclosure of personal health information in accordance with laws in either the ICF or PP. In regards to information disclosure, 73.5% of providers (n = 61) did not make any statements about mandatory information disclosure on their websites.

Consumer Rights to Their Health Information

Twenty-three providers granted consumer rights to their health information, while the rest of the providers (n = 60) kept silent in this respect. In general, three types of rights have been granted by the DTC genetic testing providers to their customers. Among these, the most frequently mentioned right is the consumers' right to decide whether providers can use their genetic data for follow-up research or provide their data to third parties (mentioned by 21 providers). Nine providers stated that consumers have the right to view and change their personal data or to remove their data from the providers' database. Only one company mentioned that consumers have the right to be informed of follow-up use of their data, which includes: (1) using users' genetic data to develop new products based on gene sequencing results; and (2) using genetic data for the latest interpretation of existing projects, interpreting the latest scientific literature, and recalculating existing projects more accurately.

Accountability and Responsibility for Data Breach

While nearly 70% of providers (n = 58) did not inform consumers of the risk of accidental information leakage, 30.1% of providers (n = 25) mentioned relevant risks on their websites, e.g.,

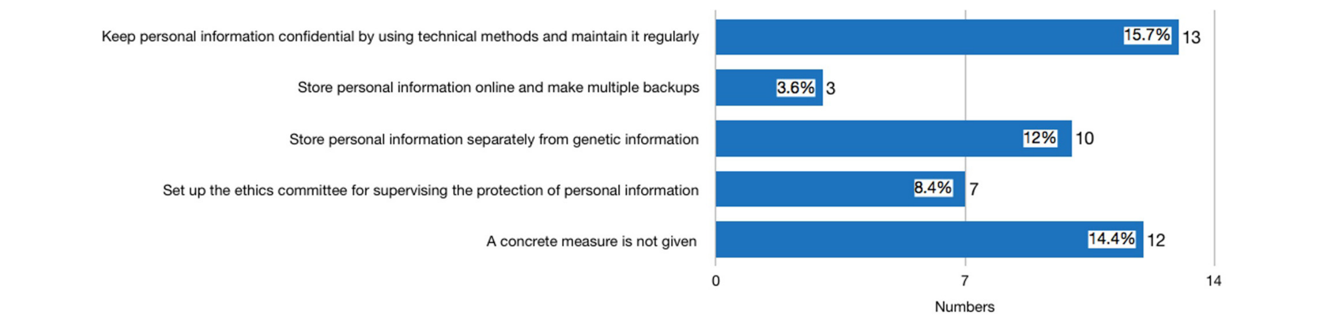


FIGURE 1 | What measures will the provider adopt to protect general personal information.

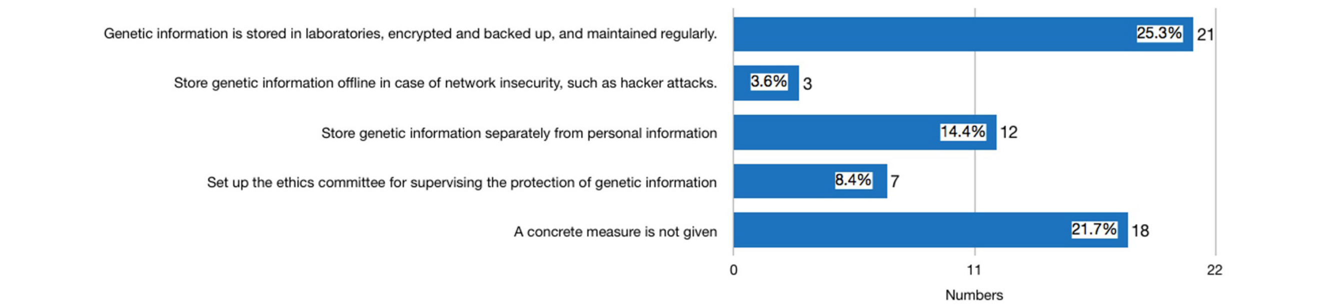


FIGURE 2 | What measures will the provider adopt to protect genetic data.

hacker attacks, internet errors, and other unforeseeable accidents. Among these 25 providers, all stated that the company would strive to protect customers' privacy, preventing their health information from being disclosed arbitrarily. However, in terms of the distribution of responsibility for the breach of privacy, fewer providers (n = 18) made explicit statements regarding whether they shall bear liability. In particular, nine providers did not address who would assume responsibility in the event of a data breach, but clearly stated that providers will be exempted from liability if the leaks are due to causes beyond their control. Eight companies specified only the consumers' responsibilities but not the obligations of providers in a breach. Just one company stated in their PP that in the event of a data leakage, that the liability would fall on the source of the breach, i.e., the source of the leak will bear the responsibility.

Law Mentioned

Twenty-three Chinese DTC genetic testing websites in their ICF, ToS, or PP stated that the company would comply with relevant Chinese laws in collecting, storing, and using consumer genetic data during and after genetic testing services. However, among these 23 websites, only four companies mentioned concrete laws, e.g., the *Cybersecurity Act* (mentioned by two companies), the *Regulation on Human Genetic Resources* (mentioned by one company), and the *Interim Measures for the Management of Human Genetic Resources* (mentioned by one company), which had been replaced by the *Regulation on Human Genetic Resources*.

It is worth noting that three companies mentioned the *Health Insurance Portability and Accountability Act* (HIPAA), the United States legislation that protects the privacy and security of Americans' medical information in the PP, ICF, or data protection agreement for dealing with issues of privacy protection. We found that HIPAA was applied in different ways. For example, one company mentioned that the storage of consumers' genetic data will strictly abide by the HIPAA requirements. One company stated that for international users, the process of the genetic testing services will follow HIPPA standards. Another company used the HIPAA as a reference for their practice of storing consumers' genetic data. In addition, one company promised that it would abide by the *Hong Kong Personal Data (Privacy) Ordinance* to carry out genetic testing services.

DISCUSSION

Our research indicates that Chinese DTC genetic testing companies have begun to take action in protecting consumer genomic data privacy. For example, many providers developed specific measures for protecting the security and privacy of consumer health information. Our research particularly indicates that some DTC genetic testing providers separated the consumer general health information from their genetic information and used different measures to protect the two types of data. This may reflect the industry's growing awareness of the sensitivity of genetic information and the need for using special measures to protect consumers' genetic data. Moreover, consumers' rights to their health information were recognized by many DTC genetic testing companies. For example, consumers have the right to access their personal data or have their information removed from company records. Nonetheless, our research has identified several legal concerns that Chinese regulatory bodies should immediately address.

First and foremost, we found that the provision of informed consent forms is not a common practice for the Chinese DTC genetic testing providers. This finding is in agreement with previous studies on international DTC genetic testing companies, where informed consent practices were found to be inadequate and sometimes misleading (Howard et al., 2010; Lachance et al., 2010; Niemiec et al., 2016). Until very recently, informed consent in China was not thoroughly implemented in general clinical practice (Wu et al., 2019). Previous studies have indicated that informed consent has grown in importance over the years as an effective strategy in softening the increased tensions between physicians and patients (Bal and Brenner, 2015). Still, consumers may not be aware of the importance of informed consent in DTC genetic testing services. For instance, many consumers may not realize that they need a supportive process to fully and properly understand the purposes and possible results of the testing, and, more importantly, the impact of testing results and genetic information on their health and other interests (Deng and Liu, 2017). In particular, as opposed to other countries where informed consent is a legal requirement for implementing genetic testing services (Knoppers et al., 2015), China has not established such legal requirements for requesting a mandatory informed consent process before receiving DTC genetic testing.

Moreover, we found that many Chinese DTC genetic testing companies offer both non-health-related tests and tests for health purposes. This raises further concerns about whether the same rules should be applied for regulating informed consent in both types of gene sequencing applications. In countries where regulatory measures for DTC genetic testing services are comparatively loose, they generally have strict requirements of informed consent for health-related genetic testing. For example, in the United States, the Presidential Commission for the Study of Bioethical Issues stated that if the genetic testing is prescribed due to clinical purposes, physicians have to present informed consent to patients due to their fiduciary duties to patients (Niemiec et al., 2016). In China, although the new *Regulation on Human Genetic Resources* specifies that informed consent should be obtained from providers of human genetic resources before collecting and using their genetic information, the purposes of the *Regulation* are mainly based on the concerns of safeguarding public health, national security, and social public interests, rather than patients or consumers' health rights and interests in the context of genetic testing. Given the lack of regulation in China that can guide and monitor informed consent procedures for both clinical and non-health-related genetic testing, consumer rights for health and information are left without adequate protection.

Compared with the number of informed consent forms provided, more Chinese DTC genetic testing companies have addressed issues related to consumer privacy protection. For example, 37 websites offered a link to a privacy policy. This finding is consistent with existing research that privacy concerns have been increasingly addressed by international DTC genetic testing providers (Webborn et al., 2015). That being said, in the Chinese DTC genetic testing market, more than half of the websites we analyzed did not offer a legal statement on consumer privacy protection. With regards to the privacy policies offered by the DTC genetic testing providers, the majority were short and incomplete, which did not cover the content required by the *Personal Information Security Specification* on privacy policy. In particular, 3 months after the *Regulation on Human Genetic Resources* took effect, we did not identify a significant change in the DTC genetic testing industry for improving the practice of informed consent and consumer privacy to comply with the new legal requirements. Only one genetic testing provider had given an update to its informed consent form on its website, in that they offered a link to an informed consent form and added one short sentence for the protection of consumers' privacy in the form. This illustrates that, without an established comprehensive personal data protection law, both the *Cybersecurity Law* and the regulatory measure on genetic resources management had very little influence on promoting and advancing the protection of privacy and implementation of informed consent in DTC genetic testing services.

Although several relevant laws, such as the *Consumer Protection Law*, *Cybersecurity Law*, and the *Regulation on Human Genetic Resources* are applicable to regulating the genetic testing market, there is no special regulatory regime covering important issues associated with the implementation of DTC genetic testing services, e.g., informed consent, privacy protection, and transparency about how consumer genetic data is used, collected, and shared. Thus, the current Chinese DTC genetic testing market operates as a self-regulated mechanism. Moreover, "soft laws," such as best practices and a code of conduct, are also missing in the regulations of DTC genetic testing. We did not identify any established voluntary best practices guidelines for genetic testing services (Park et al., 2019). BGI, the biggest player in the Chinese gene sequencing market, leveraged its role in the field by organizing a focused group meeting with other genetic testing companies for the purpose of developing a group standard for genetic testing reports. As an outcome of the discussion, the *Specification for Gene Detection and Reporting of Clinical Monogenous Hereditary Diseases* was issued in 2018, becoming the first practical standard in the genetic testing industry (Hui et al., 2018). However, as BGI stated in the article that every process of the genetic testing service requires corresponding standards, and the *Specification* for clinical genetic testing report alone is far from adequate (Hui et al., 2018).

Without a sufficient and effective regulatory framework for DTC genetic testing services, consumers may face increased risks of losing control of their genetic information and privacy breaches. Specifically, we found that many companies failed to provide meaningful information to consumers concerning the security of genetic data and how the data will be used with a third party. For example, some companies granted consumers the right to authorize the use of their data, but if thorough informed consent procedures are not provided, consumers are unlikely to know for what purposes their genetic data will be used and how their data will be handled by a third party (Tomlinson et al., 2016). This is especially problematic if there is no definition of "third party," as well as details regarding how the genetic testing company will safeguard privacy when data is transferred to a third party. As a result, the risk of consumer data leakage is very high. Moreover, many companies did not state clearly what their liability would be in the event of a data security breach. In general, the testing companies kept silent in regards to their responsibilities in the event of a privacy breach, though some were explicit in releasing themselves from any liability in cases when breach incidents were caused by events out of their control (Hazel and Slobogin, 2018). Given these concerns, consumers should be careful and diligent when choosing genetic testing services and products (Badalato et al., 2017).

Our research indicates that the United States *HIPPA* legislation was mentioned by several Chinese DTC providers. These companies highlighted their efforts to protect consumer privacy and data security by strictly complying with the legal requirements detailed in *HIPPA*. In terms of anti-genetic discrimination in the health insurance realm, the China Insurance Regulatory Commission (the former entity of the China Banking and Insurance Regulatory Commission) demonstrated the restrictions on the use of genetic testing results in health insurance in the *Measures for the Management of Health Insurance (draft for comments)* – a 2017 version for the update of *Measures for the Management of* Health Insurance 2006 (China Insurance Regulatory Commission, 2017). Articles 11 and 16 of the 2017 *Measures* require that insurance companies protect the privacy and confidentiality of policy holders, the insured, and the beneficiaries, and that insurance companies should not add premiums to policy holders based on their genetic testing data and genetic information, other than the family genetic history of a disease. Moreover, according to Article 36 of the 2017 *Measures*, insurance companies should not require policy holders or the insured to take genetic tests, and that the genetic testing results are prohibited from being used as a condition for the verification of insurance. As the 2017 *Measures* have not been passed and entered into effect, further research is needed to examine and review the impact of the new regulations on the protection of private health information within the context of DTC genetic testing.

In addition, our research indicates that social media has played an increasingly important role in promoting genetic testing services and products. On the one hand, social media, as previous studies indicated, could be a useful tool to increase patients' knowledge of genetic testing and risk assessment for certain types of cancers (Attai et al., 2015). A recent investigation by Roberts et al. (2019) on the public reactions on Twitter to the government's authorization of DTC genetic testing for BRCA1/2 variants associated with breast cancer corroborated the substantial impacts of social media in this regard. Their research revealed the potentials of social media to become the main platform for disseminating and exchanging information about genetic research and technology, as well as a powerful medium for consumer testimonials (Roberts et al., 2019). In this regard, social media platforms could be used to raise public awareness of the inadequacies of privacy measures taken by gene testing providers. On the other hand, a large number of studies on social media's role in promoting new technology has indicated that information about new biotechnology shared through social medial was usually unbalanced and misleading (Galata et al., 2014). Although we did not analyze what content had been promoted on WeChat platforms about genetic testing, we suggest regulatory agencies focus their attention on the legal and ethical issues associated with using social media for the promotion of genetic testing services and products. Given these concerns, further studies on the role of social media in DTC genetic testing services and products are needed.

CONCLUSION

Our study indicates that DTC genetic testing has become an emerging market in China. Eighty-three Chinese companies were identified as promoting genetic testing products directly to consumers. The existing applicable regulations on genetic testing are mainly focused on human genetic resource security and protection, and no special legislation has been developed to regulate DTC genetic testing offerings. Without an established legal regime, the availability for informed consent forms and policies for consumer data and privacy protection within the industry are self-imposed by DTC genetic testing companies. Moreover, the industry has not established any best practices guidelines for implementing DTC genetic testing services. As a result, the current DTC genetic testing business is running in a regulatory vacuum and is governed by a self-regulation mechanism. Our study indicates that the limits of this self-regulation model is obvious. Informed consent forms were generally not provided by DTC genetic testing companies, and a privacy policy was only available on less than half of all providers' websites we examined. For the majority of DTC genetic testing companies, consumers' autonomy for purchasing genetic testing is unable to be guaranteed, and there is a lack of transparency about how consumer genetic information is used and shared. As a result, consumers are left without adequate protection. Their genetic information might be illegally used or shared to a third party without their permission. We, therefore, urge that adequate and effective regulatory oversight over DTC genetic testing offerings should be developed. In particular, a clear and sufficient informed consent form and privacy policy should be provided on all DTC genetic testing providers' websites (Hendricks-Sturrup and Lu, 2019). Moreover, to meet the increased requirements of data protection and the demands of data sharing, regulations should be developed to render a legitimized systematic approach to the collection, use, and sharing of consumer genetic databases.

As the industry keeps evolving, some challenging issues associated with the provision of DTC genetic testing require further studies. For example, social media has been frequently used by DTC genetic testing companies as an alternative way to promote genetic testing services. The involvement of social media could bring opportunities for raising public awareness on potential privacy risks with DTC genetic testing. It may also trigger regulatory challenges in supervising the dissemination of truthful and balanced information on genetic testing via social media platforms. Additionally, some DTC genetic testing companies have distinguished consumers' general health information from genetic information and used different methods to safeguard data security. This raises the question of whether different information protection rules should be developed and applied to different types of consumer health data. Given all these potential challenges and the growing industry, the development of regulations on genetic testing call for interdisciplinary perspectives, and it is essential to examine periodically the regulatory framework on genetic testing services.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

LD designed the study. MW collected and analyzed the data. LD drafted the manuscript. MW made revisions to the manuscript.

FUNDING

This study was supported by the Faculty of Law, University of Macau under grant no. MYRG2018-00074-FLL, and the Ministry of Science and Technology of the People's Republic of China under grant Research on Precision Medicine Ethics, Policy, and Legal Framework (grant no. 2017YFC0910100).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00416/full#supplementary-material

REFERENCES

Attai, D. J., Cowher, M. S., Al-Hamadani, M., Schoger, J. M., Staley, A. C., and Landercasper, J. (2015). Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. *J. Med. Internet Res.* 17:e188. doi: 10.2196/jmir.4721

Badalato, L., Kalokairinou, L., and Borry, P. (2017). Third party interpretation of raw genetic data: an ethical exploration. *Eur. J. Hum. Genet.* 25, 1189–1194. doi: 10.1038/ejhg.2017.126

Bal, B. S., and Brenner, L. H. (2015). Medicolegal sidebar: informed consent in the information age. *Clin. Orthop. Relat. Res.* 473, 2757–2761. doi: 10.1007/s11999-015-4440-5

Caulfield, T., and McGuire, A. L. (2012). Direct-to-consumer genetic testing: perceptions, problems, and policy responses. *Ann. Rev. Med.* 63, 23–33. doi: 10.1146/annurev-med-062110-123753

Chen, Y., and Song, L. (2018). China: concurring regulation of cross-border genomic data sharing for statist control and individual protection. *Hum. Genet.* 137, 605–615. doi: 10.1007/s00439-018-1903-2

China General Administration of Food and Drug Administration, and State Health Planning Commission. (2014). *Notice on Strengthening the Management of Products and Technologies Related to Clinical Use of Gene Sequencing* [Online]. Available: http://www.nmpa.gov.cn/WS04/CL2197/324799.html (accessed February 20 2020).

China Insurance Regulatory Commission (2017). *Measures for the Management of Health Insurance (draft for comments) [Online]*. Beijing: China Insurance Regulatory Commission.

China State Council (2019). *Regulation of the People's Republic of China on Human Genetic Resources Management [Online]*. Beijing: China State Council.

China's National Information Security Standardization Technical Committee (2017). *Information Security Technology - Personal Information Security Specification [Online]*. Available: https://www.tc260.org.cn/upload/2018-01-24/1516799764389090333.pdf (accessed February 20 2020).

Covolo, L., Rubinelli, S., Ceretti, E., and Gelatti, U. (2015). Internet-based direct-to-consumer genetic testing: a systematic review. *J. Med. Internet Res.* 17, e279. doi: 10.2196/jmir.4378

Deng, Z., and Liu, S. (2017). Understanding consumer health information-seeking behavior from the perspective of the risk perception attitude framework and social support in mobile social media websites. *Int. J. Med. Inform.* 105, 98–109. doi: 10.1016/j.ijmedinf.2017.05.014

Feng, Y. (2019). The future of China's personal data protection law: challenges and prospects. *Asia Pacific Law Rev.* 27, 62–82. doi: 10.1080/10192557.2019.1646015

Galata, L., Karantininis, K., and Hess, S. (2014). "Cross-atlantic differences in biotechnology and GMOs: a media content analysis," in *Agricultural Cooperative Management and Policy: New Robust, Reliable and Coherent Modelling Tools*, eds C. Zopounidis, N. Kalogeras, K. Mattas, G. van Dijk, and G. Baourakis (Cham: Springer International Publishing), 299–314. doi: 10.1007/978-3-319-06635-6_16

Global Market Insights (2018). *Genetic Testing Market worth over $22 Billion By 2024: Global Market Insights, Inc. [Online]*. Maharashtra: Global Market Insights.

Ha, K. O. (2019). *A Reporter Took DNA Tests in the U.S. and China. The Results Left Her Worried [Online]. FORTUNE*. Available: https://fortune.com/2019/11/30/dna-tests-us-china-results/ (accessed February 20 2020).

Hall, J. A., Gertz, R., Amato, J., and Pagliari, C. (2017). Transparency of genetic testing services for 'health, wellness and lifestyle': analysis of online prepurchase information for UK consumers. *Eur. J. Hum. Genet.* 25, 908–917. doi: 10.1038/ejhg.2017.75

Hazel, J., and Slobogin, C. (2018). Who knows what, and when?: a survey of the privacy policies proffered by U.S. direct-to-consumer genetic testing companies. *Cornell J. Law Public Policy* 28, 35–66.

Hendricks-Sturrup, R. M., and Lu, C. Y. (2019). Direct-to-consumer genetic testing data privacy: key concerns and recommendations based on consumer perspectives. *J. Pers. Med.* 9:25. doi: 10.3390/jpm9020025

Hogarth, S., Javitt, G., and Melzer, D. (2008). The current landscape for direct-to-consumer genetic testing: legal, ethical, and policy issues. *Ann. Rev. Genom. Hum. Genet.* 9, 161–182. doi: 10.1146/annurev.genom.9.081307.164319

Howard, H. C., Knoppers, B. M., and Borry, P. (2010). Blurring lines. *EMBO reports* 11, 579–582. doi: 10.1038/embor.2010.105

Huang, J. (2019). Chinese private international law and online data protection. *J. Private Int. Law* 15, 186–209. doi: 10.1080/17441048.2019.1599771

Hui, H., Shen, Y. P., Gu, W. H., Wang, W., Wang, Y. M., Qi, M., et al. (2018). Discussion on the standard of clinical genetic testing report and the consensus of gene testing industry. *China J. Med. Genet.* 35, 1–8.

Jin, Y., Zhang, L., Ning, B., Hong, H., Xiao, W., Tong, W., et al. (2018). Application of genome analysis strategies in the clinical testing for pediatric diseases. *Pediatr. Investig.* 2, 72–81. doi: 10.1002/ped4.12044

Knoppers, B. M., Zawati, M. H., and Senecal, K. (2015). Return of genetic testing results in the era of whole-genome sequencing. *Nat. Rev. Genet.* 16, 553–559. doi: 10.1038/nrg3960

Lachance, C. R., Erby, L. A. H., Ford, B. M., Allen, V. C., and Kaphingst, K. A. (2010). Informational content, literacy demands, and usability of websites offering health-related genetic tests directly to consumers. *Genet. Med.* 12, 304–312. doi: 10.1097/gim.0b013e3181dbd8b2

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159. doi: 10.2307/2529310

Lenore, M., Cartwright, E., and Hardon, A. (2016). *The Routledge Handbook of Medical Anthropology*. Abingdon: Routledge.

Luo, Z., Zayts, O., and Shipman, H. (2020). "His story is truly vivid…": the role of narratives of vicarious experience in commodification and marketisation of genetic testing in Chinese social media. *J. Pragmat* 155, 111–122. doi: 10.1016/j.pragma.2019.10.009

Mostert, M., Bredenoord, A. L., Biesaart, M. C. I. H., and Delden, J. J. M. V. (2016). Big data in medical research and EU data protection law: challenges to the consent or anonymise approach. *Eur. J. Hum. Genet.* 24, 956–960. doi: 10.1038/ejhg.2015.239

National Intellectual Property Administration (2019). *Regulations on Management of China's Human Genetic Resources to be in Effect in July* [Online]. Beijing: National Intellectual Property Administration.

Niemiec, E., Borry, P., Pinxten, W., and Howard, H. C. (2016). Content analysis of informed consent for whole genome sequencing offered by direct-to-consumer genetic testing companies. *Hum. Mutation* 37, 1248–1256. doi: 10.1002/humu.23122

Niemiec, E., Kalokairinou, L., and Howard, H. C. (2017). Current ethical and legal issues in health-related direct-to-consumer genetic testing. *Pers. Med.* 14, 433–445. doi: 10.2217/pme-2017-0029

Overmaat, T., Heitner, J. A., Mischel, F. R., and Heitner, M. (2018). Consumer-facing genetic testing in China: a status report. *Lancet* 392:S50. doi: 10.1016/s0140-6736(18)32679-5

Park, J. Y., Risher, M. T., Caulfield, T., Baudhuin, L. M., and Schwab, A. P. (2019). Privacy in direct-to-consumer genetic testing. *Clin. Chem.* 65, 612–617. doi: 10.1373/clinchem.2018.298935

Roberts, M. C., Allen, C. G., and Andersen, B. L. (2019). The FDA authorization of direct-to-consumer genetic testing for three BRCA1/2 pathogenic variants: a twitter analysis of the public's response. *JAMIA Open* 2, 411–415. doi: 10.1093/jamiaopen/ooz037

Sui, S., and Sleeboom-Faulkner, M. (2015). Commercial genetic testing and its governance in Chinese society. *Minerva* 53, 215–234. doi: 10.1007/s11024-015-9279-0

The Standing Committee of the National People's Congress (2013). *Law of the People's Republic of China on the Protection of Consumer Rights and Interests (2013 Amendement) [Online]*. Beijing: The Standing Committee of the National People's Congress.

Tomlinson, A. N., Skinner, D., Perry, D. L., Scollon, S. R., Roche, M. I., and Bernhardt, B. A. (2016). "Not Tied Up Neatly with a Bow": Professionals'. challenging cases in informed consent for genomic sequencing. *J. Genet. Couns.* 25, 62–72. doi: 10.1007/s10897-015-9842-8

Webborn, N., Williams, A., McNamee, M., Bouchard, C., Pitsiladis, Y., Ahmetov, I., et al. (2015). Direct-to-consumer genetic testing for predicting sports performance and talent identification: consensus statement. *Br. J. Sports Med.* 49, 1486–1491. doi: 10.1136/bjsports-2015-095343

Wu, Y., Howarth, M., Zhou, C., Ji, X., Ou, J., and Li, X. (2019). Reporting of ethical considerations in clinical trials in Chinese nursing journals. *Nurs. Ethics* 26, 973–983. doi: 10.1177/0969733017722191

XinHuaNet (2019). *China Publishes Regulation on Management of Human Genetic Resources [Online]*. Beijing: XinHuaNet.

YiOu ZhiKu (2018). *2018 中国消费级基因检测市场研究报告 2018 China Consumer Genetic Testing Market Research Report (Translated by the authors)* [Online]. Available: https://www.iyiou.com/intelligence/report605.html (accessed February 20 2020).

Zhao, X., Wang, P., Tao, X., and Zhong, N. (2013). Genetic services and testing in China. *J. Community Genet.* 4, 379–390. doi: 10.1007/s12687-013-0144-2

Conflict of Interest: The 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.

Copyright © 2020 Du and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fgene.2020.00303

Genomic Sequencing Capacity, Data Retention, and Personal Access to Raw Data in Europe

Shaman Narayanasamy[1]*, *Varvara Markina*[1], *Adrian Thorogood*[2], *Adriana Blazkova*[1,3], *Mahsa Shabani*[4], *Bartha M. Knoppers*[2], *Barbara Prainsack*[5,6] *and Robert Koesters*[1]

[1] Megeno S.A., Esch-sur-Alzette, Luxembourg, [2] Centre of Genomics and Policy, McGill University, Montreal, QC, Canada, [3] Faculty of Language and Literature, Humanities, Arts and Education, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [4] Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium, [5] Department of Political Science, University of Vienna, Vienna, Austria, [6] Department of Global Health & Social Medicine, King's College London, London, United Kingdom

Whole genome/exome sequencing (WGS/WES) has become widely adopted in research and, more recently, in clinical settings. Many hope that the information obtained from the interpretation of these data will have medical benefits for patients and—in some cases—also their biological relatives. Because of the manifold possibilities to reuse genomic data, enabling sequenced individuals to access their own raw (uninterpreted) genomic data is a highly debated issue. This paper reports some of the first empirical findings on personal genome access policies and practices. We interviewed 39 respondents, working at 33 institutions in 21 countries across Europe. These sequencing institutions generate massive amounts of WGS/WES data and represent varying organisational structures and operational models. Taken together, in total, these institutions have sequenced ∼317,259 genomes and exomes to date. Most of the sequencing institutions reported that they are able to store raw genomic data in compliance with various national regulations, although there was a lack of standardisation of storage formats. Interviewees from 12 of the 33 institutions included in our study reported that they had received requests for personal access to raw genomic data from sequenced individuals. In the absence of policies on how to process such requests, these were decided on an *ad hoc* basis; in the end, at least 28 requests were granted, while there were no reports of requests being rejected. Given the rights, interests, and liabilities at stake, it is essential that sequencing institutions adopt clear policies and processes for raw genomic data retention and personal access.

Keywords: NGS, ELSI, policies, procedures, patient rights, research participant rights, raw, GDPR

OPEN ACCESS

Edited by:
Go Yoshizawa,
OsloMet – Oslo Metropolitan University, Norway

Reviewed by:
Shiri Shkedi-Rafid,
Hadassah Medical Center, Israel
Sarah Nelson,
University of Washington, United States

****Correspondence:***
Shaman Narayanasamy
shaman.narayanasamy@megeno.com

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *29 November 2019*
Accepted: *13 March 2020*
Published: *06 May 2020*

Citation:
Narayanasamy S, Markina V, Thorogood A, Blazkova A, Shabani M, Knoppers BM, Prainsack B and Koesters R (2020) Genomic Sequencing Capacity, Data Retention, and Personal Access to Raw Data in Europe. Front. Genet. 11:303. doi: 10.3389/fgene.2020.00303

INTRODUCTION

Whole genome sequencing (WGS) and whole exome sequencing (WES) have become widely adopted in research and, more recently, in clinical practice (Birney et al., 2017; Birney, 2019). The generated raw genomic data (i.e., WGS/WES data) include vast amounts of information of potential importance to an individual's current and future health, with implications for family members, if analytic and interpretive hurdles can be overcome. The wide availability of genomic data also offers opportunities for reuse for additional clinical, health, research, or recreational purposes. People requesting access to their own raw data, however, raises a number of legal, ethical, and practical questions. Legally, patients in many countries have a right to access their health record (Thorogood et al., 2018). Individual access rights are also being strengthened under data privacy laws. For example, the European Union General Data Protection Regulation (GDPR; European Parliament and Council, 2016), in force since May 2018, stipulates a general right of data subjects to access their personal data. GDPR leaves it to member states to decide if and how this right applies in research contexts and to raw genomic data specifically. Given the broad translational spectrum in genomics, however, it can be difficult to clearly distinguish clinical and research contexts (Schickhardt et al., 2020). Another legal uncertainty is whether or not access rights extend to raw sequence data, though broad definitions of personal (health) data support this interpretation (Thorogood et al., 2018). There are ethical arguments for and against personal genome access.

On the one hand, some argue that research participants and patients (collectively referred to as "sequenced individuals") have a moral right to access their own raw (uninterpreted) genomic data in both clinical and research contexts as something that fundamentally belongs to them (Nelson, 2016; Schickhardt et al., 2020). Access can also potentially empower sequenced individuals to direct the analysis and sharing of their own data, potentially improving their own knowledge and health, as well as accelerating research and innovation (Lunshof et al., 2014; den Dunnen, 2015; Middleton et al., 2015; Wright et al., 2017, 2019; Shabani et al., 2018; Thorogood et al., 2018). Providing data may also be a way of encouraging and engaging participants in research (Middleton et al., 2015). On the other hand, some express concerns that providing personal access is at best pointless and at worse harmful for individuals and burdensome for providers and health systems (Bredenoord et al., 2011). Individuals may not be able to do anything with the genomic data, or they may misinterpret the data. This is especially true if the data are of uncertain quality, as is often the case in research contexts. They may share it with unscrupulous researchers or unregulated service providers, exposing them to further misinterpretations and privacy harms (Guerrini et al., 2019). Of course, some of these risks can be mitigated through clear policies, oversight, education, and access to counseling services (Shabani et al., 2018; Schickhardt et al., 2020). However, this, in turn, raises practical resource questions, especially for research projects. Moreover, policies, processes, and infrastructure are required to sustainably manage and transfer large raw genomic data formats (Middleton et al., 2015; Wright et al., 2017).

To better understand current practices of personal access to raw genomic data by sequenced individuals, we conducted interviews with genomics professionals working in institutions within the EU/EEA that routinely perform WGS/WES of human individuals on a large scale (i.e., "sequencing institutions"). Sequencing institutions can be viewed as gatekeepers or enablers for sequenced individuals in accessing their personal raw genomic data. Furthermore, owing to their geographical location and/or the data they use, these institutions are expected to be directly impacted by the GDPR, which makes their practices particularly insightful and timely in light of the evolving regulatory landscape. For uninitiated readers, the following primer describes the impact of the GDPR on health research: Dove (2018). This study is the first to provide empirical insights into the policies, practices, and perspectives within sequencing institutions pertaining to individual access to raw genomic data. We also consider technical aspects of sequencing capacity and data retention practices, as these variables determine the overall availability of data. Our findings provide valuable empirical observations that can inform legal and ethical debates over personal genomic access, and indicate practical and technical solutions for sequencing institutions seeking to respond to such requests.

MATERIALS AND METHODS

Interview Guide

A semi-structured interview guide was prepared consisting of questions pertaining to practices around WES/WGS, with a particular focus on genomic data retention and provision of access to sequenced individuals. The interview guide included both closed-ended questions (aimed primarily at describing the profile and practices of the sequencing centres) and open-ended questions, intended to gauge respondents' attitudes toward a specific issue. To ensure a clear and intuitive structure of interviews, we divided the interview guide into five distinct sections ("modules") addressing the following topics: (i) organisational structure of the institution; (ii) sequencing throughput and capacity; (iii) data management and storage capacity; (iv) data retention policies and access policies for sequenced individuals (to their own data); and (v) sequencing centres' experiences with receiving requests from individuals to access their raw genomic data.

The draft versions of the interview guide underwent multiple rounds of internal review and refinement. The final round of refinement was carried out upon receiving feedback from the first 10 interviews within the study. The interview guide is available as **Supplementary File 1**.

Identification of Sequencing Institutions

This interview study specifically targeted institutions located in the EU/EEA that were member states of the EU/EEA, and that generate, process, and/or manage human WGS/WES data for research and/or clinical purposes. As such, we refer to such institutions as "sequencing institutions."

In order to identify sequencing institutions, we used all of the following methods: (i) web searches, (ii) prior knowledge, (iii) peer recommendations, (iv) media articles or announcements, and (v) personal relationships. Our research strategy identified 83 sequencing institutions from 23 member states across the EU/EEA region. Sequencing institutions were approached individually with the request to participate in the study.

Participant Recruitment

Recruitment of participants started in May 2018 and continued in parallel to interviews with early respondents. A total of 64 sequencing institutions, out of a possible 83, were invited to participate in the study via email, of which 33 eventually agreed. Interviews took place between June 2018 and April 2019. Individual respondents who participated in our study were all affiliated with sequencing institutions, within which they held various positions and responsibilities. Three interviews included multiple respondents, with a maximum of five respondents from the same institution being in the same interview. This brings the total number of respondents to 39 (**Supplementary File 2**: **Table S1**). Recruitment ended when it became clear from the interviews that data saturation had been reached and no new insights were emerging from additional interviews.

Pre-interview Communication

Respondents (representing sequencing institutions) who agreed to take part in the interview study were provided with information about the purpose of the interview, the thematic areas of focus, and the methods (**Supplementary File 3**). Potential respondents were also provided with a confidentiality statement explaining how the collected data would be treated (**Supplementary File 4**). Additional measures for data privacy, confidentiality, and security are detailed within **Supplementary File 5**.

Semi-Structured Interviews and Recordings

Interviews were carried out either with Zoom video conferencing software[1] or in person. At the beginning of the interviews, the interviewers briefly described the interview process and addressed any issues of confidentiality and privacy of the respondent (i.e., **Supplementary Files 4, 5**). The conversation then proceeded with personal introductions by the interviewers and respondents after which the interviews progressed according to the interview guide (**Supplementary File 1**). At the end of the interviews, the respondents were asked to provide concluding remarks or suggestions, if any, for further improving the interview guide.

Transcription and Review

Interviews were transcribed using otter.ai[2], an automated, artificial intelligence-based transcription software. The automatically generated transcripts were reviewed for accuracy and manually edited to correct any discrepancies with the corresponding audio files. Reviewing and editing of transcripts were carried out by VM/AB/SN and verified by RK for validation purposes.

The process of transcription was combined with the generation of interview summary documents called "review sheets" (**Supplementary File 6**). For each review sheet, a set of the most relevant quantitative and qualitative information, deemed to best reflect the respondent's views, were selected from the corresponding interview transcript. The quotes were accompanied by a concise written summary of the interview. Subsequently, the review sheets were sent to the pertinent respondents who had an opportunity to comment on the document and suggest revisions if required.

Collection and Analysis of Informed Consent Forms From Sequencing Institutions

In addition to the data collected for the primary research (i.e., semi-structured interviews), samples of informed consent forms that sequencing institutions use to consent sequenced individuals were requested via email (from the respondent) or accessed online, depending on their availability. We only collected informed consent forms from the sequencing institutions involved in the study.

Data Curation and Analyses

The process of data analysis was divided into two parts, corresponding to the nature of the data being analysed (quantitative and qualitative data analyses).

Quantitative Data Analysis and Visualisation

Quantitative data analysis was applied to questions requiring numerical responses (e.g., no. of people sequenced), binary answers (yes/no), or categorical variables (e.g., purchase year of first Novaseq) from the interview transcripts that were tabulated (**Supplementary File 2: Tables S1–S6**). The conversion of the relevant information into the aforementioned format was performed by AB and VM. The process was independently repeated and refined by SN. Quantitative data analysis was carried out using Google Sheets, as part of Google Suite, and R statistical package (R Core Team, 2013). Data visualisation was performed using ggplot2 graphical package (Wickham, 2016). Diagrams, drawings, and schemes were generated using either Google Slides (as part of Google Suite) or Adobe Illustrator. All visualisations were refined using Adobe illustrator.

Qualitative Data Analysis

To analyse qualitative data collected through this study, we employed deductive content analysis. In this approach, themes or common content categories are pre-determined before data analysis is undertaken, as opposed to being identified in the course of data analysis (Graneheim and Lundman, 2004). Qualitative data were organised under five broad themes, which reflected the overall structure of the interview guide. The process of organising the qualitative data under these themes was undertaken by SN, AB, and VM and subsequently reviewed and validated by RK.

Interview transcripts were carefully read to identify quotes referring to one or more of the predetermined themes. The relevant quotes were subsequently placed under the most suitable theme. Quotes that bore relevance to more than one theme were divided into multiple parts and the resultant sub-quotes were placed under the suitable themes. Selection and categorisation of the relevant quotes was performed by AB and VM. The process

[1] https://zoom.us/

[2] https://otter.ai/

was independently validated by SN. Discrepancies in categorising quotes were routinely discussed and resolved.

The analysis of the informed consent forms focused exclusively on individual access to genomic data and genomic data retention policies.

RESULTS

In total, the study included 33 interviews, conducted with 39 respondents within sequencing institutions, operating in 21 EU/EEA member states (**Figure 1**). They relate to more than 300,000 individuals, who underwent WGS/WES between the early 1990s to the first half of 2019. We explored current practices and policies of data management and personal raw genomic data access for sequenced individuals within these different institutions.

In line with our deductive content analysis approach, study findings were organised into the following five sections: (i) organisational structure and operational models; (ii) actual vs. potential sequencing capacity; (iii) genomic data management practices and policies; (iv) data access practices and policies; and (v) future outlook.

Respondent Profiles, Organisational Structure, and Operational Models

The first part of the interviews served as an introduction to the respondent. Overall, the respondents held various positions and responsibilities within those sequencing institutions, including technical, academic, administrative, clinical, and management (**Figure 1**).

Next, we sought to obtain a better understanding of the organisational structure and the operations of sequencing institutions, as those factors may influence processes and policies for personal access to raw genomic data by sequenced individuals. In terms of organisational structure, the study included mostly (24, ∼73%) public organisations, followed by six (∼18%) not-for-profit private organisations, two (6%) commercial organisations, and one consortium. Additionally, 19 (∼58%) of the participating institutions performed sequencing for both research and clinical purposes, while 11 (∼33%) and 3 (∼9%) institutions focused exclusively on research or clinical sequencing, respectively (**Figure 1**). The organisations included in this study varied considerably in their size and number of personnel, with the largest and the smallest institution housing approximately 3000 and 10 staff members, respectively (average ∼450). Furthermore, a total of ∼460 personnel (average ∼17, max. ∼80, min. 3) within those organisations were dedicated towards operating sequencing platforms and data analyses and management (**Supplementary File 2: Table S2**).

We then asked the respondents to elaborate on various aspects of their operations, including (but not limited to) their clientele, main activities (e.g., sequencing, or data processing), their institutional arrangements (e.g., university hospital, private laboratory), and whether they were outsourcing specific tasks or processes, related to human genome sequencing.

We found that the sequencing institutions we covered typically acted as service providers to healthcare and/or research institutions, but delivered their services in different ways. In this respect, we delineated four different "operational models" to further classify the sequencing institutions as follows: (i) dedicated, (ii) open, (iii) integrated, and (iv) outsourced. **Figure 1** provides an illustration and descriptions of the various operational models and their adoption among sequencing institutions included in our study. We found that most sequencing institutions provided services to an exclusive set of clients (i.e., dedicated), primarily made up of sequencing laboratories affiliated with a specific clinical or research institution, and performing sequencing services exclusively for those affiliated institutions. Affiliations are determined either through formal partnerships or based on geographical regions (i.e., regional). Other institutions performed sequencing for any internal or external clients as a standard service (i.e., open model). There were also "integrated" sequencing institutions that were physically located within the premises of a larger organisation (e.g., a university, hospital, consortia/network). Finally, there were 11 sequencing institutions that outsourced their sequencing (i.e., outsourced model) and focused entirely on data analysis and interpretation.

We found that 15 (45%) of the sequencing institutions combined at least two operational models, with the most prominent combination being the dedicated and integrated models (i.e., 8, 24%). This implies that most of the integrated sequencing institutions were dedicated to their "parent" institution (e.g., hospital, university, consortium, network). The second most frequent combination (i.e., 7, 21%) was the coupling of the outsourced model with any of the other operation models (**Figure 1**). Additionally, we identified two sequencing institutions operating as data hubs that did not perform any in-house sequencing, but aggregated and processed sequencing data from multiple outsourced sequencing providers.

Actual vs. Potential Sequencing Capacity

We asked respondents about their institutions' potential and actual (i.e., throughput) sequencing capacities. The potential sequencing capacity is defined as the theoretical maximum amount of in-house sequencing (in gigabases) a given sequencing institution can perform per annum, if they were to operate at full capacity. These numbers were calculated on the basis of publicly available information concerning the sequencing platforms used by the participating institutions (**Supplementary File 2: Table S3**). However, we lowered the estimates to 70% of the maximum annual capacity in order to derive more realistic assumptions, as a respondent duly noted that one should consider the capacity of a given facility, rather than the capacity of the sequencing platforms.

> *"[...] there are some practical problems like running [the sequencers] 24/7 and some working regulations. And that's why we are [not] operating [...]at the full capacity of the sequencers, but [rather] at full capacity of the facility [...]"*
>
> *Respondent 1*

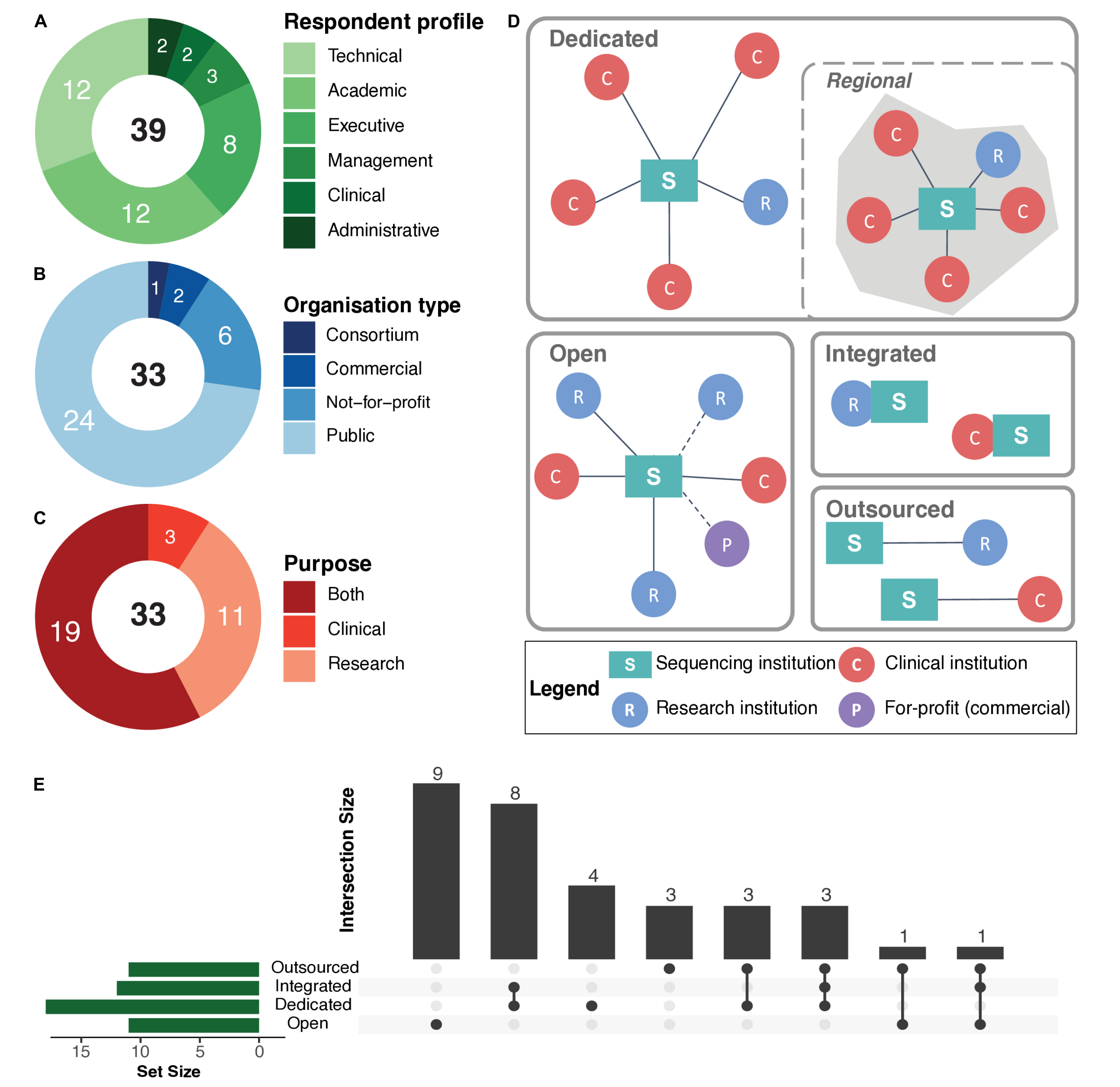


FIGURE 1 | Organisation structure and operations of sequencing institutions. **(A)** Summary of respondent profiles. **(B)** Summary of organisation types. **(C)** Use cases or purpose of human whole genome (WGS) or whole exome (WES) sequencing. In **(A–C)**, numbers in black text at the centre of the rings represent the total, while the numbers in white text on the rings represent the exact numbers for a particular category. **(D)** The various operational models practiced by the institutions participating in the interview study, in terms of human whole genome and exome sequencing: "Dedicated"—operate sequencing platforms in-house to serve an exclusive set of clinical and research clients. In certain cases, such institutions serve clients within a given region/locality (i.e., "Regional"). "Open"—a standard service-oriented institution with in-house sequencing platforms. "Integrated"—institutions/departments with in-house sequencing platforms that are embedded within a research and/or clinical unit. "Outsourced"—institutions that perform sequencing with external providers. **(E)** UpSet plot (Conway et al., 2017) represents the operational models and corresponding number of sequencing institutions that utilise those models for WGS/WES. The horizontal green bars represent the total number of a given operational model. The bottom panel represents specific combinations (or intersections) of operational models. The vertical bars represent the number of sequencing institutions using those combinations of operational models. Detailed information available in **Supplementary File 2: Table S2**.

In terms of the trends and the scale of sequencing-centered activities within the participating institutions (**Figure 2**), collectively, the 33 institutions had sequenced ∼161,899 whole genomes and ∼155,360 whole exomes (317,259 samples in total) at the time of our interviews. The sequencing of the first WGS began in the early 1990s (by a large consortium), while the first WES was generated in 2008. The sequencing coverage (or depth) for genomes (WGS) ranged from 5× to 200× (mean = ∼51×, median = 30×) across the institutions; for exomes (WES), the sequencing coverage ranged from 10× to 300× (mean = ∼107×, median = 100×; **Supplementary File 2: Table S4**). The aggregated potential sequencing capacity of the institutions was approximately 1.9×10^7 gigabases per year, which translates into ∼198,000 WGS per year at 30 × coverage. Based on this information, we additionally estimated the potential future scale of long-term genomic data retention.

We then asked respondents to estimate the future sequencing capacity of their institutions to help us form a clearer view on trends in WGS/WES in EU/EEA. Respondents representing 17 different institutions indicated plans for expansion, while 15 of them expressed a clear intention to purchase additional state-of-the-art sequencing platforms. **Figure 2** shows the estimated combined historical and future sequencing capacity. Considering this input, we estimated that the total future potential capacity would increase to 3.0×10^7 gigabases per annum, equivalent to more than 300,000 WGS at 30 × coverage. Most respondents were unable to predict actual future sequencing capacity outside of funded research projects (**Supplementary File 2: Table S4**). It is important to note that the reported numbers and projections are solely based on estimates provided by the respondents and are not meant to serve as accurate measures. We note, however, that business decisions to purchase sequencing machines may be an indicator of perceived future demand.

Genomic Data Management

The massive output from sequencing institutions generates large amounts of data and thus creates a downstream challenge in data management. When specifically addressing raw genomic data access for sequenced individuals, there were two aspects that we were interested in: (i) what is available for access in terms of raw genomic data file formats and (ii) how long will they remain accessible. Those factors, however, were constrained by (i) the capacity of the data storage infrastructure and (ii) data retention policies, which could be dictated by either internal (institutional) policies or national or EU regulations. **Figure 3** shows the storage duration of raw genomic data file formats and various policies that govern retention of genomic data.

Practices

Discussions with respondents centred on the so-called raw file formats in the genomic data processing chain, which include (listed in order of production) the following: (i) BCL, (ii) FASTQ, (iii) BAM (including all subtypes), and (iv) VCF. It is important to note that those file formats may span up to 100 GB (for WGS data), while some of those formats may be redundant (e.g., BAM and FASTQ). As such, VCF and FASTQ formats cannot be converted back to the prior data format (BAM and BCL, respectively), potentially resulting in loss of data if those prior formats are deleted.

> *"That is the benefit of storing BCLs rather than FASTQs because it's an untouched data [. . .] when you do the demultiplexing of the BCL, and you might [. . .]. I mean, and then you [. . .] have to make a decision on how do you demultiplex? How do you get out the different reads from these BCLs, and that's a decision that, [. . .] if you make the wrong decision, it cannot go back."*
>
> Respondent 2

Most respondents (26, ∼79%) were committed to at least one of the raw genomic data formats for indefinite periods (**Figure 3**). Particularly, the most widely retained are FASTQ (26, ∼79%), followed by both BAM and VCF (25, ∼76%), while four (12%) institutions committed to retaining BCL files indefinitely (**Supplementary File 2: Table S5**). In most cases, sequencing institutions maintained their own data storage (including back-ups) and computing facilities. As such, our respondents reported that a total of 284,522 (134,202 WGS and 150,320 WES) raw genomic data sets were retained by the sequencing institutions, which represented ∼90% of the total number of samples sequenced to date (see the *Actual vs. potential sequencing capacity* section and **Supplementary File 2: Table S5**). Respondents explained that raw data formats were stored for future re-analysis and re-interpretation. When specifically asked about the sustainability of storing those files indefinitely, respondents were generally confident in their capability to manage the data in the near future (e.g., 5–10 years). This is further supported by the fact that only five institutions utilise or plan to utilise state-of-the-art genomic data compression technologies (e.g., CRAM, Hsi-Yang Fritz et al., 2011; Bonfield, 2014) to save on data storage costs (**Supplementary File 2: Table S5**). However, most respondents considered that indefinite storage of genomic data sets might be unsustainable long-term (e.g., >10 years).

Policies

The genomic data retention policies of sequencing institutions ranged from 3 months to 115 years, to indefinite storage (**Figure 3**). The lower end of this spectrum is typically represented by sequencing institutions that practice the service-oriented open operational model (see the *Respondent profiles, organisational structures, and operational models* section) and therefore enforce strict internal raw genomic data retention (**Supplementary File 2: Table S5**). Consequently, the ∼10% of those so-called "unretained" genomic data sets (**Supplementary File 2: Table S4**) stem from such institutions. In contrast, seven sequencing institutions assumed the responsibility of storing all the genomic data in compliance with national laws for clinical data, under the assumption that genomic data are considered as clinical data (**Supplementary File 2: Table S5**). Moreover, those sequencing institutions support clients from healthcare and research in managing their genomic data for the time being.

> *"[...] we never removed anything, but [...] in our agreements [guarantee two years of storage]. [We] are basically waiting for healthcare to establish [...] long-term data archiving solutions. And when those are in place, we will start moving the data there for long-term storage, for archiving. But [any data that we have] in our hands, we will [store] for two years. But [. . .] because our collaborative customer [is not] ready, [we] have said [it is] too much [of] value to destroy it [...] now, so, we keep it and if [it is] not that expensive to store on tape [. . .], we can [absorb] the cost."*
>
> Respondent 2

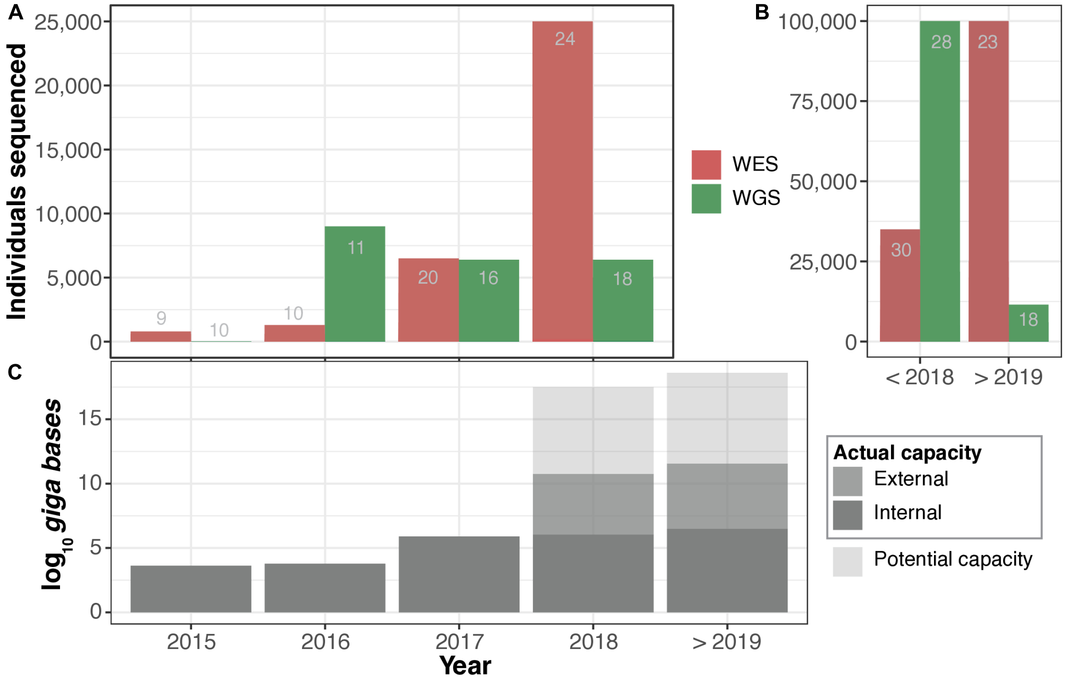


FIGURE 2 | Actual vs. potential sequencing capacity. **(A)** Number of whole exomes (WES) and whole genomes (WGS) sequenced per year. **(B)** Total number of WES and WGS sequenced to date. In **(A,B)**, the number of respondents that were able to estimate the number of WES and WGS sequenced individuals are indicated in grey text within the bars. **(C)** Annual predicted (70%) potential vs. actual sequencing capacity. Detailed information available in **Supplementary File 2: Tables S3, S4**.

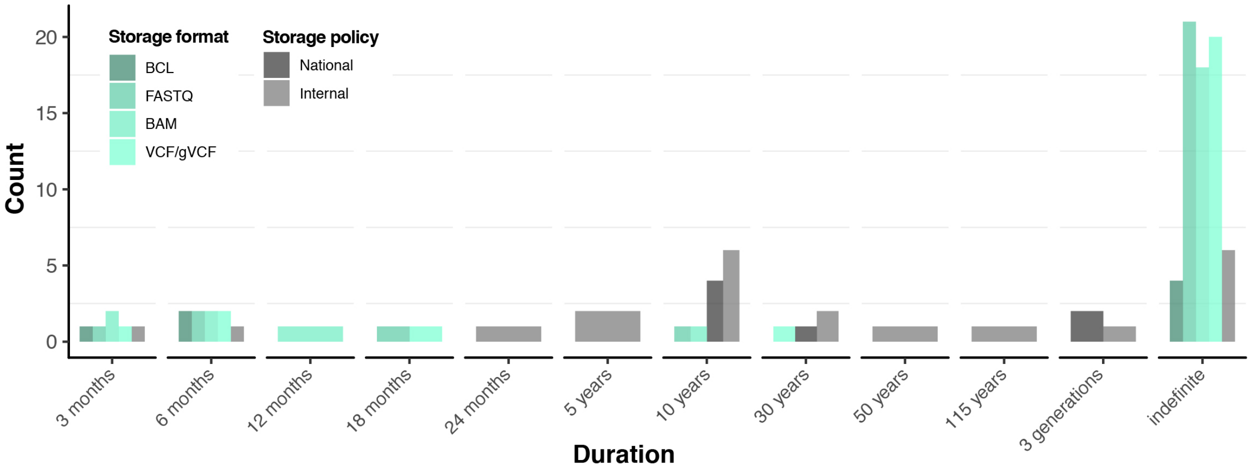


FIGURE 3 | Data retention practices and policies. Green scale represents storage practices in terms of stored file formats. Grey scale represents national and internal institutional data storage policies reported by the respondents. Detailed information available in **Supplementary File 2: Table S5**.

Thirteen respondents stated that their data retention policies are stated within their informed consent forms (**Supplementary File 2: Table S5**). Upon comparing the responses to the informed consent forms that we collected, we found that all (nine) of the informed consent forms broadly addressed data retention, with four clearly stating the duration of data retention.

Respondents were asked about the impact of the GDPR that had recently come into force in May 2018. The majority of respondents answered that necessary measures in relation to genomic data management had been in place even before the introduction of the GDPR, mostly due to existing stringent laws when dealing with personal genetic information, which includes genomic data. Only one respondent reported a change in data management strategy because of the GDPR specifically, which involved switching from the long-term storage of BCL files to FASTQ to comply with the "right to be forgotten" outlined by

GDPR. Compliance with this principle, our respondent said, was not feasible using BCL files.

> *"[...] one of the changes [we are] doing is; [switching] from BCL to FASTQs storage, because [it is easier] to remove the individuals, if they would request that."*
>
> *Respondent 2*

In summary, sequencing institutions established various measures/strategies to manage raw genomic data, in compliance with laws and regulations. The wide variation in data retention policies and practices (**Figure 3**) is surprising considering that most sequencing institutions face relatively similar technical challenges, organisational priorities, and presumably also relatively similar regulatory frameworks with regard to data retention and data protection.

Raw Genomic Data Access for Sequenced Individuals

In the final part of the interview, we asked the respondents if they had ever experienced cases of individuals seeking to access their raw genomic data. Accordingly, we documented at least 28 such cases (in 12 sequencing institutions within 10 countries) of sequenced individuals requesting and subsequently receiving access to their own genomic data (**Figure 4**). It is important to note that all sequencing institutions that received such requests ultimately granted access to the raw genomic data. We further asked those respondents who managed such cases to elaborate on how the process was carried out. We also asked all respondents about policies governing the personal raw genomic data access.

Practices

The most common raw file format provided to the individuals was VCF followed by BAM and FASTQ files (**Supplementary File 2: Table S6**). Two respondents mentioned that the sequenced individuals in question wanted their data to perform their own analyses. None of the other respondents knew the exact reasons why individuals had requested access to their own raw genomic data, but broadly speculated that those individuals were looking for second opinions.

> *"Well, it was actually a patient who wanted a second opinion on the data. And it was an individual, who [was] educated [in] bioinformatics, and wanted to have a look at the data [themselves] and have some second opinion about it [...]"*
>
> *Respondent 3*

Sequenced individuals and sequencing institutions are not in direct contact; therefore, access requests are relayed through an "intermediary contact," usually a healthcare professional, trial master, or principal investigator (**Table 1**). Additionally, the authorisation to grant access to a given sequenced individual appears to rest solely in the discretion of the aforementioned intermediary contact or, in some cases, is evaluated by a panel that may involve personnel from the sequencing institution, e.g., respondents themselves (**Table 1**). The sequencing institution (i.e., where the data resides) complies with the decision of the intermediary contact or panel and acts accordingly (**Figure 4**). In summary, those intermediary contacts may be viewed as gatekeepers for sequenced individuals to access their raw genomic data.

Requests were typically handled by institutions on a case-by-case basis, using *ad hoc* procedures. Only one institution confirmed a standardised internal procedure/process to comply with such requests. Most respondents reported handing out the data to the sequenced individual on an external hard drive, while a small number were able to provide it via download. Organisations employed measures such as pseudonymisation and encryption to ensure confidentiality and security.

Four cases of personal raw genomic data access requests occurred in sequencing institutions that were public organisations with dedicated and/or integrated sequencing platforms (**Figure 4** and section *Respondent profiles, organisational structures, and operational models*). The two commercial sequencing institutions (with open operational models) did not experience such requests. Furthermore, access cases within 10 institutions were linked to clinical utility (**Figure 4**), further highlighting the role of healthcare professionals within those cases (**Table 1**).

Policies

We compared the reported practices of the institutions by asking all respondents about their data access policies for sequenced individuals, including examples of their informed consent forms. We found that two out of nine informed consent forms provided to us included information about individual access policies. We also asked respondents if data access was granted based on a certain law or policy. In general, respondents viewed data access as a right of the individual. One respondent highlighted an organisational policy of not providing data access to minors until they are of legal age.

> *"In my own projects, we have an outspoken policy that says, they are children and we are not giving up the data to them. So, when they [turn] 18, and if they ask for the data, we will ask them instead, to give a DNA sample so we will do resequencing [. . .] for them. But we are not going to give the data to them."*
>
> *Respondent 4*

Several respondents also pointed out possible contradictions between GDPR and national laws, such as the one illustrated in the next quote:

> *"[. . .] the challenge that we have is that; there is actually, at some point, a contradiction between GDPR and [a national law pertaining to genetic testing]. Because, for example, we are not allowed to give genetic data [or] genetic results to the patients without [involving a] specialist, discussing the data with the patient first. So, when the patients request their data, we have to make sure that we [. . . just can't. . .] give them the data and [say], 'so here's all the variants'. Even worse, if it's a child, we cannot just give the data to the parents [and] say, okay, we'll do whatever you want. So that is one [challenge], and we haven't had a case yet that somebody asked for that data according to GDPR, but this is an ongoing discussion internally [on] what's the best way [. . .]."*
>
> *Respondent 5*

We also observed opposing opinions on who should bear the cost(s) associated with providing personal access

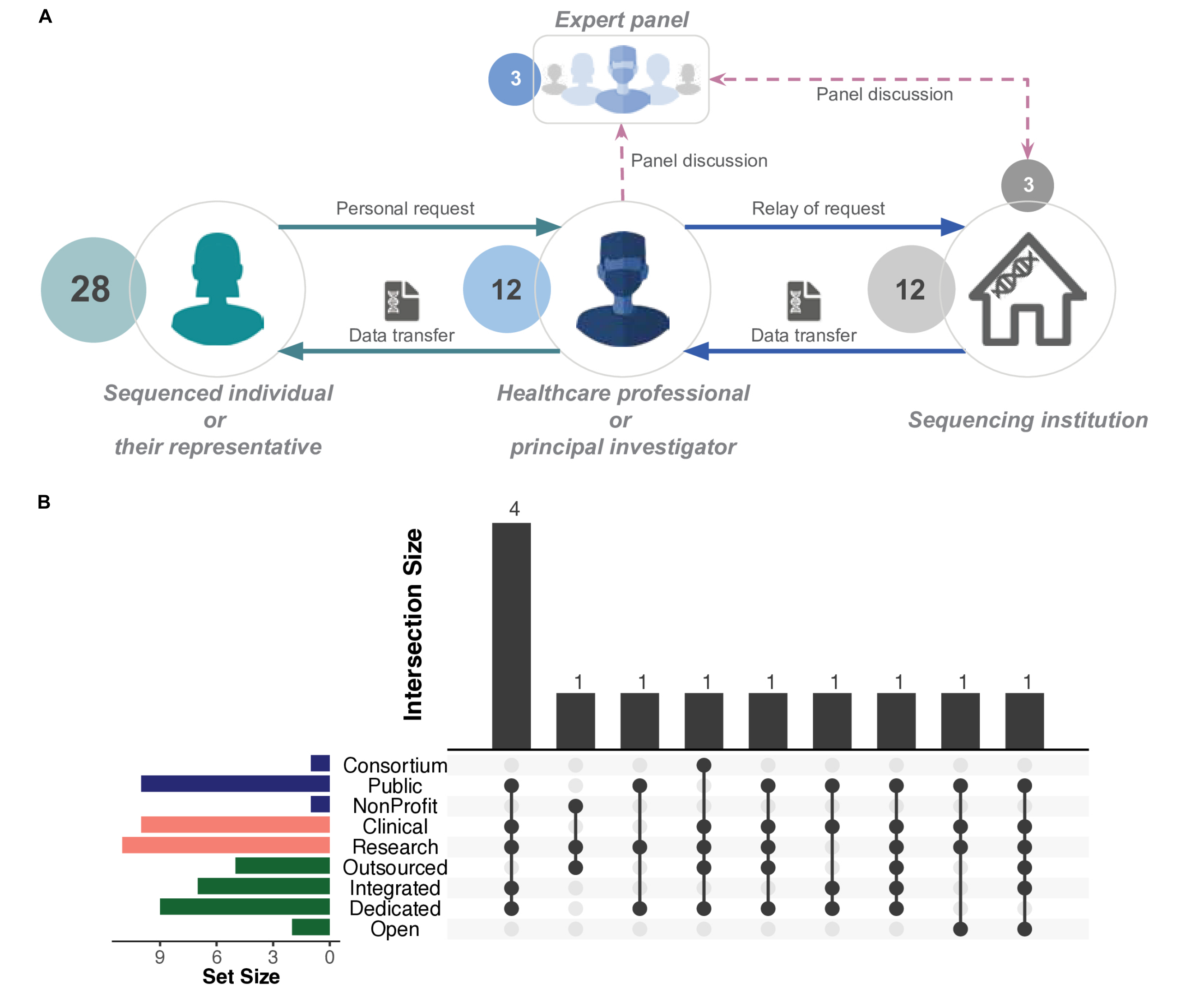


FIGURE 4 | Generalised illustration of access in practice. **(A)** Typical scenario of personal whole genome (WGS) or whole exome (WES) sequencing data access. Individuals request their data through a healthcare professional or a principal investigator. Those parties authorise the access to the genomic data. In certain cases, a panel of experts, which may include personnel from a sequencing institution, jointly decide and authorise the access. The final decision is relayed to the sequencing institution, which initiates the data transfer process, mediated by a healthcare professional or principal investigator. Sequenced individuals do not directly interact with the sequencing institution. For detailed descriptions, refer to **Table 1**. **(B)** UpSet plot (Conway et al., 2017) represents the number and types of sequencing institutions that successfully enabled personal access to raw genomic data of patients and/or research participants. The horizontal bars represent the elements of organisational structure, purpose of sequencing, and operational model (**Figure 1**). The bottom panel represents the intersections between the aforementioned elements. The vertical bars represent the size of those intersections. Detailed information available in **Supplementary File 2: Tables S2, S6**.

to raw genomic data, e.g., infrastructure, hardware, and staff/administration. Specifically, one respondent noted that the institution will not be able to cover potential costs.

> *I can only say [that] the institute will not be able to pay [for the hard drives to store the raw genomic data]. [Especially], when [...] people are coming [to us], requesting for the data, [but] the download speed is not fast enough, via internet. [Therefore] it is, of course, [should] not [be] the responsibility of the institution to pay [for] it."*
>
> *Respondent 6*

Another respondent highlighted that it was possible for their institution to fund the associated cost(s), despite the law allowing them to reject requests if costs were too high.

> *"If they want to access [to the data in] electronic format, [then] they [also have] the right to get [it] in [the] electronic format. [In a national law] there's also a caveat that says, if the effort and the financial cost would be too high, then you can refuse the request. But we see that it's possible [. . .]."*
>
> *Respondent 1*

TABLE 1 | Communication, decision, authorisation, and data transfer related to personal access to genomic data.

Communication	*“[...]. In both cases, the request came via the collaborator, [because] the individuals don’t know us. And they are not even aware that we are the ones processing their sample. [...]”* *Respondent 2*
	“[...]. Two other cases came through the actual hospital where the patients were on treatment because they came with a metastatic condition. The primary case or prior cases have been investigated with us. [. . .] the patients have been asking through the hospital, where they could get access to their data to make the most use of the old data together with the new ones out there.” *Respondent 1*
	“Well, [...] I don’t have any direct [contact] with the individuals. It is done by the [researchers] or by healthcare.” *Respondent 2*
	“Yes, it’s a medical doctor, who will send us a query and he will manage all the things with this patient and [. . .], we are not allowed to send the data directly to people.” *Respondent 7*
Decision and authorisation	*“[...] the [trial] coordinator; that’s usually the first point of contact, and then they contact for the reply of clinical heads of this program, and they decide it.”* *Respondent 1*
	“Well, you know, since we do the service [for] other [researchers], these questions are the responsibility of the researcher and not us, except for those projects that are [for] my research groups – internal projects.” *Respondent 4*
	“So, it’s discussed in committees that are set up, and then it’s kind of communicated directly to the person-owner of the data.” *Respondent 1*
	“[...] we came to the agreement that,[. . .] this was obviously a person who was very interested and who had some prior knowledge of what [he asked for]. So, we decided on giving out both [the BAM and VCF].” *Respondent 3*
Data transfer	*“[. . .] we just need to put them on to some external hard drive or whatever, and we hand this out.”* *Respondent 1*
	We send the data on hard drives to the collaborator, and then they send it further to [the sequenced individuals], and we don’t know the name of individuals. We use numbers. *Respondent 8*
	“[...]. Right now, we had like two or three times, and we sent hard drives with the data.” *Respondent 9*

Finally, it is unclear if such data access is fully compliant with the right of individuals to access their health record or their personal data.

Outlook

At the end of the interviews, we asked all the respondents to provide their future outlook on such cases of personal raw genomic data access. It must be noted that answers include both institutional policies and personal views of respondents, whereby the latter does not represent institutional policies. Twenty-one (∼64%) responses were supportive of the right and providing an option for individuals wanting to obtain access to their own sequencing data. Fourteen (∼42%) believed that the number of personal raw genomic data access requests from sequenced individuals will grow in the foreseeable future. Finally, 10 (∼30%) respondents indicated that their respective organisations were currently developing processes to manage those requests (**Supplementary File 2: Table S6**).

DISCUSSION

This study represents the first empirical study of genomic data management and personal access to raw genomic data for sequenced individuals. It demonstrates the frequency of access requests, the overwhelming tendency of sequencing institutions to grant such access, as well as technical and procedural complexities involved.

Organisational Structure and Operational Models

This study covered a diverse set of sequencing institutions within the EU/EEA genomics ecosystem, with varying organisational structures and operational models, thus providing a good representation of the current landscape, especially in the context of evolving regulations (i.e., GDPR).

Accountability for Data Retention and Personal Access

The institution or department responsible for genomic data management are typically physically, if not legally, separate from the institution or department responsible for interaction with sequenced individuals. However, certain organisations had units performing both data management and communicating with sequenced individuals within the same physical location, but were in fact independent, yet highly collaborative entities (e.g., departments, units, or groups). On the other end of the spectrum were institutions within which these two functions were located in clearly distinct institutions, with a clear service-based relationship and minimal collaboration. Particularly for institutions where several functions are fulfilled by the same units, it is necessary to clarify how these functions correspond with roles and responsibilities that are of legal relevance. The GDPR, for example, distinguishes between "data controllers," who determine the purpose of processing and who are primarily responsible for respecting individual rights, and "data processors," who carry out data processing services on behalf of controllers. The GDPR also recognises the possibility of joint controllership, where more than one party is responsible for protecting data and meeting demands from the individual right of access.

In light of these complex organisational structures, we recommend that there should be clear instructions for individuals regarding how to request access to their raw genomic data, including a clear point of contact, when they consent to having their genome sequenced. If requests for an individual's raw genomic data access are made directly to the sequencing institution/unit, they may need to be directed back to the appropriate access point. It should also be clear which organisation is responsible for determining if access should be provided, and according to what criteria. A failure to respond to access requests under the GDPR could lead to legal liability for both the requesting party and the sequencing institution.

Clinical Versus Research Data

One important consideration in discussion about both data retention and right to access to raw genomic data is the distinction between research and clinical data. Most notably, research data may not be considered of sufficiently high quality to enable meaningful consumer reuse (Shevchenko and Bale, 2016). Consumers may insist, however, that it is them, and not the sequencing institution, who should be able to make this determination, if necessary under the guidance of relevant experts (e.g., genetic counselors). In Europe, however, the right of data subjects to access their own raw genomic data in the research context may be restricted by member states, under Article 89 of the GDPR. While access rights are typical in the clinical setting, it remains unclear whether or not raw genomic data are considered part of patients' medical records (Thorogood et al., 2018). Even if there is no legal requirement, research projects may still opt for ethical or engagement reasons to provide access. Complicating things further is that a significant number of sequencing institutions provide both types of sequencing (research and clinical). The emergence of numerous national clinical genomics projects designed as learning health systems that routinely collect clinical data for the purposes of both care and research is also eroding this distinction (Stark et al., 2018; Price and Cohen, 2019).

Sequencing institutions may need processes to distinguish between research and clinical data for the purposes of retention or personal access. Alternatively, they may decide to adopt a single policy on personal access for all data in favor of the strictest requirements (i.e., to provide access).

Actual and Potential Sequencing Capacity

Our respondents reported rapid increases in WES/WGS potential and actual sequencing capacity from previous years. Moreover, growing competition between manufacturers in producing the most cost-efficient sequencing technology platforms was seen by our respondents as providing sequencing institutions with ample choices for further expansion of their sequencing capacity, and thus leading them to "stock up" on sequencing capacity. It must be noted that the unused sequencing capacity is typically due to various limiting factors such as (i) research funding, (ii) capacity of facility/institution, (iii) consortia activity, and (iv) clinical demand. The latter is especially true as respondents could not predict future actual capacity related to clinical genomic data, likely because of (i) unclear public healthcare allocation (budget) for WES/WES, (ii) emerging state-of-the-art sequencing platforms, which may result in (iii) falling costs of WES/WGS. Yet, sequencing institutions foresee performing more WES in the near future as it currently is and it will be considerably cheaper compared to WGS, indicating that the lowering costs of WGS is still insufficient to justify its cost for all use cases, though it will be important in certain niches (e.g., rare diseases).

The increasing amounts of genomic data produced in the clinical and in the research domain will have important ramifications for both data retention and the provision of raw genomic data access to sequenced individuals. WES/WGS is a platform technology, which generates rich and stable information that can be used for multiple clinical, research, and recreational purposes over time. Reuse of sequences has potential value not only for sequenced individuals, but also for healthcare systems, science, and commerce. Of course, reuse of data depends on deployment of standard sequencing platforms, analysis pipelines (where applicable), and file formats to ensure both interoperability and quality. Our results suggest significant variation in sequencing practices and pipelines. High interoperability and quality standards are needed to ensure that sequenced individuals can access raw genomic data for consumer use or for redistribution to other service providers and researchers (patient-centric data sharing; Kish and Topol, 2015). This is to ensure that data are meaningful and trustworthy for a number of downstream, distributed users. The right of consumer portability (closely related to the "right to access") has been recognised by GDPR by stating that data controllers should provide personal data "in a structured, commonly used, machine-readable and interoperable format" (Recital 68). While there does already appear to be a relatively high level of standardisation and reproducibility for human WGS/WES data generation and processing to enable medical and research reuse (DePristo et al., 2011; Auwera et al., 2013), most sequencing institutions do not currently provide levels of standardisation aiming to enable meaningful consumer reuse. This may change with the growing frequency and awareness of personal access to raw genomic data.

Genomic Data Management

With the rapid increase in sequencing capacity, questions arise as to who will store raw genomic data, in what form, and for how long. Our study, the first of its kind to review data retention, reveals uncertainty over who was responsible for storing data. In some cases, sequencing institutions were storing data as a stop-gap measure until requesting organisations developed sufficient capacity to do so. There was also a general lack of clear institutional policies about the duration of data retention, and significant variation between the policies that do exist (from a couple of months to indefinite). Unclear and varied retention policies are surprising considering legal requirements of data retention that may apply, particularly in clinical contexts. Retention policies were also not consistently described in the consent forms we reviewed. This is in line with the findings of previous reviews (Shabani et al., 2018). On the basis of these findings, we recommend that sequenced individuals should be provided transparent information about the length and location of storage at the time of consenting to their DNA being sequenced. A further area for exploration would be to determine if these requirements apply, or should apply, to raw genomic data.

Data retention practices present important challenges. On the one hand, longer-term storage of data can provide practical opportunities for quality control, re-interpretation, and reuse for secondary research purposes, and also allows individuals a greater span of time to request personal genome access. On the other hand, given the potential increase in sequencing capacity, storage may soon start to pose a bottleneck and sustainability challenge (especially as we move to WGS), though respondents did not suggest this was an immediate problem. In the view of advancements in sequencing technologies and the decreasing costs of sequencing, long-term storage of data may not seem cost-efficient. Moreover, data privacy principles such as data minimisation, which dictates that personal data should only be kept as long as necessary to carry out a specific purpose, could pose challenges to long-term storage of genomic data. The implementation of such principles into practice is still an ongoing process among many sequencing institutions, as also highlighted in our study. Finding ways to ensure compliance with emerging regulatory requirements without giving up the benefits of long-term genomic data retention is one of the key challenges currently facing sequencing institutions in the EU/EEA (Wagner et al., 2014). Sequencing institutions could be supported in balancing these interests through the development of standard storage technologies (e.g., compressed file formats, electronic health records) and practices. This could be pursued initially through voluntary standards organisations (e.g., Global Alliance for Genomics and Health, GA4GH; Health Level Seven, HL7) and through professional guidelines and best practices (e.g., American College of Medical Genetics and Genomics, ACMG; Advancing Human Genetics & Genomics, ASHG; The European Society of Human Genetics, ESHG), and could eventually be incorporated into laboratory regulations (Botkin et al., 2015; Deignan et al., 2019).

That said, in determining the period for data retention in the context of raw genomic data, other existing relevant regulations, such as those concerning minimum/maximum length for storage of medical information in the healthcare setting may apply. Consequently, it would be important to clarify the status of raw genomic data, namely, whether they would be considered as part of patient medical records or not.

Requests to Access Personal Genomic Data

Previous work has found that individuals are typically interested in obtaining access to their own genomic/genetic data (Lunshof et al., 2014; Middleton et al., 2015). It may help that various national regulations and the GDPR require data controllers to inform data subjects of their right to access data [Article 13(2)(b)]. Moreover, as more third-party service providers emerge, rising consumer awareness may lead to more individuals requesting access.

At present, however, the overall number of requests for genomic data is very modest in comparison to the number of sequenced individuals. Possible reasons for this are the relatively recent adoption of WES/WGS, as well as low interest, awareness, and consumer utility. Moreover, many genomic data sets in the EU/EEA are currently generated in research contexts that anonymise data, thus precluding return. The plausibility of this explanation is supported by our finding that only two access requests were related to research genomic data. Moreover, the complex structures and operations of sequencing institutions may lead to a lack of clear coordination of responsibility within and between organisations. For instance, service-oriented commercial sequencing institutions did not see any access cases possibly due to the lack or limited communication between such sequencing institutions and the intermediary contact (i.e., clients), beyond data generation and processing. Finally, most of the informed consent forms we analysed do not consistently mention access rights, as addressed by previous work on this topic (Shabani et al., 2018).

Our study found that in all cases where access was requested, the sequencing institution gave them access, despite a lack of formal internal policies and procedures. This is a general indication that sequencing institutions recognise their ethical responsibility and the rights (legal) of sequenced individuals to access their raw genomic data. It is therefore important for sequencing institutions to establish clear policies and procedures for personal raw genomic data access. As such, one noteworthy finding of this study is the observation that most sequencing institutions make decisions to grant access on an *ad hoc*, case-by-case basis. It is unclear who within the institution authorises the access and according to what criteria. Similarly, there are no standards or best-practice guidelines for protecting the privacy, security, and well-being of sequenced individuals during the access process. Provision of access through an intermediary with appropriate genetics expertise can help sequenced individuals better understand the meaning and limits of genetic data. It appears that the interaction between research and healthcare personnel is quite common in sequencing units integrated within healthcare institutions. Staff of dedicated and/or integrated sequencing institutions (e.g., respondents) are able to communicate directly with respective intermediary contacts, which are, in turn, able to communicate with the patient, thus creating a conducive environment for providing raw genomic data access to sequenced individuals. However, present *ad hoc* and case-by-case-based practices may not be scalable.

Once institutional policy about when to provide personal genome access has been formulated, there are additional technical and practical questions about how data will be accessed. Technically, should data be provided on a hard disk, through a web portal, or through the cloud? Who must bear the cost for such access, the individual, the requesting institution, the sequencing institution, or the healthcare system? Security and privacy measures are also important elements that need to be adequately protected when retaining, sharing, and accessing sensitive data.

It should be noted that, as of yet, there is no evidence available on what individuals do, or intend to do, with their raw genomic data after access, and it is a matter of an ongoing debate how much support they should receive in understanding/interpretation of such data, and from whom. Seeking a second opinion might be one reason for patients to request access to their raw genomic data. Currently, there are some third-party online services that also offer interpretation services to the individuals (Guerrini et al., 2019). They may also opt to share their data with interested third parties such as biotech or pharma companies in exchange for monetary or non-financial incentives (Ahmed and Shabani, 2019). However, it is not clear if patients should receive professional support when using such online services. This will, of course, depend on the context—is the interpretation for healthcare purposes (e.g., serious disease predispositions), or for more general preventative, well-being or recreational purposes? If individuals are seeking medical interpretation, the best option for individuals would be to reuse their raw genomic data within third-party healthcare institutions, which includes the guidance of qualified professionals by design (Wright et al., 2017; Middleton, 2018). However, to the best of our knowledge, this particular use case of interoperability between EU/EEA healthcare institutions is yet to be explored and may be highly complex given varying infrastructure, resources, and capabilities of different healthcare institutions. Most importantly, the aforementioned uncertainties of third-party reuse of raw genomic data should first be explored through empirical research to distinguish the concrete needs and risks from hypothetical ones (Middleton, 2018). This should further guide the development of interoperability channels specific to the reuse of genomic data.

Limitations

A general limitation to this study is that we may not have covered all possible types of sequencing institutions (e.g., commercial institutions and consortia) and personal access requests, including those that were possibly overlooked, did not respond, or declined to participate. Furthermore, we did not cover ordering institutions (clients) that may have received access requests that were not passed on to the sequencing institutions in the study, but rather handled by the clients themselves. Most importantly, given the complexity of organisational relationships and structures, we were unable to directly interview the organisation or the gatekeepers of access requests, including healthcare professionals and expert panels. Neither did we interview those persons within an organisation most knowledgeable about its infrastructure (e.g., IT specialist) and policies (e.g., lawyer or data steward) as this was not a criterion for the selection of interviewees (we spoke to whoever from the organisation that agreed to speak with us).

We would also like to highlight potential bias between respondent sequencing institutions and those declining to be interviewed. It may be that institutions within our professional networks were more likely to agree to interview. Decliners can be characterised as follows: most of the negative responses stemmed from people who did not respond to our communication at all or failed to schedule an interview, while five outright declined to interview, with four of them providing reasons for rejecting (see **Supplementary File 5**). If given, the most common reason for rejection was due to the preference in answering the questions in a written format. However, we decided against it (i.e., written questionnaires) to maintain consistency of our data collection methodology and also due to the nature of the open-ended questions, which are more suited within an interview setting. However, given that we successfully surveyed a large proportion of identified institutions (63 of 83), it is likely that we achieved saturation and so this bias is expected to be limited.

We also did not systematically analyse differences in national regulatory frameworks, written institutional policies, and governance documents (if they exist), outside the limited number of informed consent forms. In that regard, we were unable to compare information from statements within the interviews with a complete set of informed consent forms from the organisations, outside the limited number of informed consent forms obtained from those organisations, used for validation. It is also crucial to investigate other potential professional concerns from the perspectives of the healthcare professionals that may disfavor personal access to raw genomic data. Future research may also want to consider cross-country comparisons of sequencing institution structure, data retention, and personal access. Our exploratory study aimed at identifying general trends rather than making these granular comparisons. Moreover, our central finding, that few institutions have formally addressed retention or personal access, seems to have general implications across Europe. Our study remains the first exploratory study providing empirical evidence on the organisational structures, current and future sequencing capacity, and approaches to genomic data retention and personal genome access of sequencing institutions located in the EU/EEA.

Outlook

We find that sequencing capacity in Europe is growing and that some sequenced individuals are requesting access to their raw genomic data. Despite these trends, we also find that sequencing institutions are largely unprepared to handle questions of retention and personal access, and have yet to develop clear policies and practices. In a broader context, this study gives insight into the complexity and the general direction of the emerging genomics ecosystem. In that regard, we hope that this study becomes a catalyst for future explorations of similar nature with other stakeholders of the genomics ecosystem, and enhances the development policies and best practices in the context of personal access to raw genomic data.

DATA AVAILABILITY STATEMENT

The data sets generated and analysed for this study are included as part of the manuscript as **Supplementary Files 1–6**. Code for analysis and visualisations are available in GitHub repository (https://github.com/shaman-narayanasamy/sequencing_institution_analysis). Please contact authors for in case of any inquiries about the raw data.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by National Data Protection Commission of Luxembourg. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SN and RK conceptualised and designed the study. AT, BP, BK, and MS provided critical feedback on the study design. SN and VM coordinated the study. VM, SN, and RK prepared the confidentiality statement, pre- and post-interview material. SN, VM, and RK performed the interviews (i.e., data collection). RK and VM managed communication with respondents before and after the interviews. VM, AB, and SN performed the transcriptions. SN, VM, AB, and RK carried out the data curation, analysis, and visualisation. AT, MS, BK, and BP interpreted the data. SN, AT, MS, VM, and AB drafted the manuscript while critical revision was provided by BK, BP, and RK. All the authors approved the manuscript for publication.

FUNDING

This research was sponsored by Megeno S.A.

ACKNOWLEDGMENTS

We thank Megeno S.A. for funding this work. We also thank the Megeno research team, especially David Rubio Hermosa, Clara García Obelleiro, Dheeraj Bobbili, Tobias Hasenberg, Thanh Phuong Nguyen, Davit Chokoshvili, and Erich Felber for supporting various tasks related to this work. We duly thank the University of Luxembourg Incubator for providing the premises to carry out the work. Finally, our gratitude goes out to all the respondents and institutions for their participation in this interview study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00303/full#supplementary-material

REFERENCES

Ahmed, E., and Shabani, M. (2019). DNA data marketplace: an analysis of the ethical concerns regarding the participation of the individuals. *Front. Genet.* 10:1107. doi: 10.3389/fgene.2019.01107

Auwera, G. A. V., der Carneiro, M. O., Hartl, C., Poplin, R., Angel, G., del Levy-Moonshine, A., et al. (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Birney, E. (2019). The convergence of research and clinical genomics. *Am. J. Hum. Genet.* 104, 781–783. doi: 10.1016/j.ajhg.2019.04.003

Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. *bioRxiv* [Preprint], doi: 10.1101/203554

Bonfield, J. K. (2014). The scramble conversion tool. *Bioinform. Oxf. Engl.* 30, 2818–2819. doi: 10.1093/bioinformatics/btu390

Botkin, J. R., Belmont, J. W., Berg, J. S., Berkman, B. E., Bombard, Y., Holm, I. A., et al. (2015). Points to consider: ethical, legal, and psychosocial implications of genetic testing in children and adolescents. *Am. J. Hum. Genet.* 97, 6–21. doi: 10.1016/j.ajhg.2015.05.022

Bredenoord, A. L., Kroes, H. Y., Cuppen, E., Parker, M., and van Delden, J. J. M. (2011). Disclosure of individual genetic data to research participants: the debate reconsidered. *Trends Genet.* 27, 41–47. doi: 10.1016/j.tig.2010.11.004

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinform. Oxf. Engl.* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Deignan, J. L., Chung, W. K., Kearney, H. M., Monaghan, K. G., Rehder, C. W., and Chao, E. C. (2019). Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the american college of medical genetics and genomics (ACMG). *Genet. Med. Off. J. Am. Coll. Med. Genet.* 21, 1267–1270. doi: 10.1038/s41436-019-0478-1

den Dunnen, J. T. (2015). The DNA bank: high-security bank accounts to protect and share your genetic identity. *Hum. Mutat.* 36, 657–659. doi: 10.1002/humu.22810

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Dove, E. S. (2018). The EU general data protection regulation: implications for international scientific research in the digital era. *J. Law. Med. Ethics* 46, 1013–1030. doi: 10.1177/1073110518822003

European Parliament and Council, (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council - of 27 April 2016 - on the Protection Of Natural Persons With Regard To The Processing Of Personal Data And On The Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation*. Brussels: European Parliament and Council.

Graneheim, U. H., and Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Educ. Today* 24, 105–112. doi: 10.1016/j.nedt.2003.10.001

Guerrini, C. J., Wagner, J. K., Nelson, S. C., Javitt, G. H., and McGuire, A. L. (2019). Who's on third? Regulation of third-party genetic interpretation services. *Genet. Med.* 22, 4–11. doi: 10.1038/s41436-019-0627-6

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

Kish, L. J., and Topol, E. J. (2015). Unpatients—why patients should own their medical data. *Nat. Biotechnol.* 33, 921–924. doi: 10.1038/nbt.3340

Lunshof, J. E., Church, G. M., and Prainsack, B. (2014). Raw personal data: providing access. *Science* 343, 373–374. doi: 10.1126/science.1249382

Middleton, A. (2018). Society and personal genome data. *Hum. Mol. Genet.* 27, R8–R13. doi: 10.1093/hmg/ddy084

Middleton, A., Wright, C. F., Morley, K. I., Bragin, E., Firth, H. V., Hurles, M. E., et al. (2015). Potential research participants support the return of raw sequence data. *J. Med. Genet.* 52, 571–574. doi: 10.1136/jmedgenet-2015-103119

Nelson, S. (2016). Geneticists should offer data to participants. *Nat. News* 539:7. doi: 10.1038/539007a

Price, W. N., and Cohen, I. G. (2019). Privacy in the age of medical big data. *Nat. Med.* 25, 37–43. doi: 10.1038/s41591-018-0272-7

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schickhardt, C., Fleischer, H., and Winkler, E. C. (2020). Do patients and research subjects have a right to receive their genomic raw data? An ethical and legal analysis. *BMC Med. Ethics* 21:7. doi: 10.1186/s12910-020-0446-y

Shabani, M., Vears, D., and Borry, P. (2018). Raw genomic data: storage, access, and sharing. *Trends Genet.* 34, 8–10. doi: 10.1016/j.tig.2017.10.004

Shevchenko, Y., and Bale, S. (2016). Clinical versus research sequencing. *Cold Spring Harb. Perspect. Med.* 6:a025809. doi: 10.1101/cshperspect.a025809

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfied, M. J., et al. (2018). Integrating genomics into healthcare: a global responsibility. *Am. J. Hum. Genet.* 104, 13–20. doi: 10.1016/j.ajhg.2018.11.014

Thorogood, A., Bobe, J., Prainsack, B., Middleton, A., Scott, E., Nelson, S., et al. (2018). APPLaUD: access for patients and participants to individual level uninterpreted genomic data. *Hum. Genom.* 12:7. doi: 10.1186/s40246-018-0139-135

Wagner, J. K., Mozersky, J. T., and Pyeritz, R. E. (2014). "Use it or lose it" as an alternative approach to protect genetic privacy in personalized medicine. *Urol. Oncol.* 32, 198–201. doi: 10.1016/j.urolonc.2013.09.016

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wright, C. F., Middleton, A., Barrett, J. C., Firth, H. V., FitzPatrick, D. R., Hurles, M. E., et al. (2017). Returning genome sequences to research participants: policy and practice. *Wellcome Open Res.* 2:15 doi: 10.12688/wellcomeopenres.10942.1

Wright, C. F., Ware, J. S., Lucassen, A. M., Hall, A., Middleton, A., Rahman, N., et al. (2019). Genomic variant sharing: a position statement. *Wellcome Open Res.* 4:22. doi: 10.12688/wellcomeopenres.15090.2

Conflict of Interest: This study may guide the development of future products/services of Megeno S.A. SN, RK, VM, and AB are employees of Megeno S.A. AT does paid consulting work for Megeno.

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.

The reviewer SN declared a past co-authorship with two of the authors AT and BP to the handling Editor.

Copyright © 2020 Narayanasamy, Markina, Thorogood, Blazkova, Shabani, Knoppers, Prainsack and Koesters. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

PERSPECTIVE
published: 08 May 2020
doi: 10.3389/fgene.2020.00446

How Can We Not Waste Legacy Genomic Research Data?

Susan E. Wallace[1], Emily Kirby[2] and Bartha Maria Knoppers[2]*

[1] Department of Health Sciences, University of Leicester, Leicester, United Kingdom, [2] Centre of Genomics and Policy, McGill University, Montreal, QC, Canada

Enabling genomic and biomedical data to be shared for secondary research purposes is not always straightforward for existing "legacy" data sets. Researchers may not know whether their data meet ethical and regulatory requirements for sharing. As a result, these data, collected using public funds and the good will and efforts of the donors and investigators, may not be used beyond their original purpose. Single-use plastics are now being banned in many countries; single-use research should be avoided if possible. This paper describes a filter developed through the driver projects of the Global Alliance for Genomics and Health that can be used by researchers to help them determine the extent of sharing possible for their legacy data and actions to be taken to enable further sharing.

Keywords: consent, data sharing, policy, secondary research, genomic research

OPEN ACCESS

Edited by:
Stephan Beck,
University College London,
United Kingdom

Reviewed by:
Gabriele Werner-Felmayer,
Innsbruck Medical University, Austria
Stamatina Liosi,
University College London,
United Kingdom

****Correspondence:***
Susan E. Wallace
sew40@leicester.ac.uk

Specialty section:
This article was submitted to
ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

Received: *25 November 2019*
Accepted: *14 April 2020*
Published: *08 May 2020*

Citation:
Wallace SE, Kirby E and Knoppers BM (2020) How Can We Not Waste Legacy Genomic Research Data? Front. Genet. 11:446. doi: 10.3389/fgene.2020.00446

INTRODUCTION

Sharing of research data between institutions and across national and international borders is an expectation for many involved in genomic research studies. Too often, though, datasets languish because, amongst other reasons, researchers are unaware of whether the original consent given includes further data sharing or whether existing ethical, legal, and institutional requirements allow such sharing. The Global Alliance for Genomics and Health (GA4GH) "…both advocates for responsible data sharing and produces the practical standards to enable such a future." (Birney, 2019). Through its driver projects, "real-world genomic data initiatives" that help guide and implement data sharing activities[1], and workstreams, stakeholders work together to develop policies, tools and standards that follow the GA4GH *Framework for Responsible Sharing of Genomic and Health-Related Data* which provides, within a human rights framework, "a set of foundational principles for responsible research conduct and oversight of research data systems in the realm of genomic and health-related data sharing." (Knoppers, 2014). The recently revised GA4GH Consent Policy[2], which was written, "…to guide the sharing of genomic and health-related data in a way that supports the autonomous decision-making of data subjects," states that tools should be developed to support

[1] Global Alliance for Genomics and Health. Driver Projects. Available at https://www.ga4gh.org/how-we-work/driver-projects/. Accessed 30Mar20.

[2] Global Alliance for Genomics and Health Consent Policy (Sept. 2019). Available at https://www.ga4gh.org/wp-content/uploads/GA4GH-Final-Revised-Consent-Policy_16Sept2019.pdf. Accessed 24Nov19.

data donors[3]' understanding of data sharing plans and to ensure data are shared as was agreed in the consent. This filter is one example of a flexible tool that, as part of a larger governance framework, can help researchers determine if legacy datasets can be shared, within applicable ethical, and legal requirements while respecting patients and participants' wishes.

MATERIALS AND METHODS

Policy Background

There are many obstacles that stand in the way of wide-spread data sharing, yet there are also great incentives and rewards (**Figure 1**). Many existing "legacy" datasets from research or datasets generated from legacy or archival biological samples were created before widespread data sharing was encouraged. In these cases, research proposals and consent materials commonly did not include plans and language to enable further sharing, and often included conditions that limited the way in which a researcher could share, for example, across international borders or for research in other disease types than studied in the original research. At one time, sharing datasets could not be done easily so it was normal to not think about these possibilities. With technological changes and the genomics and big data revolutions, interrogating large datasets is now the norm, and in some cases, the only way in which the fundamental causes of disease could be found. Funders in many countries now require data sharing as part of grant conditions, and groups such as the GA4GH have worked tirelessly to develop tools and policies to share data for research purposes in a scientifically sound, ethical and lawful way. However, there are still barriers to overcome. The "…sharing of data and samples through global collaborative research networks…" has raised fears of a loss of privacy (Kaye, 2012). New legislation, such as the recent European General Data Protection Regulation (GDPR)[4], has caused many to be unsure as to what can and cannot be done[5]. In addition, there are those who might feel they cannot share, for many reasons, such as a misplaced commitment to "protect the privacy" of their participants or the need for secrecy in order to be the one to publish that ground-breaking, and promotion-securing, academic paper (Linek et al., 2017).

For those who actively seek to share legacy data, by for example joining a national or international consortium, the options may be reduced to not sharing, in effect wasting the opportunity to achieve the best from the data collected, or sharing in a way that might not fully reflect the original wishes of the data donors. A stepwise approach to assessing legacy datasets would allow researchers to decide how they could share their data to the greatest extent possible. The, "…ethical and legal interoperability process…" for assessing retrospective or "legacy" studies, proposed by Tassé et al. (2016) was chosen as a framework. This process asks the researchers to (1) identify the legal and ethical restrictions inherent in a data set, (2) determine whether these would allow or prevent participation in research collaborations, and (3) identify any options that would help to resolve these in an ethical and lawful way. Two GA4GH driver projects have now taken these steps and used them to construct a "legacy filter" to determine ethical and legal interoperability. While the use of the filter is different for each of the driver projects described below, the approach for creating or tailoring a filter can be used by anyone seeking to identify the requirements for sharing and using legacy data. The expectation is that any filter would be used within any existing governance framework and would inform, not exclude, other measures such as the use of data access agreements (DAAs) and other appropriate safeguards.

The International Cancer Genome Consortium (ICGC) and the Accelerated Research in Genomic Oncology Project (ICGC-ARGO)

The International Cancer Genome Consortium (ICGC) was established in 2008 to broadly and comprehensively map the structural aberrations of genomes and begin to understand the molecular basis of cancer (Hudson et al., 2010). Data are now available for over 88 projects across 17 jurisdictions (16 countries and the European Union) with >20,000 tumor genomes for 26 cancer types. The results of the analyses of these data are available through the Data Coordination Centre (DCC) via the ICGC website[6]. The ICGC Accelerated Research in Genomic Oncology (ARGO) project follows on from ICGC and, "…aims to analyze biospecimens from at least 100,000 cancer patients with high quality clinical data to address current key outstanding questions that are vital to our quest to defeat cancer."[7] This GA4GH driver project is an international research consortium of public studies and private commercial entities. Because research and clinical data from individuals will be contributed from many different countries with differing ethical, cultural and legal norms, ethical, and legal interoperability across studies is key.

Upon joining the consortium, ICGC members agreed to make the data as broadly available as possible under appropriate governance with minimal restrictions to expedite cancer and related research. It was recognized early that a core set of ethics elements were needed for researchers to include in consent materials given to, and in discussions with, prospective research participants. Two lists were created: a set of core elements member projects must agree to and a list of elements where there would be flexibility. For example, sharing with colleagues internationally was core, while decisions on whether to return individual research results were given over to the

[3] The term "data donor" has been chosen to be used throughout and is defined in the GA4GH Lexicon as, "The individual whose data have been collected, held, used and shared." Available at: https://www.ga4gh.org/wp-content/uploads/GA4GH_Data_Sharing_Lexicon_Mar15.pdf. Accessed 08Apr20.

[4] REGULATION (EUhile) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN. Accessed 24Nov19.

[5] The GA4GH has launched the GDPR & International Health Data Sharing Forum policy briefs to answer important questions in order to assuage concerns that might be blocking sharing. These are available at: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/. Accessed 02Apr20.

[6] ICGC Data Portal. Available at: https://dcc.icgc.org/. Accessed 24Nov19.

[7] Accelerating Research in Genomic Oncology. Available at: https://icgc-argo.org/. Accessed 24Nov19.

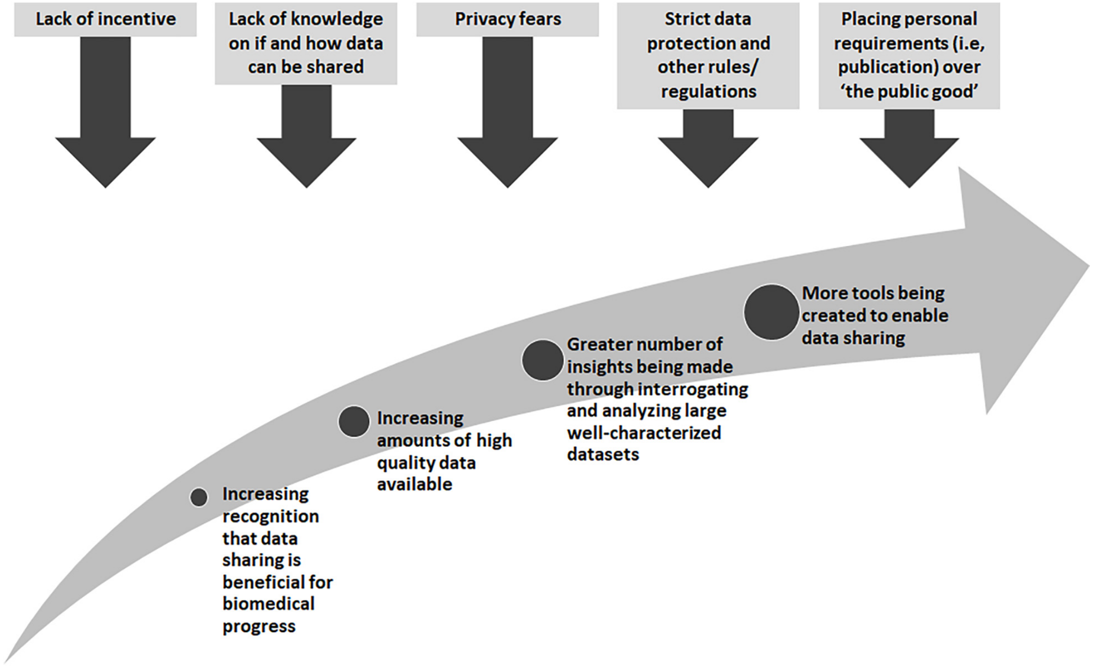
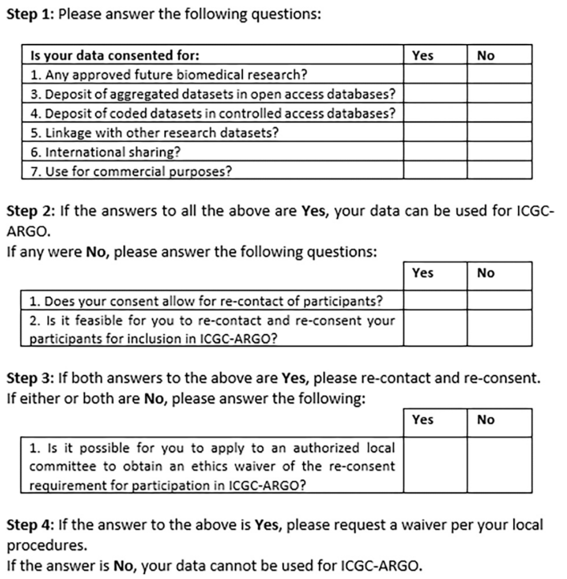

Lack of incentive
Lack of knowledge on if and how data can be shared
Privacy fears
Strict data protection and other rules/ regulations
Placing personal requirements (i.e, publication) over 'the public good'
Increasing recognition that data sharing is beneficial for biomedical progress
Increasing amounts of high quality data available
Greater number of insights being made through interrogating and analyzing large well-characterized datasets
More tools being created to enable data sharing

FIGURE 1 | Incentives for and pressures against greater sharing of biomedical data.

local member project to make. A later analysis of member study consent materials showed that, due to projects using their own institutionally approved consent materials in many different languages, it was very difficult to ascertain whether the core ICGC elements were being clearly communicated to participants (Wallace and Knoppers, 2011). While no concerns were raised (and to date this continues to be the case), anecdotal discussions highlighted that ambiguous consent language in ICGC member consents could preclude participation. For example, originally ICGC core elements stated that data would be used for cancer research but once requests for the data began to be received, it became clear that the ICGC data were useful for research related to, but not specifically for, cancer. It was decided that the scope of research in the core elements should be broadened to "cancer and related research" and later to "any approved biomedical research." This raised the question of whether the member projects were still in compliance. A letter was sent to each project leader asking them to confirm if their project had consent for two key elements: broad research use and international data sharing. If the project leader could not answer yes to these, they were instructed to speak with their ethics committee to see if it was possible to re-consent their participants or obtain a waiver (if possible and appropriate under local legal and ethics requirements.) Throughout the project, all member project leaders have completed this form. At least one ICGC project re-consented its participants for the broader scope of research (Wallace and Knoppers, 2011).

For ICGC-ARGO, a set of core ethics elements was again agreed on. It was decided that it would be beneficial if the projects could confirm whether their consent adhered to the details of the consortium at the beginning of the process of becoming a member, rather than retrospectively seeking this confirmation as with ICGC. Using the filter process, the authors reviewed the thirteen elements that had been drafted for ICGC, and translated them into a limited number of process-related elements (six) that it was felt required confirmation; other elements would be followed up on through other processes. For example, one core element is that data users will attest that they will not attempt to re-identify participants. As this is a provision in the legally binding DAA used by ICGC-ARGO, it was felt that enforcing this should be governed through the data access process which requires institutions to take legal responsibility for the actions of its researchers.

Figure 2 shows an early version of the filter for use by ICGC-ARGO as part of its Expression of Interest (EOI) process. This shows the seven core research consent elements of participation in this consortium but also provides further steps that researchers can take to enable their dataset to be shared respecting research ethics requirements. Step 1 takes one through the six and if these points cannot be met, Step 2 asks if re-contact and re-consent are possible. If not, applying for a waiver from an appropriate body, such as a research ethics board, is suggested. This will be useful in cases when re-contact was not foreseen or for when it is may be impracticable to re-consent participants. Consent language would need to be interpreted to judge whether it fits with the items in the tool, so contact information of an ICGC-ARGO team member is available if guidance is needed. A possible further option would be to anonymize the dataset. This is not the preferred option as de-linked data would be of limited



Step 1: Please answer the following questions:

Is your data consented for:	**Yes**	**No**
1. Any approved future biomedical research?		
3. Deposit of aggregated datasets in open access databases?		
4. Deposit of coded datasets in controlled access databases?		
5. Linkage with other research datasets?		
6. International sharing?		
7. Use for commercial purposes?		

Step 2: If the answers to all the above are **Yes**, your data can be used for ICGC-ARGO.
If any were **No**, please answer the following questions:

	Yes	**No**
1. Does your consent allow for re-contact of participants?		
2. Is it feasible for you to re-contact and re-consent your participants for inclusion in ICGC-ARGO?		

Step 3: If both answers to the above are **Yes**, please re-contact and re-consent.
If either or both are **No**, please answer the following:

	Yes	**No**
1. Is it possible for you to apply to an authorized local committee to obtain an ethics waiver of the re-consent requirement for participation in ICGC-ARGO?		

Step 4: If the answer to the above is **Yes**, please request a waiver per your local procedures.
If the answer is **No**, your data cannot be used for ICGC-ARGO.

FIGURE 2 | ICGC-ARGO consent assessment tool for participation.

value in clinical care and may pre-empt projects from updating datasets longitudinally.

An early version of the retrospective assessment filter was developed and informally piloted with a small number of ICGC projects that were considering joining ICGC-ARGO and this showed that ambiguous consent text could preclude participation. For example, if there is research consent to share data with one other named country, outside the one in which the research is being conducted, can the consent be interpreted as allowing "international data sharing" with any country? The wording of point 6: "Use by industry partners" has been changed as a result of discussions with potential project representatives and the ARGO Ethics and Governance Committee – should it be specifically aimed at commercial entities or made broader, such as "Use by bona fide researchers from institutions including not-for profit and commercial?"

Those completing the ICGC-ARGO EOI application must mark their agreement to this statement: We agree that their participant (donor) consents meet the requirements of inclusion in ICGC-ARGO as outlined in the ICGC-ARGO Participant Assessment tool in Appendix III. By including the retrospective assessment filter tool at this point in the formation of the consortium, we have tried to place consent harmonization at the heart of the recruitment process, and not as an afterthought. However, in anecdotal discussions with project representatives it appears that boxes may have been ticked without a full understanding of the specific project's informed consent materials. Discussions continue as to how to when and where would be the best place to introduce the filter. There are plans to automate this process by including a requirement to complete the filter as part of the online data submission process. This could allow greater scope for explaining the specific elements and for recording acceptance of its provisions.

The Human Cell Atlas

A similar "legacy assessment filter" is currently being developed in the context of the Human Cell Atlas (HCA) driver project. Given the diversity of tissues types and cells required to map the human body, the HCA presents an interesting scenario, since several contributors to the Atlas will need to consider both the use of legacy tissue samples (for e.g., in the case of rare specimens, or tissues collected prior to the creation of the HCA), as well as legacy datasets. In this perspective, the first draft of the HCA research consent assessment filter was divided into four steps, and namely:

1. Can the legacy tissue sample be used to generate datasets for the HCA?
2. Is the tissue donor's consent adequate to deposit datasets in the HCA data coordination platform?
3. What is the appropriate data tier for the datasets?
4. If requirements for previous steps are not met, is it possible to re-consent donors or seek an ethics consent waiver?

Steps 2 and 4 are similar to the elements used in the ICGC-ARGO filter. However, Step 1 was added in light of the complexities involved in the tissue sampling sources and scenarios envisaged by HCA contributors[8]. Furthermore, Step 3 was added to account for the potential levels of permission on data sharing for example, based on consent language, data protection requirements, source of tissue (e.g., paediatric, disease cohorts), or other policy requirements, such as open (public) versus registered versus managed access[9].

Although at the time of writing, the assessment filter is still being discussed within the HCA Ethics Working Group, it is hoped that the final filter will provide an educational guidance tool, pointing to different layers of considerations involved in the use of legacy tissues samples and datasets. We expect that pure legal compliance will depend on more than simply this assessment tool (for example, on data protection regulation, institutional policies, and ethics approvals, etc.). Nonetheless, dissemination of this tool to the HCA community aims at fostering an understanding of transparent, and responsible data governance, while maximizing legacy data sharing and use.

DISCUSSION

When initially prepared, the main objective of the legacy filter was to provide guidance on assessing whether research consent language used by member projects was sufficient to allow sharing within consortia, in response to the authors' experience with seeking interoperability between the ethical, legal and social issues (ELSI) linked to research studies. Because the teams

[8] Building the Human Cell Atlas: Issues with Tissues. (2019) Available at: http://www.genomicsandpolicy.org/Ressources/Issues-with-Tissues_2019.pdf. Accessed 24Nov19.

[9] The GA4GH Lexicon (https://www.ga4gh.org/wp-content/uploads/GA4GH_Data_Sharing_Lexicon_Mar15.pdf) defines open access as "Making data available without restriction." An example of managed access is registered access, defined in the Lexicon as, "A system of authentication and self-declaration prior to providing access to data."

organizing most scientific consortia are not "legal entities" they cannot enforce decisions across consortia, instead they must rely on each participating study to be able participate based on their own local legal requirements and cultural norms. Better harmonization of these, such as around sharing legacy data, would be beneficial to consortia, but has been shown that it would be difficult to achieve (Tassé et al., 2010). Data protection regulations, such as the GDPR, can add an additional layer of complexity to this reliance on local practices and knowledge. Consortia need to be aware that use of the filter does not in itself verify compliance. It is always contingent on the researcher being compliant with their own locally applicable data protection regulations. However, in their local adaptations of legacy filter tools, regional consortia may eventually consider adding additional steps to provide guidance on jurisdiction-specific data protection requirement (e.g., GDPR).

The filter can help with clarifying the consent elements needed for participation but familiarity with consent materials is needed. When principle investigators of a research study seek to be part of research consortia they may not have the in-depth knowledge of the ethical, social and legal requirements under which they must act, leaving the ELSI representatives (if there are such individuals) to raise concerns about whether participation conforms with the rules under which the data were gathered. Therefore, it is crucial that all researchers understand that they are taking responsibility for knowing, not only the content of their consent materials, but what their local (institutional or national) rules and regulations are, so that when they tick the boxes they do so in full knowledge of the commitment being made. Groups and individuals, such as data protection officers within institutions and research ethics committees, have a role in educating and working with their research teams, as well as learning themselves about working in national and international consortia.

ACTIONABLE RECOMMENDATIONS

It is vital that research data is shared for purposes that adequately match the understanding and consent given by data donors and that conform with applicable ethical, social, and legal requirements. While it is well-known that individuals may not remember the exact provisions in any given consent form that they have signed, it is also known that one of the most important considerations underlying agreement to participate in research is that researchers and academic institutions are worthy of their trust (Dixon-Woods and Tarrant, 2009). In addition, neither of the filter examples presented have been in place long enough to critically evaluate their success. Empirical evidence will be needed to validate the approach taken.

Therefore, we recommend that:

1. International consortia agree on a set of core elements for participation, design a filter to reflect these to be provided to study leaders considering participation.
2. Project leads attest that their consent materials meet the requirements for participation and that this attestation be recorded either on paper or through electronic means.
3. Consideration be given to the best way to present the filter, such as through EOIs or data submission processes.
4. Consortia use this and similar tools to educate their communities and raise awareness with respect to the complexities involved in the ethical governance of legacy datasets.
5. Local data protection officers, research ethics committees and others, such as legal experts, work with researchers to educate them on the ethical, social, and legal requirements surrounding data sharing.
6. Consortia that have used the filter share their experiences in order to enable improvements to be recommended.

CONCLUSION

This filter is proposed as one part of a larger governance framework to support research consortia. Its aim is not to place barriers in the way of researchers, but instead provide a way for them to know what contributing data to a consortium entails and to have a simple way to confirm that their consents meet the requirements for participation. If there are conditions that block participation, researchers will know what avenues they can take to share their data according to ethical and legal requirements. Datasets, like plastics, cannot continue to be single use. This filter is one way to encourage data sharing to the widest extent possible, in a responsible, ethical and lawful way that respects the wishes of the original data donor.

AUTHOR CONTRIBUTIONS

SW drafted the original draft of this manuscript. All authors contributed to the writing and editing of the text, and approved the final text.

FUNDING

BK, SW, and EK are supported by the Genome Quebec, Genome Canada, the Government of Canada, the Ministère de l'Économie, Innovation et Exportation du Québec (Can-SHARE grant 141210). BK and EK are supported for the HCA project by the Klarman Family Foundation, the Leona M. and Harry B. Helmsley Charitable Trust and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (grant number 2018-190754).

ACKNOWLEDGMENTS

We wish to thank those involved with the ICGC-ARGO Ethics and Governance Committee, Scientific Advisory Committee and Member Projects for their help in creating the ARGO filter. We would also like to thank the members of the Human Cell Atlas (HCA) Ethics Working Group for their ongoing feedback on the HCA filter. We would also like to thank the Global Alliance for Genomics and Health for its support of this work.

REFERENCES

Birney, E. (2019). The convergence of research and clinical genomics. *Am. J. Hum. Genet.* 104, 781–783.

Dixon-Woods, M., and Tarrant, C. (2009). Why do people cooperate with medical research? Findings from three studies. *Soc. Sci. Med.* 68, 2215–2222. doi: 10.1016/j.socscimed.2009.03.034

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabe, R. R., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Kaye, J. (2012). The tension between data sharing and the protection of privacy in genomics research. *Annu. Rev. Genomics Hum. Genet.* 13, 415–431. doi: 10.1146/annurev-genom-082410-101454

Knoppers, B. (2014). Framework for responsible sharing of genomic and health-related data. *HUGO J.* 8:3. doi: 10.1186/s11568-014-0003-1

Linek, S. B., Fecher, B., Friesike, S., and Hebing, M. (2017). Data sharing as social dilemma: influence of the researcher's personality. *PLoS One* 12:e0183216. doi: 10.1371/journal.pone.0183216

Tassé, A. M., Budin-Ljosne, I., Knoppers, B. M., and Harris, J. R. (2010). Retrospective access to data: the ENGAGE consent experience. *Eur. J. Hum. Genet.* 18, 741–745. doi: 10.1038/ejhg.2010.30

Tassé, A. M., Kirby, E., and Fortier, I. (2016). Developing an ethical and legal interoperability assessment process for retrospective studies. *Biopreserv. Biobank.* 14, 249–255. doi: 10.1089/bio.2015.0122

Wallace, S. E., and Knoppers, B. M. (2011). Harmonised consent in international research consortia: an impossible dream? *Genom. Soc. Policy* 7, 35–46.

Conflict of Interest: The 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.

Copyright © 2020 Wallace, Kirby and Knoppers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fgene.2020.00283

Public Attitudes Toward Precision Medicine: A Nationwide Survey on Developing a National Cohort Program for Citizen Participation in the Republic of Korea

Hannah Kim[1,2], Hye Ryun Kim[3], Sumin Kim[2], Eugene Kim[3], So Yoon Kim[1,2] and Hyun-Young Park[3]**

[1] Division of Medical Law and Ethics, College of Medicine, Yonsei University, Seoul, South Korea, [2] Asian Institute of Bioethics and Health Law, Yonsei University, Seoul, South Korea, [3] National Biobank of Korea, Center for Genome Science, Korea National Institute of Health, Cheonju-si, South Korea

This nation-wide survey was conducted among Korean adults to examine the public interest in and attitudes toward establishing a citizen participation cohort model and to collect data to support and determine the future policy and research directions of the Resource Collection Project for Precision Medicine Research (RCP-PMR) before the project proceeds. The demographic framework of the survey population was established based on the statistical standards of the Ministry of the Interior and Safety. An online survey was carried out using web panels between 14 May 2018 and 23 May 2018. Sampling was performed using a simple proportional allocation method considering region, gender, and age. From this survey, the RCP-PMR received very high support (94.5%) and the intention to participate was as high as 83.5%. Respondents had a very positive attitude toward providing their samples and information to the study (84.5–89.9%). In terms of incentives to participate, respondents wanted to receive health information (80.2%), monetary compensation (51.4%), and smart devices (41.3%). Most participants responded that it was appropriate to carry out the project at governmental research institutes (66.9%). Respondents also had a positive attitude toward sharing their information and samples as long as it was only shared with the governmental researchers who run the project (88.0%). However, the survey participants expressed concerns about the study being time consuming or a hassle (38.1%), privacy breaches (33.6%), and the lack of returning benefits of participation (25.1%). Participants had a negative attitude toward sharing their data with researchers who are not directly involved in the RCP-PMR. Considering the future use of the database derived from this project, it will be important to communicate with the lay public as well as the RCP-PMR participants to understand their needs in participating in the forthcoming study and to improve their understanding of the goals of the project, and how data sharing can contribute to disease research and prevention. The RCP-PMR should consider building an efficient citizen-participation program and privacy protection for the research participants.

Keywords: precision medicine cohorts modeling, participant engagement, public attitude, ELSI, benefits for participation, data sharing

OPEN ACCESS

Edited by:
Mahsa Shabani,
Ghent University, Belgium

Reviewed by:
Ma'n H. Zawati,
McGill University, Canada
Richard Milne,
Wellcome Genome Campus,
United Kingdom

****Correspondence:***
So Yoon Kim
syoonkim@yuhs.ac
Hyun-Young Park
hypark65@nih.go.kr

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *24 October 2019*
Accepted: *09 March 2020*
Published: *12 May 2020*

Citation:
Kim H, Kim HR, Kim S, Kim E, Kim SY and Park H-Y (2020) Public Attitudes Toward Precision Medicine: A Nationwide Survey on Developing a National Cohort Program for Citizen Participation in the Republic of Korea. Front. Genet. 11:283. doi: 10.3389/fgene.2020.00283

INTRODUCTION

Precision medicine aims to understand how a person's genetics, environment, and lifestyle can help determine the best approach to prevent or treat diseases (Collins and Varmus, 2015; Alzu'bi et al., 2019; Genetics Home Reference, 2019). Precision medicine integrates advanced technologies with enriched biomedical big data, including multi-omics; physiological, clinical, mobile, and remote health; and external environmental information to provide transformed healthcare services to one or more people (Collins and Varmus, 2015). For instance, the Precision Medicine Initiative All of Us Research Program (PMI-AURP) in the United States collects specimens and a wide range of personal health information including clinical data, genomic data, and lifelog data from at least 1 million Americans (National Institutes of Health, 2019). The 100,000 Genomes Project funded by the National Institute for Health Research and NHS England involves sequencing 100,000 genomes, including genomic, phenotypic, and other clinical data, from 85,000 patients with rare diseases or cancer (Peplow, 2016; Haga, 2017; Genomics England, 2019). These large-scale precision medicine cohort models necessitate public participation and collective engagement in conjunction with longitudinal collection, access, and use of data (Kaufman et al., 2016).

The Korea Centers for Disease Control & Prevention (KCDC) is planning to carry out the Resource Collection Project for Precision Medicine Research (RCP-PMR) from 2020 onward. This project is expected to collect clinical information, specimens, genetic data, environmental information, and lifelog data – which are essential for research and technological development – from individuals who agree to participate in the RCP-PMR. The collection, storage, and sharing of individuals' data are expected to be conducted mainly under the *Personal Information Protection Act* and the *Bioethics and Safety Act* (Kim et al., 2018). Qualified researchers who obtain approval to access the database by a proper authority will be able to use the information to conduct a variety of biomedical studies.

However, public acceptance of building a citizen participation national cohort model has not yet been studied in the Republic of Korea (ROK). In the case of the United States large-scale prospective cohort, the nationwide precision medicine initiative cohort study conducted surveys of United States adults to identify public concerns and problems that had to be addressed before the study (Kaufman et al., 2016; Okita et al., 2018). To benchmark a precision medicine cohort program such as the PMI-AURP, we conducted an Ethical, Legal, and Social Implications (ELSI) study in the form of a nationwide survey to confirm public attitude toward precision medicine and to collect opinions on the RCP-PMR before implementing it. The survey identified the social acceptance of the specimen and information provisions and the issues that must be addressed before the project can proceed.

MATERIALS AND METHODS

Survey Methods

We conducted online surveys to collect basic data for future policy directions and research by confirming public attitudes toward and opinions on the RCP-PMR. The online survey participants were recruited based on the statistics of the resident registration of the Ministry of the Interior and Safety at the end of January 2018 (men and women aged 20 and older). Sampling was performed using a simple proportional allocation method considering region, gender, and age. The sample selection and online administration of the survey were managed by the Nielsen Korea online survey firm. During the field period, 1,500 potential respondents of at least 20 years old were randomly sampled from Nielsen's web-enabled master panel of 500,000 Korean residents. The survey was fielded online between 14 May 2018 and 23 May 2018 (10 days).

Questionnaire Development

The questionnaire used in this study was developed in reference to the Kaufman et al. survey (Kaufman et al., 2016). The questionnaire was written in simple Korean and included 17 carefully selected multiple-choice questions about the RCP-PMR and eight items on social/demographic variables. The KCDC, which leads this precision medicine national cohort program, provided the draft description of precision medicine, and the RCP-PMR and the authors of this paper completed the description by including a comparison to the PMI-AURP. Respondents answered questions about precision medicine awareness and then confirmed a brief description of precision medicine and the RCP-PMR. Respondents were then asked several questions about the need for the project, their concerns and willingness to participate in the project, and the use of their data. See the questionnaire in the **Supplementary Material**. After completing the survey, participants received 4,000 South Korean Won (equivalent of USD 3.50) for their time.

Ethics Approval

The survey was approved by the Institutional Review Board (IRB) of the Yonsei University (approval number: Y-2018-0039). Under the Bioethics and Safety Act, written consent was exempted by the judge of the IRB because, due to the nature of the survey form, the respondents should read the survey information before starting the survey and thus were perceived to have agreed to participate in the survey. The survey was also designed for participants to withdraw their own participation at any time during or after the survey.

Statistical Analysis

Data were analyzed using the IBM SPSS 20 statistical software. Missing data were excluded by this online survey design. Respondents' socio-demographic characteristics were analyzed

Abbreviations: ELSI, Ethical, Legal and Social Implications; IRB, Institutional Review Board; KCDC, Korea Centers for Disease Control & Prevention; PMI-AURP, Precision Medicine Initiative All of Us Research Program; RCP-PMR, Resource Collection Project for Precision Medicine Research; ROK, Republic of Korea; SNS, Social Networking Sites.

TABLE 1 | Respondents' characteristics.

Variables		***N* (%)**
Gender	Men	743 (49.5)
	Women	757 (50.5)
Age	20–29	261 (17.4)
	30–39	259 (17.3)
	40–49	305 (20.3)
	50–59	389 (25.9)
	60+	286 (19.1)
Household Income (KRW)	<\2,000,000	153 (10.2)
	\2,000,000–\3,990,000	458 (30.5)
	\4,000,000–\5,990,000	508 (33.9)
	\6,000,000≤	381 (25.4)
Education	Less than middle school	46 (3.1)
	High school	342 (22.8)
	College and more	1,112 (74.1)
Social networking service	No use	138 (9.2)
	Former use	221 (14.7)
	Current use	1,141 (76.1)

using the variables of gender, age, region, household income, and education (**Table 1**). In addition, two multiple logistic regressions were examined (**Table 2**). The attitude toward the RCP-PMR and willingness to participate in the project were the dependent variables.

RESULTS

The Respondents

A total of 52,000 people were invited to participate in the survey *via* email and 4,271 connected to the website. Among them, 1,500 people fully responded, resulting in an invitation-response rate of 2.9% and an access-response rate of 35.1%. The demographic characteristics of the survey population are shown in **Table 1**. The gender distribution was nearly equal, 50.5% female and 49.5% male. The age ranges were 50–59 (25.9%), 40–49 (20.3%), 60+ (19.1%), 20–29 (17.4%), and 30–39 (17.3%). The distribution of household income per month in Korean Won (KRW) was <\2,000,000 (10.2%), \2,000,000 – \3,990,000 (30.5%), \4,000,000 – \5,990,000 (33.9%), and \6,000,000 ≤ (25.4%). The education level ranges were less than middle school graduate (3.1%), high school graduate (22.8%), and college and more (74.1%). Experience with social networking sites (SNS), such as Facebook, Twitter, Instagram, and KakaoTalk, were no use (9.2%), former use (14.7), and current use (76.1%). The margin of error on opinion estimates based on the sample of 1,500 is ± 2.53% in a 95% confidence interval.

Awareness of Precision Medicine

Participants were asked, "Have you ever heard of precision medicine?" Of the respondents, 11.5% answered, "I have heard of it and I know what it is"; 58.2% answered, "I have heard of it, but I

TABLE 2 | Results of two multiple logistic regressions examining socio-demographic variables related to survey participants' attitude toward the RCP-PMR and their willingness to participate in the project (*n* = 1,500).

	Demographic Group	**Unweighted *N* (weighted percent)**	**% who said the project definitely or probably should be done**	**Beta**	**SE**	***p*-value**	**% who are definitely or probably willing to participate in the project**	**Beta**	**SE**	***p*-value**
Total		1,500 (100)	94.5				83.5			
Gender	Men	743 (49.5)	95.0	0.197	0.234	0.401	87.6	0.641	0.149	0
	Women	757 (50.5)	94.1	ref			79.5	ref		
Age	20–29	261 (17.4)	92.7	−0.810	0.334	0.015	78.5	−0.695	0.207	0.001
	30–39	259 (17.3)	94.6	−0.474	0.359	0.186	79.9	−0.649	0.208	0.002
	40–49	305 (20.3)	93.4	−0.655	0.32	0.041	83.6	−0.404	0.203	0.047
	50+	675 (45.0)	95.7	ref			86.8	ref		
Household Income	<\2,000,000	153 (10.2)	90.8	−0.659	0.421	0.117	69.3	−1.311	0.263	<0.0001
	\2,000,000–\3,990,000	458 (30.5)	94.5	−0.313	0.353	0.376	79.9	−0.805	0.217	<0.0001
	\4,000,000–\5,990,000	508 (33.9)	94.3	−0.412	0.337	0.222	86	−0.386	0.219	0.078
	\6,000,000 ≤	381 (25.4)	96.3	ref			90.3	ref		
Education	Less than middle school	46 (3.1)	84.8	−1.155	0.5	0.021	76.1	−0.272	0.393	0.489
	High school	342 (22.8)	95.6	0.249	0.318	0.434	81	−0.066	0.181	0.716
	College and more	1,112 (74.1)	94.6	ref			84.6	ref		
Social networking service	No use	138 (9.2)	95.3	−0.792	0.338	0.019	84.8	−0.483	0.234	0.039
	Former use	221 (14.7)	93.7	−0.270	0.314	0.390	81	−0.215	0.196	0.274
	Current use	1,141 (76.1)	89.9	ref			76.8	ref		

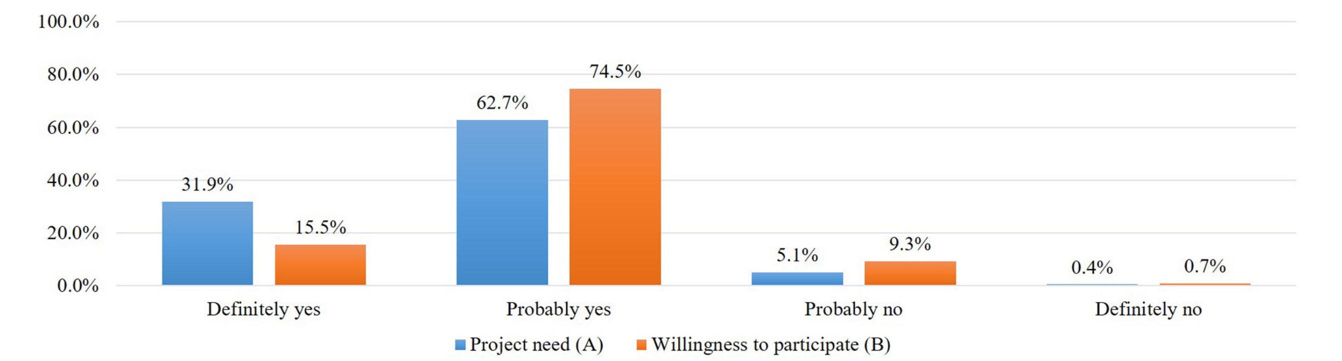


FIGURE 1 | Participants' attitude toward the need to implement the RCP-PMR **(A)** and their willingness to participate in the study **(B)**.

do not know what it is"; and 30.3% answered, "I have never heard of it." Among the 1,046 respondents who have heard of "precision medicine," respondents learned about it through media such as TV and radio (58.9%), the internet (49.8%), magazines and newspapers (21.5%), and hospitals (13.6%).

After reading the definition and a brief example of precision medicine, 96.1% responded that precision medicine is important for prevention and treatment of disease. The higher the education (less than middle school [89.1%], high school [95.9%], and college and more [96.4%]) or the higher the household income (<\2,000,000 [94.8%], \2,000,000 – \3,990,000 [95%], \4,000,000 – \5,990,000 [96.1%], and \6,000,000 ≤ [97.9%]) of the respondent, the higher the rating of the importance of precision medicine. Of the 69.9% of respondents who said they did not know what precision medicine was before the survey, 91.5% agreed on the importance of precision medicine.

Attitude Toward the Need to Implement the RCP-PMR

After introducing the concept of precision medicine, the plan of the project, and data sharing policies, we asked about the need for the RCP-PMR (**Figure 1A**). Most respondents (94.5%) agreed on the need to implement the study and, among variables, men (95.0%), older adults (older than 50, 96.1%), those with a higher education (high school [95.6%], and college and more [94.6%]), and current SNS users (95.3%) highly supported implementing the study (**Table 2**).

Adjusting for the other factors in **Tables 1**, **2** showed no significant differences between genders, region groups, and household income groups by a multiple logistic regression treating the need of the RCP-PMR as a binary independent variable. Younger age (20–29 [$p = 0.015$]) and lower education level (less than middle school [$p = 0.021$]) were independently associated with lower levels of support for the study. The SNS non-user group was significantly associated with higher levels of support for the study ($p = 0.019$).

Willingness to Participate in the Study

When asked about their intention to participate in the project, 83.5% of respondents said they would participate and 16.5% said they would not (**Figure 1B**).

Adjusting for the other factors in **Table 1**, age 20–49 (20–29: $p = 0.001$; 30–39: $p = 0.002$; 40–49: $p = 0.047$) and lower household income (< \2,000,000: $p < 0.0001$; \2,000,000 – \3,990,000: $p < 0.0001$) were independently associated with lower levels of willingness to participate in the study (**Table 2**). As a group, those who did not have experience with SNS were significantly more likely to say they would participate in the study if asked ($p = 0.039$).

Concerns About Participating in the Study

Among 247 respondents who said they would not participate in the study, we asked about the reasons why they have no intention to participate in the study. Of the respondents, 94 (38.1%) expressed concerns about it being time consuming or a hassle. The leakage of personal information was a concern for 83 respondents (33.6%), and 62 respondents (25.1%) were concerned about the lack of returning benefits of participation.

Willingness to Provide Personal Information and Samples for the Study

We asked all respondents, including those who said they would not participate, about their willingness to provide various types of samples and data to this project (**Figure 2**). Most respondents replied that they would provide clinical information ($n = 1,311$, 87.4%), samples ($n = 1,328$, 88.5%), genetic information ($n = 1,268$, 84.5%), and data on lifestyle ($n = 1,349$, 89.9%) and would link their data with existing national statistics from the Meteorological Administration and the Ministry of Environment (85.9%). Many of the respondents who would not provide specimens or personal health information were concerned about personal information leakage and privacy violations. Most respondents had a positive attitude toward providing specimens and information, and 84.5% to 89.9% of participants said they would provide certain types of samples and information.

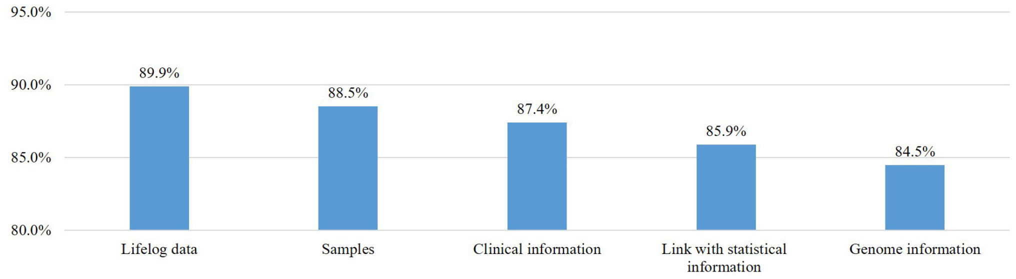


FIGURE 2 | Willingness to provide personal information and samples if asked.

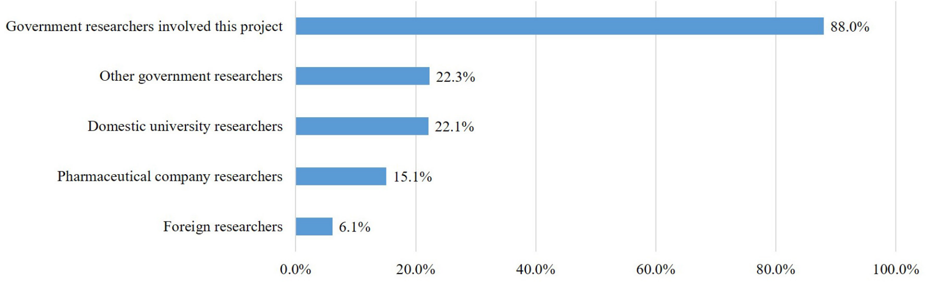


FIGURE 3 | Willingness to share personal information and samples with researchers.

Appropriate Research Institute to Undertake the Study

We asked all respondents about their opinion on what type of research institute would be suitable in initiating the RCP-PMR. The majority of respondents (66.9%) said that it should be carried out by government research institutes. Less than 20% of respondents agreed that the study should be undertaken by government-funded research institutes (19.7%), other non-profit institutes (8.8%), and industry and private research institutes (4.5%).

Using the Collected Samples and Personal Information

The RCP-PMR plans to authorize qualified researchers to use data and specimens collected from cohort participants to perform various research activities. In the questionnaire, we asked about their willingness regarding the range of researchers allowed to use the personal information and samples provided by participants. Most respondents responded negatively to their specimens and information being used by researchers who are not directly involved in the RCP-PMR. The approval rate for their own data being used by the government researchers running this project was quite high (88%), but the approval rates for its use by other government researchers (22.3%), domestic university researchers (22.1%), pharmaceutical researchers (15.1%), and foreign researchers (6.1%) were all low (**Figure 3**).

Participation Benefits

Respondents were asked about the importance of incentives behind their decision of whether to participate. Respondents said incentives are very important (31.4%) or rather important (61.0%). The incentives for participation were receiving health information (80.2%), monetary compensation less than 50,000 KRW (about 42 USD) per year (51.4%), and smart devices (41.3%) (**Figure 4A**). When asked about the information they wanted to receive, laboratory results (cholesterol, blood sugar, etc.) was the highest (73.7%), followed by health information based on family history and genetic testing (67.7%), genetic testing results (66%), health-related research results that use their own information (49.8%), nutrition information (48.9%), health information based on lifelog results (46.7%), and environment-based health information (38.9%) (**Figure 4B**).

Participation in Decision-Making

In large-scale cohort projects that receive a variety of samples and information, it is important to communicate with participants to identify stakeholder needs. Participants were asked whether the opportunity to comment on the project design or operation was important or not, and most respondents (89.7%) said it was important. When we subsequently asked about the phases of the study in which they wanted to participate in decision-making, the rates of participation in each phase were all under 40%. They mainly wanted to be involved in three phases: the questionnaire development and design of personal data collection (39.6%), study participant recruitment (38.9%), and

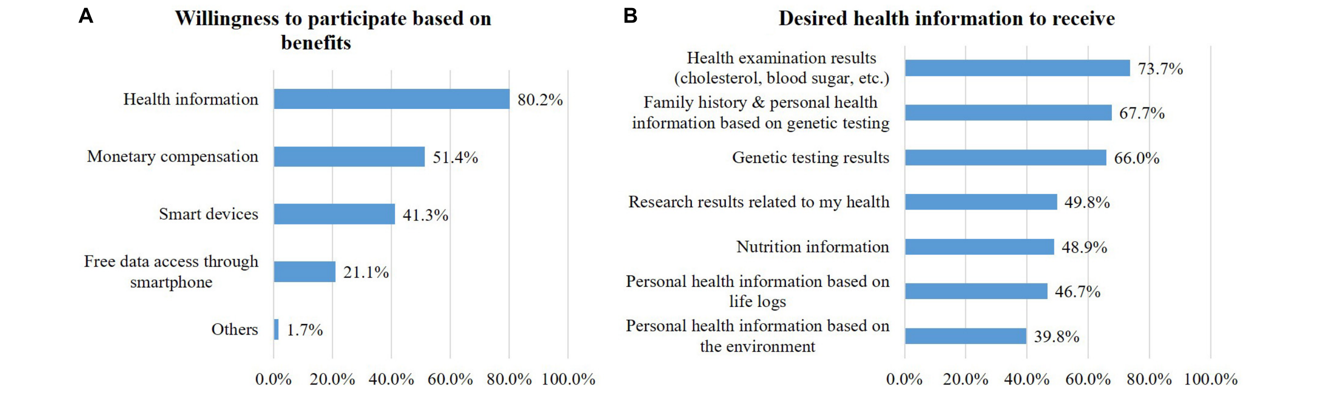


FIGURE 4 | Participation benefits: Willingness to participate based on benefits **(A)** and desired health information to receive **(B)**.

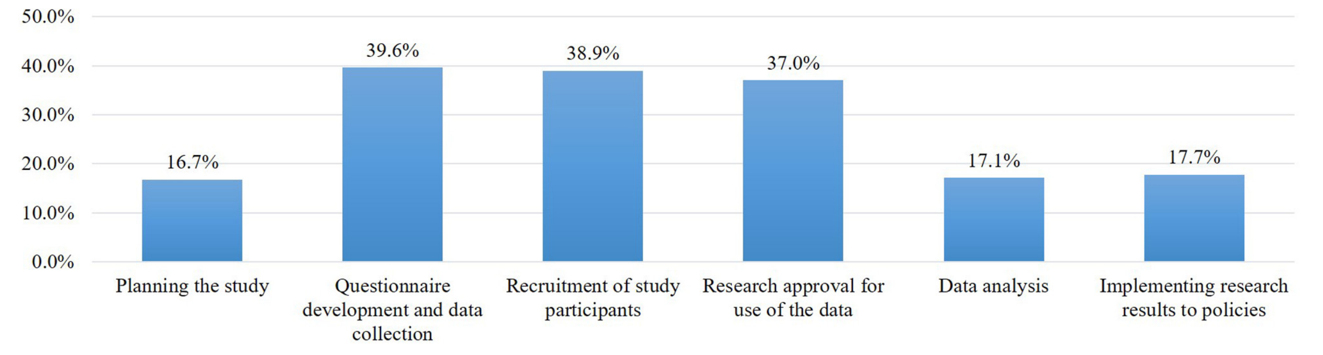


FIGURE 5 | Desired participation in decision making of the study by phase.

deciding which research projects will use the collected data (37.0%) (**Figure 5**).

Change in Perception Regarding Project Participation

We expected that participants who completed the survey would have a better understanding of the project, and this was assumed to have an impact on their willingness to participate in the project. In order to confirm the change in perception regarding project participation, near the end of the survey, respondents were once again asked, “Would you participate in the project?” Of the respondents, 83.5% said they would participate in the project, and their intention to participate in the project was almost unchanged from the beginning to the end of the survey.

DISCUSSION

A large-scale research cohort study is important to ensure that the RCP-PMR develops in a manner that respects public values and interests as well as the instrumental goals of recruiting and retaining participants. We conducted a survey of Korean adults who are potential participants to understand their opinions on the RCP-PMR. Five points of discussion regarding the results analysis are presented below.

The first point is regarding the public attitude toward the citizen participation cohort model for precision medicine. Our survey results showed that 94.5% of participants answered that we need this project, which is 15.5% higher than the Kaufman et al. (2016) study in the United States (Kaufman et al., 2016). The intention to participate is very high (83.5%), which is much higher than that of the Kaufman et al. study (54%) (Kaufman et al., 2016). We can extrapolate that this result is because of the sense of the nationalism of and the familiarity with cutting-edge technology among Koreans. Kim (2007) showed that the majority of participants had very positive attitudes toward the life sciences industry as the most important pillar of the South Korean economy (Kim, 2007) and Okita et al. explained that a sense of responsibility to their families and society may have a positive impact on increasing the willingness to participate in genomics research in Japan (Okita et al., 2018). Bak and Kim (2016) also provided evidence that the higher the trust in scientific expertise, the higher the level of public support for social problem-solving research in South Korea (Bak and Kim, 2016). This positive attitude of the general public also corresponds with the Ishiyama et al. (2008) finding that 69.4% of Japanese participants favored the promotion of genomic studies related to medicine (Ishiyama et al., 2008).

More specifically, this result also showed that a lower level of support for the project is associated with lower education levels, consistent with Okita et al. in Japan and a focus group study in China (Chen et al., 2013; Okita et al., 2018). To establish strategies to recruit and retain participants in the research program, it is essential to obtain more evidence on prejudice against or misconceptions of this project results from low levels of genomic literacy and to explore favorable ways of knowledge translation and transfer in order to identify appropriate health and educational interventions (Etchegary et al., 2013; Nakamura et al., 2017). On the other hand, those who do not have experience with SNS were significantly more likely to agree on the need to implement the study and answered that they would participate in the study if asked. While SNS was known as a potential source for participant recruitment and research data for those who are supportive toward this kind of large-scale precision medicine project (Reaves and Bianchi, 2013), this survey result requires more elaborative strategy to encourage them to use SNS and to improve their knowledge about SNS. The evidence can also support the development of educational and awareness programs to familiarize people with genomics and health beyond the scope of the RCP-PMR (Reaves and Bianchi, 2013).

Second, 24.7% of the respondents who were not willing to participate said they would get involved if they had better protection from privacy leakage and if personalized health services were provided. Gaskell et al. in Europe provided a possible reason for participants' privacy concern that people explicitly expressed that privacy violations are an issue not only in biobanks but also in wider society (Dorey et al., 2018). A recent South Korean governmental report also supports the finding of Gaskell et al. – 4.6% of 4,000 internet users experienced information security incidents, and 97.3% were aware of the importance of personal information security in 2018 (Gaskell et al., 2013). Health information security, such as leakage of personal information in the National Health Insurance Corporation database in the ROK, remains a serious issue (Korean Internet & Security Agency, 2018). Although it is crucial to utilize the RCP-PMR database for data sharing in research and commercial sectors, this emerging public view of privacy protection suggests that a well-established model for privacy protection and communication with the public are needed.

Third, regarding benefits of participation, 80.2% of respondents wanted health information including health examination and genetic testing results as an appropriate compensation for participating in the project. People in South Korea already receive health results from regular medical checkups such as general health exams and cancer screenings, which are covered by the national health insurance program, as they are in Japan, Taiwan, and Singapore. However, the possibility of receiving genetic testing results in particular could be the reason participants preferred a return of health information to other forms of compensation. This public interest in return of results should be specifically implemented in the RCP-PMR.

Next, in terms of public willingness to allow different types of researchers to use their data and samples, a significant gap was found between researchers running the program and researchers outside the program, particularly private companies. This project will be carried out by researchers who are qualified to research various specimens and information provided by the participants. More than 85% of respondents said they would provide a variety of specimens and personal information, and 88.0% of participants agreed that government researchers conducting the project should be able to use their specimens and information. However, they had a negative attitude toward its use by other government researchers, non-government researchers such as private companies, and foreign researchers. The United States survey showed that participants agreed to use various researchers such as researchers at the National Institutes of Health (79%), other government researchers (44%), university researchers in the United States (71%), pharmaceutical or drug company researchers (52%), and university researchers in other countries (39%) (Kaufman et al., 2016). In the ROK, however, only 6.1% – 22.3% of the participants agreed to have their data used by researchers other than the government researchers running the project. This result suggests that sharing the provided samples and information with non-government researchers such as private companies may reduce willingness to participate. For the success of the project, however, it is necessary to communicate with the lay public as well as the RCP-PMR participants to improve their understanding of the purpose of the project and how data sharing can contribute to disease research and prevention.

Finally, in terms of the citizen partnership model in the decision-making process of the RCP-PMR, the findings addressed a gap between the 89.7% of respondents willing to participate generally in the decision-making process of the RCP-PMR and their low response rate of willingness to participate in each individual study phase (16.7–39.6%). We can interpret that researchers in Korea did not consider the engagement of research participants in the decision-making process of biobanking or personalized medicine research, so the lay public were not experienced in how to involve themselves in the decision-making process of this kind of research project. Communication with civil society and patient organizations at various stages of the research would be important in improving understanding and reaching consensus to achieve the goals of precision medicine.

This study also found that around 40% of respondents would want to provide their opinions on questionnaire development and data collection (39.6%), recruitment (38.9%), and approval of research for data use (37.0%). In particular, this finding is consistent with the United States study, which found that the respondents wanted to be involved in helping decide what kinds of research are appropriate (45%) and what to do with the study results (45%) (Kaufman et al., 2016).

LIMITATIONS

The findings of this study have to be seen in light of some important limitations. First, the participants highly supported the RCP-PMR but the results may not reflect actual participation rate of South Korean population. The full response rate from the invitation is 2.9%, so the survey results do not reflect actual willingness to participate. In addition, all invited participants were on panels of a survey firm, so they possibly responded more favorably than the general population.

Second, this survey focused on the social acceptance of the implementation of and participation in the RCP-PMR, meaning that participants' perceptions of what can be derived from this study were not examined. For instance, ethical concerns over sharing genomic data, including that of the patient's family members, is an emerging issue in East Asian countries (Johnsson et al., 2010; Yoshizawa et al., 2017). Families have played significant roles in genetic research, and their value is re-illuminated in the era of genomic medicine. It is important to make progress in data sharing while simultaneously protecting the privacy and interests of patients and families and returning its benefits to them (Yoshizawa et al., 2017). More empirical evidence to identify interrelated and cross-cultural factors of the social acceptance of government-led biomedical research is also required.

In addition, public awareness of the risks of health-related data sharing has not been fully investigated in this research. The perceptions of potential risks of data sharing are influenced by attitudes to genomic data sharing (Takashima et al., 2018). The applications of health-related data sharing on the grounds of research and public interest, without due regard for the perspective of patients and the lay public, could run the risk of fostering distrust toward healthcare data collection (Chen et al., 2013). Further studies to investigate the issues, such as discrimination, are essential to promote voluntary participation by the public.

Finally, our finding that lower levels of education population correspond to lower levels of support for the project is possibly associated with genomics-related literacy. However, this study is limited by the fact that the survey participants do not know how much they understand about the project and genomics and responded to the survey based on the guidance given. Thus, as discussed above, we suggest further empirical studies to evaluate genomics-related literacy as a basis of establishing health and educational strategies.

CONCLUSION

These survey results will influence the policy development of the precision medicine national cohort program in South Korea. We found that the need for the project, the willingness to participate in the project, and the willingness to provide specimens and information to the project are higher than the results of the United States and Japan. This survey also found a strong willingness of the public to participate in the decision-making process of the RCP-PMR. However, more importantly, this survey also revealed that participants who have negative attitudes toward the RCP-PMR are concerned about privacy violations and the majority of participants disagreed with specimen and data sharing with researchers other than the government researchers who run the project.

For the success of this national project, such findings will determine the public engagement policy for precision medicine in South Korea. As a crucial point, the policy will focus on communicating with the general public and patients how the project and sharing of data with other researchers can help healthcare. This project will also establish a system of the governance that respects the opinions of various stakeholders, including civil society organizations, patient groups, and researchers in the project planning and execution.

DATA AVAILABILITY STATEMENT

The datasets generated in this study can be accessed from the corresponding author upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board (IRB) of the Yonsei University. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

HK and HRK equally contributed to framing the research design, analyzing the data, and drafting and revising the manuscript. SK contributed to conducting the survey and drafting the manuscript. EK contributed to the interpretation of the findings. SYK participated in the framing of the initial research design and interpretation of the findings. H-YP continuously provided feedback on conducting the survey and revising the manuscript. All the authors read and approved the final manuscript.

FUNDING

This study was supported by the Research Program funded by the Korea Centers for Disease Control and Prevention (2017NE740100).

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of the KCDC for supporting the survey that led to this publication. We would like to thank Jae-Pil Jeon, National Biobank of Korea, for his valuable opinions. We also appreciate Jihyun Yang, Sujin Hong, Subin Lee, and Jiwon Choi, from the Asian Institute of Health Law and Ethics, Yonsei University for their assistance in the survey process.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00283/full#supplementary-material

REFERENCES

Alzu'bi, A. A., Zhou, L., and Watzlaf, V. J. M. (2019). Genetic variations and precision medicine. *Perspect. Health Inf. Manag.* 16:1a.

Bak, H. J., and Kim, M. (2016). The relationship between public support for scientific research and political orientations: the case of research for social problem-solving. (in Korean). *J. Technol. Innov.* 24, 107–137. doi: 10.14383/SIME.2016.24.3.107

Chen, H., Gottweis, H., and Starkbaum, J. (2013). Public perceptions of biobanks in China: a focus group study. *Biopreserv. Biobank.* 11, 267–271. doi: 10.1089/bio.2013.0016

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/NEJMp1500523

Dorey, C. M., Baumann, H., and Biller-Andorno, N. (2018). Patients data and patient rights: swiss healthcare stakeholders' ethical awareness regarding large patient data sets - a qualitative study. *BMC Med. Ethics* 19:20. doi: 10.1186/s12910-018-0261-x

Etchegary, H., Green, J., Dicks, E., Pullman, D., Street, C., and Parfrey, P. (2013). Consulting the community: public expectations and attitudes about genetics research. *Eur. J. Hum. Genet.* 21, 1338–1343. doi: 10.1038/ejhg.2013.64

Gaskell, G., Gottweis, H., Starkbaum, J., Gerber, M. M., Broerse, J., Gottweis, U., et al. (2013). Publics and biobanks: pan-European diversity and the challenge of responsible innovation. *Eur. J. Hum. Genet.* 21, 14–20. doi: 10.1038/ejhg.2012.104

Genetics Home Reference (2019). *What is Precision Medicine?* Available online at: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition (accessed June 1, 2019).

Genomics England (2019). *The 100,000 Genome Project*. Available online at: http://www.genomicsengland.co.uk/ (accessed June 1, 2019).

Haga, S. B. (2017). Update: looking beyond the 100,000 Genome Project. *Per. Med.* 14, 85–87. doi: 10.2217/pme-2016-0101

Ishiyama, I., Nagai, A., Muto, K., Tamakoshi, A., Kokado, M., Mimura, K., et al. (2008). Relationship between public attitudes toward genomic studies related to medicine and their level of genomic literacy in Japan. *Am. J. Med. Genet.* 146, 1696–1706. doi: 10.1002/ajmg.a.32322

Johnsson, L., Helgesson, G., Rafnar, T., Halldorsdottir, I., Chia, K. S., Eriksson, S., et al. (2010). Hypothetical and factual willingness to participate in biobank research. *Eur. J. Hum. Genet.* 18, 1261–1264. doi: 10.1038/ejhg.2010.106

Kaufman, D. J., Baker, R., Milner, L. C., Devaney, S., and Hudson, K. L. A. (2016). 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

Kim, H., Kim, S. Y., and Joly, Y. (2018). South Korea: in the midst of a privacy reform centered on data sharing. *Hum. Genet.* 136, 627–635. doi: 10.1007/s00439-018-1920-1

Kim, J. S. (2007). A survey research on the social and ethical implication of life science and regenerative medicine. (in Korean). *J. Health Med. Sociol.* 21, 157–196.

Korean Internet & Security Agency (2018). *2017 Survey on Information Security - Summary Report. [Internet]. Ministry of Science and ICT*. Available online at: https://www.kisa.or.kr/eng/usefulreport/surveyReport_List.jsp (accessed August 20, 2019).

Nakamura, S., Narimatsu, H., Katayama, K., Sho, R., Yoshioka, T., Fukao, A., et al. (2017). Effect of genomics-related literacy on non-communicable diseases. *J. Hum. Genet.* 62, 839–846. doi: 10.1038/jhg.2017.50

National Institutes of Health (2019). *All of US Research Program*. Available online at: http://www.allofus.nih.gov (accessed June 2, 2019).

Okita, T., Ohashi, N., Kabata, D., Shintani, A., and Kato, K. (2018). Public attitudes in Japan toward participation in whole genome sequencing studies. *Hum. Genomics* 12:21. doi: 10.1186/s40246-018-0153-7

Peplow, M. (2016). The 100,000 genomes project. *BMJ* 353:i1757. doi: 10.1136/bmj.i1757

Reaves, A. C., and Bianchi, D. W. (2013). The role of social networking sites in medical genetics research. *AJMG* 161, 951–957. doi: 10.1002/ajmg.a.35903

Takashima, K., Maru, Y., Mori, S., Mano, H., Noda, T., and Muto, K. (2018). Ethical concerns on sharing genomic data including patients' family members. *BMC Med. Ethics* 19:61. doi: 10.1186/s12910-018-0310-5

Yoshizawa, G., Sasongko, T. H., Ho, C. H., and Kato, K. (2017). Social and communicative functions of informed consent forms in East Asia and beyond. *Front. Genet.* 8:99. doi: 10.3389/fgene.2017.00099

Conflict of Interest: The 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.

Copyright © 2020 Kim, Kim, Kim, Kim, Kim and Park. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

BRIEF RESEARCH REPORT
published: 30 June 2020
doi: 10.3389/fgene.2020.00619

Workflow for the Implementation of Precision Genomics in Healthcare

*Sanja Mehandziska[1], Aleksandra Stajkovska[2], Margarita Stavrevska[1], Kristina Jakovleva[1], Marija Janevska[2], Rodney Rosalia[1], Ivan Kungulovski[2], Zan Mitrev[1] and Goran Kungulovski[2]**

[1] Zan Mitrev Clinic, Skopje, Macedonia, [2] Bio Engineering LLC, Skopje, Macedonia

OPEN ACCESS

Edited by:
*Manuel Corpas,
Cambridge Precision Medicine,
United Kingdom*

Reviewed by:
*Chiara Fabbri,
King's College London,
United Kingdom
Karyn Megy,
University of Cambridge,
United Kingdom*

Correspondence:
*Goran Kungulovski
goran@bioengineering.mk;
goran.kungulovski@zmc.mk*

Specialty section:
*This article was submitted to
Genomic Medicine,
a section of the journal
Frontiers in Genetics*

Received: *14 October 2019*
Accepted: *21 May 2020*
Published: *30 June 2020*

Citation:
*Mehandziska S, Stajkovska A,
Stavrevska M, Jakovleva K,
Janevska M, Rosalia R, Kungulovski I,
Mitrev Z and Kungulovski G (2020)
Workflow for the Implementation
of Precision Genomics in Healthcare.
Front. Genet. 11:619.
doi: 10.3389/fgene.2020.00619*

To enable the implementation of precise genomics in a local healthcare system, we devised a pipeline for filtering and reporting of relevant genetic information to healthy individuals based on exome or genome data. In our analytical pipeline, the first tier of filtering is variant-centric, and it is based on the selection of annotated pathogenic, protective, risk factor, and drug response variants, and their one-by-one detailed evaluation. This is followed by a second-tier gene-centric deconstruction and filtering of virtual gene lists associated with diseases, and VUS-centric filtering according to ACMG pathogenicity criteria and pre-defined deleteriousness criteria. By applying this filtering protocol, we were able to provide valuable insights regarding the carrier status, pharmacogenetic profile, actionable cardiovascular and cancer predispositions, and potentially pathogenic variants of unknown significance to our patients. Our experience demonstrates that genomic profiling can be implemented into routine healthcare and provide information of medical significance.

Keywords: exome, genome, personalized medicine, precision genomics, clinical practice, implementation

INTRODUCTION

Personalized medicine is a proactive medical approach, which in general seeks to stratify patients in risk groups and tailor treatments, medical decisions, health promotion, or preventive measures according to the individual's omics baseline profile combined with lifestyle and environmental factors (Ashley, 2016). The advent of cost-effective next-generation sequencing (NGS) technologies, such as the sequencing of whole genomes (WGS), whole and clinical exomes (WES and CES), combined with the accumulation of genetic knowledge and easy-to-use bioinformatics tools, has paved the way for genomics-based personalized medicine into clinics (Manolio et al., 2013; Goodwin et al., 2016; Doble et al., 2017; Vassy et al., 2017; Bylstra et al., 2019; Zoltick et al., 2019). Nowadays, these technologies have already started to transform healthcare by enabling precise disease screening, actionable diagnostics, treatment, and management. Despite this, precision genomics has not been fully implemented in the vast majority of healthcare systems yet. In order to facilitate its implementation, practical and user-friendly workflows and pipelines are required.

Abbreviations: ACMG59, incidental findings in 59 genes recommended by American College of Medical Genetics and Genomics; CES, clinical exome sequencing; GWAS, genome-wide association studies; indel, insertion or deletion; SNV, single nucleotide variant; VUS, variants of uncertain significance; WES, whole-exome sequencing; WGS, whole-genome sequencing.

In this study, we primarily aimed to describe a pipeline for balanced CES, WES, or WGS reporting of called variants in healthy individuals interested in proactive genetic testing. A batch of datasets taken from symptomatic patients was included only as a proof-of-principle. This approach of variant filtering helped us to initialize the process of implementing precision genomics in clinical practice at a tertiary healthcare institution (**Figure 1**). By applying this workflow, we were expecting to find actionable variants of clinical relevance or variants that might aid reproductive decisions. Our current experience demonstrates that the implementation of genomic profiling following our filtering pipeline into real-life clinical practice can provide information of medical significance. The pipeline could be used in future systematic and longitudinal studies focusing on the translational aspects of genomic medicine.

MATERIALS AND METHODS

Patients

We evaluated 94 patients meeting our inclusion/exclusion criteria with a median age of 34.5 years (range 2 to 65) of which 51/94 (54%) were males and 43/94 (46%) were females (**Table 1**). All individuals/patients were recruited at the Zan Mitrev Clinic either through regular pro-active check-ups or TV/social media. The sole inclusion criterion for symptomatic patients was a referral from a medical specialist; patients who were not able to provide written informed consent and complete medical history were excluded from the study. In addition, healthy individuals who were unable to provide written informed consent were excluded from the study (**Figure 1A**). The analysis was done according to the workflow described in **Figure 1B**. The vast majority of patients were of Macedonian descent 74/94 (78.72%), followed by Albanian 11/94 (11.70%), Serbian 4/94 (4.25%), American 3/94 (3.19%), Turkish 1/94 (1.06%), and Bulgarian 1/94 (1.06%). The aforementioned protocol was used only as a proof-of-principle for analyzing genetic data from symptomatic patients (n = 15); we only communicated the mutations associated with the clinical phenotype. In contrast, full reports following this protocol were disclosed to all healthy individuals (n = 79). Symptomatic patients were informed that additional unrelated information concerning their carrier status and pharmacogenetic profile could be provided as well. Although the WGS analysis has advantages over WES and CES in the respect of providing more comprehensive and uniform coverage of the whole genome, most of the patients 89/94 (94.7%) underwent WES or CES testing, due to cost-effectiveness.

TABLE 1 | Description of the cohort.

Gender	
Male	51 (54%)
Female	43 (46%)
Healthy/Affected	
Healthy	79 (84%)
Affected	15 (16%)
Age	
<18	9 (10%)
>18	85 (90%)
Method	
CES	29 (31%)
WES	60 (64%)
WGS	5 (5%)

Ethics Statement

Written and signed informed consent for participation and publication of data was obtained from all subjects or their legal guardians (for patients under the age of eighteen) in this study. The ethics committee of the Zan Mitrev Clinic waived the need for IRB approval, deeming written and signed informed consent sufficient.

DNA Extraction, Library Preparation, NGS Sequencing

Around 5 ml of whole blood was collected in K_2-EDTA tubes, following accepted principles for blood drawing and blood collection. DNA was extracted from 400 μl of whole blood in a SaMag-12 automatic nucleic acid extraction system (Sacace Biotechnologies, Como, Italy), yielding between 5 and 15 μg of pure DNA, measured by NanoDrop spectrometry (A260/280 ratio 1.7–1.9). Clinical exome enrichment was carried out by using the TruSight One sequencing panel (Illumina, San Diego, United States) or in-house developed CES enrichment protocol (Sophia Genetics, Saint-Sulpice, Switzerland). Whole exome enrichment was carried out by using the SureSelect Human All Exon V6 kit (Agilent Technologies, Santa Clara, United States) or Human Core Exome kit (Twist Bioscience, San Francisco, United States). The entire wet lab work (DNA QC, enrichment, library preparation, and sequencing) for CES, WES, and WGS were carried out in the Sophia Genetics, Wuxi Nextcode or DNA link, or Beijing Genomics Institute facilities, respectively.

Primary Bioinformatic Analysis

For CES, between 13 and 30 million reads were obtained with a NextSeq machine (Illumina, San Diego, United States), with a coverage of at least 50x for average 81% of all sequences. Sequence quality control was done with FastQC[1], and sequences were mapped to hg19 with BWA (Li and Durbin, 2009). single nucleotide variant (SNV) and indel calling, together with advanced variant annotation, were done with the Sophia DDM platform (Sophia Genetics, Saint-Sulpice, Switzerland).

For WES and WGS, between 40 and 120 million reads or ∼950 million were obtained with a HiSeq X–10 machine or NovaSeq 6000 (Illumina, San Diego, United States), respectively. The coverage of WES or WGS was >75x or >40x, respectively. Alignment, variant calling, and annotation were done on the Genoox platform (Palo Alto, United States).

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

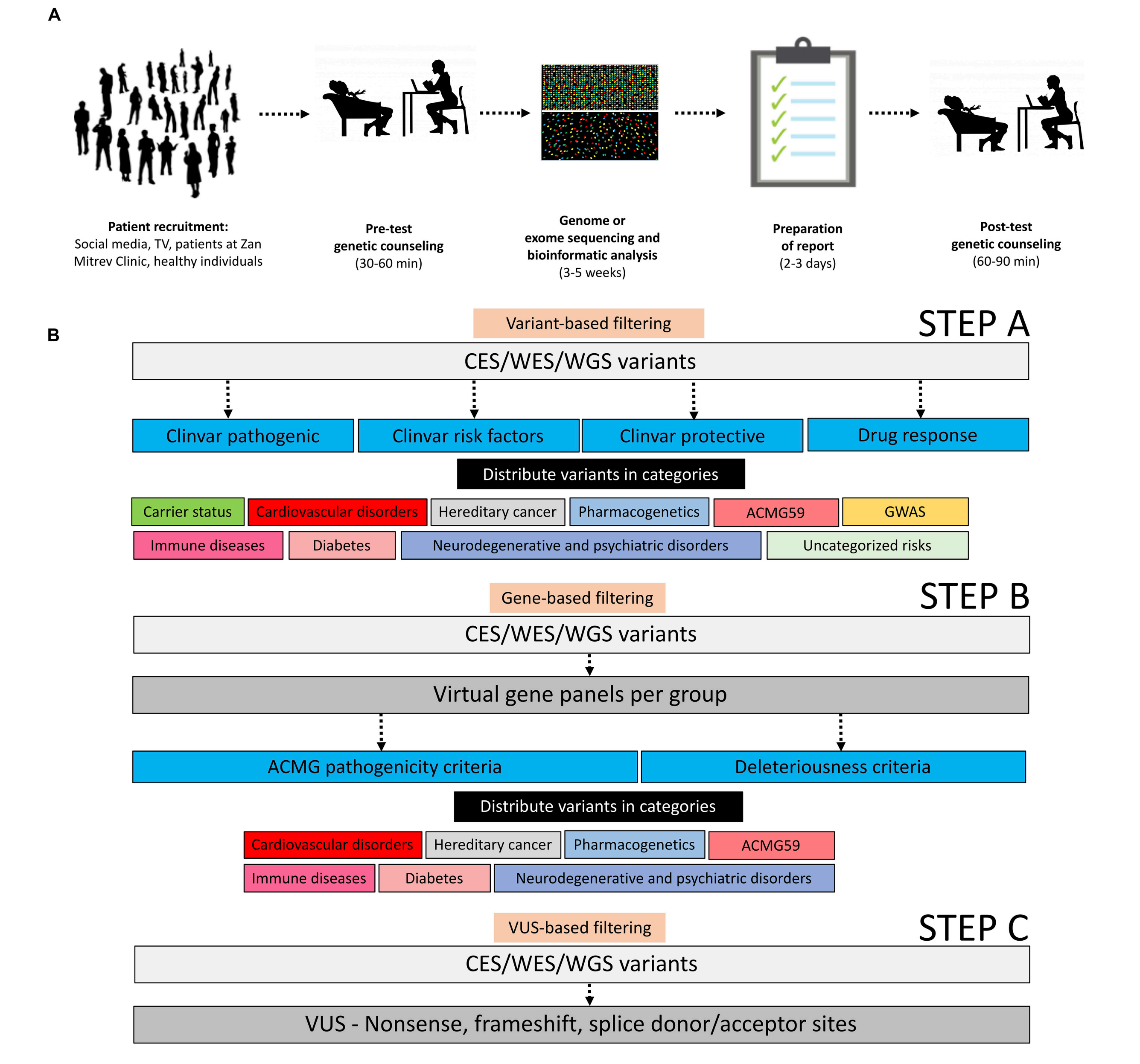

A
Patient recruitment:
Social media, TV, patients at Zan Mitrev Clinic, healthy individuals
Pre-test genetic counseling (30-60 min)
Genome or exome sequencing and bioinformatic analysis (3-5 weeks)
Preparation of report (2-3 days)
Post-test genetic counseling (60-90 min)
B
Variant-based filtering
STEP A
CES/WES/WGS variants
Clinvar pathogenic
Clinvar risk factors
Clinvar protective
Drug response
Distribute variants in categories
Carrier status
Cardiovascular disorders
Hereditary cancer
Pharmacogenetics
ACMG59
GWAS
Immune diseases
Diabetes
Neurodegenerative and psychiatric disorders
Uncategorized risks
Gene-based filtering
STEP B
CES/WES/WGS variants
Virtual gene panels per group
ACMG pathogenicity criteria
Deleteriousness criteria
Distribute variants in categories
Cardiovascular disorders
Hereditary cancer
Pharmacogenetics
ACMG59
Immune diseases
Diabetes
Neurodegenerative and psychiatric disorders
VUS-based filtering
STEP C
CES/WES/WGS variants
VUS - Nonsense, frameshift, splice donor/acceptor sites

FIGURE 1 | Operational and analytical workflows applied in this study. **(A)** Operational workflow starting with patient recruitment, pre-test genetic counseling, genome, or exome sequencing and bioinformatic analyses, representative sample of reports, and post-test genetic counseling. All adapted images used in this figure have a CC license. **(B)** Description of the analytical workflow for variant filtration.

All CES, WES, and WGS variant lists were additionally annotated with Annovar, which provides more annotation notes than Sophia Genetics and Genoox (Yang and Wang, 2015). All detected variants were taken into consideration in the subsequent filtering steps.

Secondary Bioinformatics Analysis, Filtering, and Interpretation of Variants

The list of high confidence annotated variants was downloaded directly from the Sophia DDM/Genoox platforms in a.txt or.csv format and analyzed further in a spreadsheet program such as Microsoft Excel. In the primary variant-based selection step the entire list of variants was filtered (either in Excel or in-platform) based on ClinVar terms “pathogenic,” “protective,” “risk factor,” and “drug response” followed by manual curation, manual filtration, and manual function attribution, and then distributed in the following categories: carrier status, cardiovascular disorders, hereditary cancer, pharmacogenetics, ACMG59 (Kalia et al., 2017), immune diseases, diabetes, neurodegenerative and psychiatric disorders, uncategorized risks, and genome-wide association studies (GWAS; MacArthur et al., 2017).

As a secondary gene-based filtering approach, by using the "virtual panel" capability within the Sophia DDM/Genoox platforms, we have created an array of standardized virtual gene panels encompassing genes associated with (1) cardiovascular disorders, (2) hereditary cancer, (3) neurodegenerative and psychiatric disorders, (4) diabetes, (5) immune diseases, and (6) ACMG59 genes (Kalia et al., 2017). The virtual gene lists per panel can be found in **Supplementary File S1**. Variants were selected from these virtual panels based on ACMG pathogenicity criteria (Richards et al., 2015) or computationally defined deleteriousness criteria (SIFT, Polyphen2, Mutation Taster, Mutation Assessor, FATHMM, dbscSNV Ada, GERP, GeneCanyon, and fitCons), and distributed in the above-mentioned categories.

Finally, in a third filtering variants of uncertain significance (VUS)-based approach, we selected all the non-sense/frameshift VUS in exons/splice donor-acceptor sites, and we reported them in the annex without further interpretation in order to make sure their significance is reassessed in the future when their function is determined (**Figure 1B**).

All selected variants were evaluated (selected or discarded) according to information in ClinVar and literature[2]. Also, further evaluation in other databases such as the human gene mutation database[3], CentoMD database (Trujillano et al., 2017), and Clinvitae[4] was carried out. Pathogenic variants, especially for medium and high penetrance alleles, were interpreted according to the latest available literature and ClinGen guidelines[5]. Disease risk for non-Mendelian, lower-penetrance variants associated with common diseases was assessed based on GWAS. Typically, the odds ratio or relative risk was reported, or in rare cases P-value or chi-square statistic, respectively.

To increase the detection rate and minimize the rate of false positives, first, we applied the entire protocol for all detected variants (low and high confidence), and then we did the same only for high confidence retained variants. Retained variants had a quality score >100, read depth >10, and quality by depth >10. To further reduce the level of false-positive variants, we retained only variants detected by multiple variant callers (GATK Haplotype Caller and FreeBayes). Alleles with a representation of >25% of the total read coverage were defined as heterozygous. All discrepancies were solved by additional manual evaluation.

Patient Reports

Concerning the bioinformatics strategies for data analysis, filtering, and interpretation, we wanted to strike a balance between under- and over-reporting of variants; in other words, we wanted to maximize the benefit of the provided genetic analysis while reducing the costs and unnecessary follow-ups. Patient reports were divided into two sections, main report and annex, following the same line of reasoning as described previously (McLaughlin et al., 2014). The main report is typically 5–6 pages long and encompasses all the major findings (typically related to medium and high penetrance diseases), patient information, and methodology in a clear and concise language. In addition to the main report, we attached an annex consisting of 20–30 pages of all the selected (relevant) variants assigned to the above categories, in a tabular form accompanied with additional information such as:

- An in-depth reference to studies and major findings of the studies, especially for high penetrance alleles
- Odds ratio/relative risk/chi-square/p-value for common alleles (if available)
- Name of gene
- Type of variant (SNV, indel, etc.)
- Functional consequence (non-sense, missense, etc.)
- Genomic region (exonic, intronic, 5'UTR, etc.)
- Chromosome and chromosomal coordinates
- dbSNP or ClinVar rsID
- Population frequency (G1000, ExAC, esp5400)
- ClinVar signature (pathogenic, benign, drug response, etc.). This section reports the overall interpretation of a variation based on aggregating data from submitters.
- Inheritance (autosomal recessive, autosomal dominant, etc.)
- Level of evidence in the pharmacogenetics section (based on[6] grading system)

Genetic Counseling

All the patients underwent pre- and post-test genetic counseling. At the time of pre-test genetic counseling patients were informed about the potential implications of the genetic results to themselves and their families. We discuss the pros and cons of the test, and the current state of genetic/genomic research, as well as the basic principles of inheritance and penetrance. Following the test, we held in-depth discussions where the patients were reacquainted with the basics of DNA biology, genetic variants, types of inheritance, penetrance, and implications. With regard to common medical conditions and diseases, patients were told that these conditions are multifactorial and may include other known or unknown genetic, lifestyle, or environmental components. The genetic counselors stressed the meaning of phrases "no known pathogenic mutation causing/associated with [name of condition]" and the meaning and gravity of known pathogenic mutations.

Prior to testing, all of the patients signed written informed consent. All patients having actionable variants underwent a further examination or consultation with a relevant specialist. Hence, whenever mentioned further on that the patient was advised for a specific medical procedure, the advice came from relevant specialists and not the genetic counselors alone.

RESULTS

In our efforts to provide the most relevant genomic information to our patients, we divided variants into two categories: variants

[2] http://clinvar.com/

[3] http://www.hgmd.cf.ac.uk/

[4] http://clinvitae.invitae.com/

[5] https://www.clinicalgenome.org/

[6] www.pharmgkb.org

of known significance (VKS) obtained from the filtering steps A and B and VUS obtained from the filtering step C. Out of the VKS basket, we only selected variants with a defined clinical value (e.g., high penetrance, level 1a/1b pharmacogenetic association, direct involvement in a clinically relevant pathway). Out of the VUS basket, we reported only variants fulfilling ACMG pathogenicity criteria in medium and high penetrance disease-causing genes. For better reporting, every filtered variant was distributed in a fitting category (**Figure 1B**).

Evaluation of Known Variants

Carrier Status of Rare Diseases

By selecting genetic variants annotated as "pathogenic" or "likely pathogenic" in the ClinVar database, supplemented by further manual curation, available literature evaluation, and filtration of low-frequency variants, the rare disease carrier status of each patient for known variants was derived. The vast majority of patients (96.2%) were carriers of at least one known rare disorder/condition with some of them carrying multiple pathogenic variants, median = 4 (**Supplementary File S2**).

Drug Response

Regarding "drug response," the list of variants was interpreted using information from the pharmGKB[7] database. Only level 1A and level 1B clinical annotations were added to the main report, while the rest of them remained in the annex. Most of the patients were carriers of multiple Level 1A/1B variants. For example, we observed that 34/94 (36.2%) patients were "poor/intermediate *CYP2D6* metabolizers," which is relevant for the metabolism of many drugs, including anti-depressants, opioids, and tamoxifen. In order to reduce the misclassification rate of CYP2D6 metabolizers, we are currently implementing approaches for the detection of *CYP2D6* (and other genes) copy number variants from WES data. In addition, 23/94 (24.5%) patients were "poor/intermediate *CYP2C19* metabolizers" highly relevant for antiplatelet therapy with clopidogrel, in line with observations from our cohort of >3,000 patients (Klinceva et al., 2018). Similarly, 26/94 (27.7%) patients had a risk of statin-induced myopathy, with 2/26 (7.7%) being of a very high risk of myopathy and rhabdomyolysis as well (**Supplementary File S2**), in line with our internal observations from a cohort of >1,500 patients.

Actionable Variants Involved in Cardiovascular Diseases and Cancer

According to our protocol, we proceeded with the analysis of "actionable" variants (**Table 2** and **Supplementary File S2**). Interestingly, more than one third of healthy individuals were carriers of pathogenic/potentially pathogenic variants leading to different types of arrhythmias and hereditary cancer, which are known to have incomplete penetrance.

For instance, patient 2 was a carrier of a pathogenic variant (c.566G > T, p.Arg189Ile; rs199473381) in the *KCNJ2* gene, which has been associated with congenital long QT syndrome (Goldenberg and Moss, 2008). Follow-up EKG revealed visible abnormalities in the heart rhythm and the patient underwent further diagnostics. Patient 20 harbored a rare pathogenic/likely pathogenic variant (c.839C > T, p.Ala280Val; rs72552291) in the *GPD1L* gene that has been shown to decrease inward SCN5A Na + current and cause Brugada syndrome (Pfahnl et al., 2007). The patient underwent a regular cardiac exam, and EKG showed no visible abnormalities; since the patient was taking lithium therapy (which could conceivably unmask Brugada syndrome), advice was given to discuss this with their clinical psychiatrist and cardiologist. Furthermore, in patient 32 we discovered a variant (c.253G > A, p.Asp85Asn, and rs1805128) in the *KCNE1* gene that has been reported to be associated with long QT syndrome (Paulussen et al., 2004). During the post-test genetic counseling, the patient disclosed that a member of their close family had passed away "due to complications from arrhythmia." Other patients were also found to carry variants associated with long QT or other channelopathies; patient 12 and their parent carried a pathogenic variant (c.914G > C; p.Trp305Ser; rs120074186) in the *KCNQ1* gene, whilst in patient 30 we identified a potentially pathogenic variant (c.5434C > T; p.Arg1812Trp; rs121912706) in the *ANK2* gene associated with sudden death of the young (Methner et al., 2016). Finally, in patient 43 and patient 56 we found a combination of variants (Marburg I and *F5* Leiden, *F2* and *F5* Leiden, respectively) that might significantly increase the risk of thrombosis (Voorberg et al., 1994; Poort et al., 1996; Hoppe et al., 2005). All patients were advised to consult a specialist and conduct follow-up studies if deemed necessary.

In regard to cancer, in patient 52, we discovered the presence of variants (c.1437_1439delGGA, p.Glu480del, and rs587778541) in the *MUTYH* gene and (c.470T > C, p.Ile157Thr, and rs1787996) in the *CHEK2* gene. The same *CHEK2* mutation was detected in patient 92. The variant in *MUTYH* is pathogenic and leads to MUTYH-Associated Polyposis in a recessive manner. The presence of the variant in a heterozygous format might slightly (1.5 times) increase the risk of colorectal cancer (Nielsen et al., 1993). The *CHEK2* variant has been reported to increase the risk of different types of cancer 2–3 times (Han et al., 2013). The patient was advised to consult a specialist and discuss a screening protocol. Next, patient 84 harbored the variant (c.511A > G, p.Ile171Val, and rs61754966) in the *NBN* gene, which is a low penetrance risk factor for cancer development (Gao et al., 2013); the patient reported having a family history of breast and pancreatic cancer. Finally, we detected a potentially pathogenic variant (c.3920T > A, p.Ile1307Lys, and rs1801155) in the *APC* gene in patient 91 (Leshno et al., 2016), who is currently undergoing follow-up diagnostics. All patients were advised to consult a specialist and conduct follow-up studies if deemed necessary.

Evaluation of Potentially Pathogenic Variants of Uncertain Significance (VUS)

By analyzing VUS with rare population frequency meeting ACMG pathogenicity criteria (Richards et al., 2015) we uncovered many VUS in medium- or high-penetrance genes (**Supplementary File S2**). For instance, patient 56 is a

[7] https://www.pharmgkb.org/

TABLE 2 | Representative list of patients with actionable variants.

Patient	Associated conditions with gene	Gene (ClinVar)	Nucleotide	Protein	rsID	Clinvar signature (interpretation)
2	Congenital long QT syndrome	*KCNJ2*	c.566G > T	p.Arg189Ile	rs199473381	Likely pathogenic(1)
12	Long QT syndrome	*KCNQ1*	c.914G > C	p.Trp305Ser	rs120074186	Pathogenic(2);Likely pathogenic(1)
29	Hereditary pancreatitis	*SPINK1*	c.101A > G	p.Asn34Ser	rs17107315	Risk factor(2);Pathogenic(4);Uncertain significance(3)
	Cystic fibrosis; Hereditary pancreatitis	*CFTR*	c.3154T > G	p.Phe1052Val	rs150212784	Likely pathogenic(3);Pathogenic(2);Uncertain significance(4);Drug-response(1)
30	Cardiac arrhythmia; Long QT syndrome	*ANK2*	c.11716C > T	p.Arg3906Trp	rs121912706	Likely benign(4);Pathogenic(2);Uncertain significance(2)
32	Malignant tumor of prostate; Hereditary cancer-predisposing syndrome	*MSR1*	c.877C > T	p.Arg293*	rs41341748	Pathogenic(1);Uncertain significance(3);Benign(1)
	Hereditary prostate cancer	*RNASEL*	c.793G > T	p.Glu265*	rs74315364	Pathogenic(1);Likely benign(1);Uncertain significance(1)
	Long QT syndrome	*KCNE1*	c.253G > A	p.Asp85Asn	rs1805128	Benign(5);Likely benign(5);risk factor(3);Pathogenic(1);Likely pathogenic(1);Uncertain significance(2)
43	Thrombophilia, Thyroid cancer	*HABP2*	c.1601G > A	p.Gly534Glu	rs7080536	Risk factor(2);Likely benign(1);Benign(1)
	Thrombophilia	*F5*	c.A1601G	p.Q534R	rs6025	Pathogenic(4);Risk factor(4);Benign(1)
52	Hereditary cancer risk	*CHEK2*	c.470T > C	p.Ile157Thr	rs17879961	Likely pathogenic(8);Pathogenic(9);Risk factor(3);Uncertain significance(2)
	MYH-associated polyposis; Hereditary cancer-predisposing syndrome	*MUTYH*	c.1437_1439delGGA	p.Glu480del	rs587778541	Pathogenic(14)
56	Prothrombin deficiency, congenital; Thrombophilia	*F2*	c.*97G > A		rs1799963	Pathogenic(4);Risk factor(4)
	Thrombophilia	*F5*	c.A1601G	p.Q534R	rs6025	Pathogenic(4);Risk factor(4);Benign(1)
81	Hereditary cancer-predisposing syndrome	*RAD50*	c.2801del	p.Asn934fs	rs748536322	Pathogenic(1)
84	Hereditary cancer-predisposing syndrome	*NBN*	c.511A > G	p.Ile171Val	rs61754966	Benign(3);Likely benign(1);Uncertain significance(11);Pathogenic(1);Risk factor(1)
91	Familial adenomatous polyposis	*APC*	c.3920T > A	p.Ile1307Lys	rs1801155	Likely benign(1);Likely pathogenic(3);Pathogenic(1);Uncertain significance(10);Risk factor(9)
92	Hereditary cancer risk	*CHEK2*	c.470T > C	p.Ile157Thr	rs17879961	Likely pathogenic(8);Pathogenic(9);Risk factor(3);Uncertain significance(2)
94	Prothrombin deficiency, congenital; Thrombophilia	*F2*	c.*97G > A		rs1799963	Pathogenic(4);Risk factor(4)

The full table is given in ***Supplementary File S2****. The asterisk denotes the variant in Clinvar https://www.ncbi.nlm.nih.gov/clinvar/variation/13310/.*

carrier of (c.2423A > G, p.Tyr808Cys; rs746368140) in the *TGFBR3* gene. The involvement of the TGF-beta pathway has been reported in pathologies such as familial thoracic aortic aneurysm and dissection (familial TAAD; Milewicz and Regalado, 1993). The father of patient 56 was diagnosed with a thoracic and abdominal aortic aneurysm and underwent valve-sparing root replacement (Tirone-David procedure). The patient was advised to follow regular cardiovascular check-ups. Patient 75 is a carrier of VUS (c.1755dupA, p.Glu586fs, and rs751465048) in the *MLH3* gene, which is part of the MMR machinery associated with Lynch syndrome (Peltomaki, 2003). The patient already had benign tumors removed from their breast and nose, in the past. The patient was advised to consult a specialist. Finally, patient 84 is a carrier of (c.3145G > A, p.Gly1049Ser, and rs778181932) in the *FBN1* gene, which could be possibly associated with TAAD; the patient reported a history of sudden death in their close family.

DISCUSSION

The central tenet of personalized medicine is proactive care of patients based on the combined information and insights provided by omics approaches, lifestyle, environmental factors, and family history. To aid the implementation of precision genomics locally into our hospital, we have outlined a workflow centered around filtering, stratification in groups, and interpretation of genetic variants that can be readily applied in any genetic lab. By applying these strategies of variant-centric, gene-centric, and VUS-centric filtering, we were able to peer into the genetic constitution of 94 patients and make initial assessments of their carrier status, pharmacogenetics profile, and genetic risk of developing rare and common disorders.

Our experience demonstrates that the implementation of genomic profiling into real-life clinical practice can provide molecular and physiological information of medical significance, although many challenges remain to be addressed (Carter and He, 2016). To begin, serious efforts should be made to improve the knowledge of physicians and raise awareness for patients and the general public about the benefits and pitfalls of pre-emptive genomic testing, especially in the context of the current genetic knowledge. Second, standardization and defined guiding principles are necessary for both the technical and interpretational side of genomics in medicine. A list of guidelines (benchmarks) should be set for the minimal quality and coverage of sequencing data. For instance, currently WES is the most cost-effective approach, but it is limited to the protein-coding regions of the genome; as WGS sequencing costs continue to plummet this will likely lead to a rise in the popularity of WGS, which generates more uniform coverage of both coding and non-coding regions of the genome, relevant for monogenic as well as polygenic disorders. In addition, standardized algorithms for variant calling in clinical settings should be recommended. Moreover, more standardized approaches for filtering and distillation of relevant information, especially methods for calculation of polygenic scores, as well as balanced reporting of valuable information and VUS, and support tools for clinical interpretation, should be designed (Carter and He, 2016). Third, our analytical workflow based on filtering and virtual gene panels is readily applicable but still has a lot of space for improvement. For example, the virtual gene lists should undergo a process of constant curation and improvements from experts in the relevant subspecialties in order to get better informed, non-redundant, and more optimal lists of genes. Another limitation is that our focused study did not provide insights in regard to the cost-effectiveness of genetic testing, as well as the perceived value by both physicians and patients in a controlled and systematic manner. In order to objectively quantify the value of proactive genetic testing, longitudinal follow-up approaches are necessary.

Finally, many complex diseases, such as diabetes, cancer, and some neurological, cardiovascular, and psychiatric disorders, likely involve a large number of different genes and environmental factors (Hindorff et al., 2009; Ashley et al., 2010; De La Vega and Bustamante, 2018; Torkamani et al., 2018). These caveats sometimes might lead to unnecessary follow-up diagnostic measures and wastefulness of resources. Currently, the greatest value of genomic approaches lies in the detection of lower frequency moderate to high penetrance variants, which are easier to interpret and are better characterized due to their more resonant effects (Doble et al., 2017).

In conclusion, by establishing a balanced filtering pipeline, we set the foundation for the integration of genomics in mainstream clinical practice. The valuable insights and experiences we have obtained can have a bearing in future systematic and longitudinal follow-up studies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in Zenodo doi: 10.5281/zenodo.3830498 and url: https://doi.org/10.5281/zenodo.3830498.

ETHICS STATEMENT

Written and signed informed consent for participation and publication of data was obtained from all subjects or their legal guardians in this study. Written and signed informed consent for participation and publication of data was obtained from the legal guardians of all patients under the age of eighteen. The ethics committee of the Zan Mitrev Clinic waived the need for IRB approval, deeming written, and signed informed consent sufficient.

AUTHOR CONTRIBUTIONS

GK conceived and designed the study. GK, SM, AS, MS, KJ, and MJ analyzed and interpreted the data. SM and AS participated in the genetic counseling process. IK and ZM contributed to the recruitment of patients and contributed intellectually. GK and RR wrote the manuscript. All authors contributed to the improvement of the manuscript and read the final version of the manuscript.

FUNDING

This study was carried out as part of the routine clinical work at the Zan Mitrev Clinic.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00619/full#supplementary-material

FILE S1 | Contains lists of genes for each category of virtual panels.

FILE S2 | Contains variant information related to patient information, methodology, rare diseases, drug-response, actionable conditions, and VUS.

REFERENCES

Ashley, E. A. (2016). Towards precision medicine. *Nat. Rev. Genet.* 17, 507–522. doi: 10.1038/nrg.2016.86

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., et al. (2010). Clinical assessment incorporating a personal genome. *Lancet* 375, 1525–1535. doi: 10.1016/S0140-6736(10)60452-7

Bylstra, Y., Davila, S., Lim, W. K., Wu, R., Teo, J. X., Kam, S., et al. (2019). Implementation of genomics in medical practice to deliver precision medicine for an Asian population. *NPJ Genom. Med.* 4:12. doi: 10.1038/s41525-019-0085-8

Carter, T. C., and He, M. M. (2016). Challenges of identifying clinically actionable genetic variants for precision medicine. *J. Healthc Eng.* 2016, 3617572. doi: 10.1155/2016/3617572

De La Vega, F. M., and Bustamante, C. D. (2018). Polygenic risk scores: a biased prediction? *Genome Med.* 10:100. doi: 10.1186/s13073-018-0610-x

Doble, B., Schofield, D. J., Roscioli, T., and Mattick, J. S. (2017). Prioritising the application of genomic medicine. *NPJ Genom. Med.* 2:35. doi: 10.1038/s41525-017-0037-0

Gao, P., Ma, N., Li, M., Tian, Q. B., and Liu, D. W. (2013). Functional variants in NBS1 and cancer risk: evidence from a meta-analysis of 60 publications with 111 individual studies. *Mutagenesis* 28, 683–697. doi: 10.1093/mutage/get048

Goldenberg, I., and Moss, A. J. (2008). Long QT syndrome. *J. Am. Coll Cardiol.* 51, 2291–2300. doi: 10.1016/j.jacc.2008.02.068

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Han, F. F., Guo, C. L., and Liu, L. H. (2013). The effect of CHEK2 variant I157T on cancer susceptibility: evidence from a meta-analysis. *DNA Cell Biol.* 32, 329–335. doi: 10.1089/dna.2013.1970

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

Hoppe, B., Tolou, F., Radtke, H., Kiesewetter, H., Dorner, T., and Salama, A. (2005). Marburg I polymorphism of factor VII-activating protease is associated with idiopathic venous thromboembolism. *Blood* 105, 1549–1551. doi: 10.1182/blood-2004-08-3328

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). 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. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Klinceva, M., Mehandziska, S., Idoski, E., Stajkovska, A., Stavrevska, M., Jakovleva, K., et al. (2018). CYP2C19-dependent clopidogrel resistance is a possible cause for coronary stent and peripheral bypass occlusion. *Eur. Heart J.* 39:ehy564.961.

Leshno, A., Shapira, S., Liberman, E., Kraus, S., Sror, M., Harlap-Gat, A., et al. (2016). The APC I1307K allele conveys a significant increased risk for cancer. *Int. J. Cancer* 138, 1361–1367. doi: 10.1002/ijc.29876

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (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

Manolio, T. A., Chisholm, R. L., Ozenberger, B., Roden, D. M., Williams, M. S., Wilson, R., et al. (2013). Implementing genomic medicine in the clinic: the future is here. *Genet. Med.* 15, 258–267. doi: 10.1038/gim.2012.157

McLaughlin, H. M., Ceyhan-Birsoy, O., Christensen, K. D., Kohane, I. S., Krier, J., Lane, W. J., et al. (2014). A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Med. Genet.* 15:134. doi: 10.1186/s12881-014-0134-1

Methner, D. N., Scherer, S. E., Welch, K., Walkiewicz, M., Eng, C. M., Belmont, J. W., et al. (2016). Postmortem genetic screening for the identification, verification, and reporting of genetic variants contributing to the sudden death of the young. *Genome Res.* 26, 1170–1177. doi: 10.1101/gr.195800.115

Milewicz, D. M., and Regalado, E. (1993). "Heritable thoracic aortic disease overview," in *GeneReviews*®, eds M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. Stephens, et al. (Seattle, WA: University of Washington, Seattle).

Nielsen, M., Lynch, H., Infante, E., and Brand, R. (1993). "MUTYH-Associated Polyposis," in *GeneReviews*®, eds M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. Stephens, et al. (Seattle, WA: University of Washington, Seattle).

Paulussen, A. D., Gilissen, R. A., Armstrong, M., Doevendans, P. A., Verhasselt, P., Smeets, H. J., et al. (2004). Genetic variations of KCNQ1, KCNH2, SCN5A, KCNE1, and KCNE2 in drug-induced long QT syndrome patients. *J. Mol. Med.* 82, 182–188. doi: 10.1007/s00109-003-0522-z

Peltomaki, P. (2003). Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J. Clin. Oncol.* 21, 1174–1179. doi: 10.1200/JCO.2003.04.060

Pfahnl, A. E., Viswanathan, P. C., Weiss, R., Shang, L. L., Sanyal, S., Shusterman, V., et al. (2007). A sodium channel pore mutation causing Brugada syndrome. *Heart Rhythm.* 4, 46–53. doi: 10.1016/j.hrthm.2006.09.031

Poort, S. R., Rosendaal, F. R., Reitsma, P. H., and Bertina, R. M. (1996). A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. *Blood* 88, 3698–3703. doi: 10.1182/blood.v88.10.3698.bloodjournal88103698

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19, 581–590. doi: 10.1038/s41576-018-0018-x

Trujillano, D., Oprea, G. E., Schmitz, Y., Bertoli-Avella, A. M., Abou Jamra, R., and Rolfs, A. (2017). A comprehensive global genotype-phenotype database for rare diseases. *Mol. Genet. Genomic Med.* 5, 66–75. doi: 10.1002/mgg3.262

Vassy, J. L., Christensen, K. D., Schonman, E. F., Blout, C. L., Robinson, J. O., Krier, J. B., et al. (2017). The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. *Ann. Intern. Med.* 167, 159–169. doi: 10.7326/M17-0188

Voorberg, J., Roelse, J., Koopman, R., Buller, H., Berends, F., ten Cate, J. W., et al. (1994). Association of idiopathic venous thromboembolism with single point-mutation at Arg506 of factor V. *Lancet* 343, 1535–1536. doi: 10.1016/s0140-6736(94)92939-4

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10, 1556–1566. doi: 10.1038/nprot.2015.105

Zoltick, E. S., Linderman, M. D., McGinniss, M. A., Ramos, E., Ball, M. P., Church, G. M., et al. (2019). Predispositional genome sequencing in healthy adults: design, participant characteristics, and early outcomes of the PeopleSeq Consortium. *Genome Med.* 11:10. doi: 10.1186/s13073-019-0619-9

Conflict of Interest: GK, AS, MJ, and IK are currently employed by the company Bio Engineering LLC. SM, RR, and ZM are currently employed by the private Zan Mitrev Clinic. MS was formerly employed by the Zan Mitrev Clinic. KJ did an internship at the Zan Mitrev Clinic.

Copyright © 2020 Mehandziska, Stajkovska, Stavrevska, Jakovleva, Janevska, Rosalia, Kungulovski, Mitrev and Kungulovski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fgene.2020.00578

Impute.me: An Open-Source, Non-profit Tool for Using Data From Direct-to-Consumer Genetic Testing to Calculate and Interpret Polygenic Risk Scores

Lasse Folkersen[1]*, Oliver Pain[2], Andrés Ingason[1], Thomas Werge[1†], Cathryn M. Lewis[2,3†] and Jehannine Austin[4,5†]

[1] Institute of Biological Psychiatry, Mental Health Centre Sankt Hans, Copenhagen, Denmark, [2] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom, [3] Department of Medical & Molecular Genetics, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom, [4] Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada, [5] Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

OPEN ACCESS

Edited by:
*Gustavo Glusman,
Institute for Systems Biology (ISB),
United States*

Reviewed by:
*Hervé Michel Chneiweiss,
Centre National de la Recherche
Scientifique (CNRS), France
S. Hong Lee,
University of South Australia, Australia*

***Correspondence:**
*Lasse Folkersen
lasse.folkersen@regionh.dk;
lassefolkersen@gmail.com*

† These authors have contributed equally to this work

Specialty section:
*This article was submitted to
ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics*

Received: *29 November 2019*
Accepted: *11 May 2020*
Published: *30 June 2020*

Citation:
Folkersen L, Pain O, Ingason A, Werge T, Lewis CM and Austin J (2020) Impute.me: An Open-Source, Non-profit Tool for Using Data From Direct-to-Consumer Genetic Testing to Calculate and Interpret Polygenic Risk Scores. Front. Genet. 11:578. doi: 10.3389/fgene.2020.00578

To date, interpretation of genomic information has focused on single variants conferring disease risk, but most disorders of major public concern have a polygenic architecture. Polygenic risk scores (PRSs) give a single measure of disease liability by summarizing disease risk across hundreds of thousands of genetic variants. They can be calculated in any genome-wide genotype data-source, using a prediction model based on genome-wide summary statistics from external studies. As genome-wide association studies increase in power, the predictive ability for disease risk will also increase. Although PRSs are unlikely ever to be fully diagnostic, they may give valuable medical information for risk stratification, prognosis, or treatment response prediction. Public engagement is therefore becoming important on the potential use and acceptability of PRSs. However, the current public perception of genetics is that it provides "yes/no" answers about the presence/absence of a condition, or the potential for developing a condition, which in not the case for common, complex disorders with polygenic architecture. Meanwhile, unregulated third-party applications are being developed to satisfy consumer demand for information on the impact of lower-risk variants on common diseases that are highly polygenic. Often, applications report results from single-nucleotide polymorphisms (SNPs) and disregard effect size, which is highly inappropriate for common, complex disorders where everybody carries risk variants. Tools are therefore needed to communicate our understanding of genetic vulnerability as a continuous trait, where a genetic liability confers risk for disease. Impute.me is one such tool, whose focus is on education and information on common, complex disorders with polygenetic architecture. Its research-focused open-source website allows users to upload consumer genetics data to obtain PRSs, with results reported on a population-level normal distribution. Diseases can only be browsed by *International Classification of Diseases*, 10th Revision (ICD-10) chapter–location or alphabetically, thus prompting the user to consider genetic risk scores in a medical context of relevance to the individual. Here, we present an overview of the implementation of the impute.me site, along with analysis of typical usage patterns, which may advance public perception of genomic risk and precision medicine.

Keywords: genetics, polygenic risk scores, direct-to-consumer, personal genomes, risk prediction

INTRODUCTION

In clinical genetics, testing for rare strong-effect causal variants is routinely performed in the health-care system to confirm a diagnosis or to evaluate individual risk suspected from anamnestic information (Baig et al., 2016), and in such instances, the use of genome sequencing is expanding (Byrjalsen et al., 2018). Meanwhile, outside of the health-care system, direct-to-consumer (DTC) genetics expands rapidly, providing the public with access to individual genetic data profiles and to interpretation of common genetic variants derived from genotyping microarrays (Kaye, 2008; Greshake et al., 2014). This is developing as a sprawling industry of consumer services with widely diverging standards, including third-party genome analysis services. These services typically provide individual results from analysis of common single-nucleotide polymorphisms (SNPs) with (at best) weak effects. They are therefore severely mis-aligned with current state-of-the-art, which at least for common, complex disease is to use polygenic risk scores (PRSs) to estimate the combined risk of common variation in the genome (Lee et al., 2008; Lewis and Vassos, 2017).

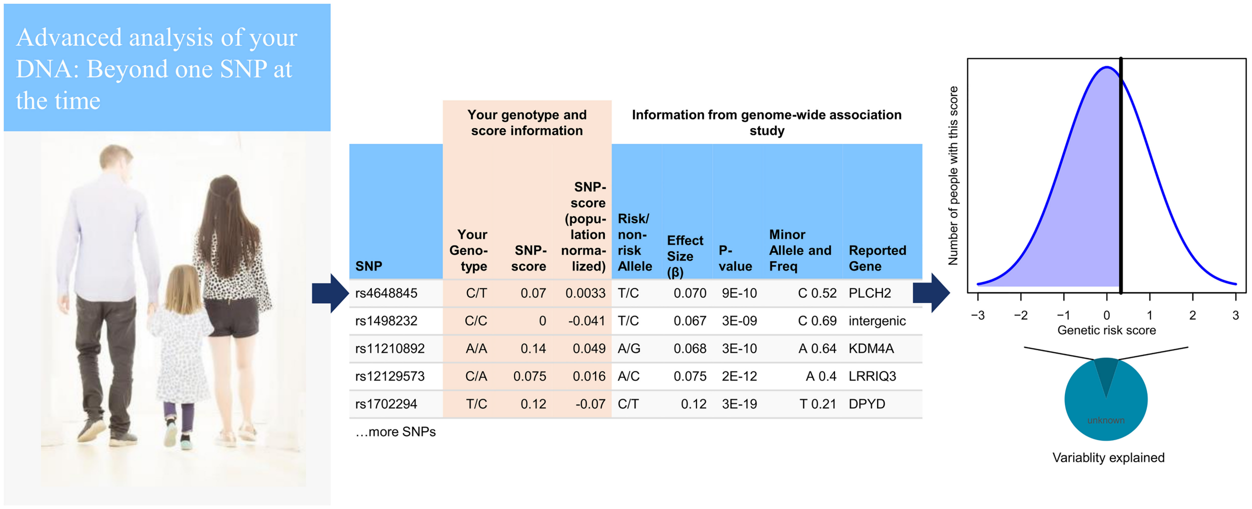
We believe that the goal of the academic genetics community should extend beyond theory. This means engaging with the public and assisting those who seek information, even when it means helping them to interpret their own genomic data. We therefore developed impute.me as an online web-app for analysis and education in personal genetic analysis. The web-app is illustrated in **Figure 1**. Using any major DTC vendor, a user can download their raw data and then upload it at impute.me. Uploaded files are checked and formatted according to procedures that have been developed to handle most types of microarray-based consumer genetics data, including an imputation step. These data are then further subjected to automated analysis scripts including PRS calculations. This includes more than 2,000 traits, browsable in different interface types (modules). Each module is designed with the goal of putting findings in as relevant a context as possible, prompting users to see common variant genetics as a support tool rather than a diagnosis finder. The aim is to provide information as broadly as possible to offer a real alternative to the widespread practice of reporting on weak SNP genotypes for any trait, even though that means generation of reports that are below any sensible threshold for clinical usability. We hope that having this as an open and accessible resource for everyone will be of help to the debate on what exactly constitutes clinical usability beyond high-risk pathogenic variants.

In this article, we will describe the (i) development and setup, (ii) validation and testing, (iii) evaluation of usage, (iv) communication of risk scores, and (v) ethics and implications. In the section *Development and Setup*, we discuss some of the challenges faced when developing a full personal-genome scoring pipeline. The goal of this section is to motivate and explain the choices made in development. In the second section, *Validation and Testing*, we use public Biobank data from individuals who consented for genetic research to test the effect of the impute.me scores on known disease outcomes. The purpose of this section is to test and validate scores, as well as to investigate consequences of some of the challenges that were raised in the first section. In the third section, *Evaluation of Usage*, we evaluate usage metrics of impute.me users. The goal of this section is to shed light on behavioral patterns of individuals who use DTC genetics for health questions and to offer recommendations that may be of use in other personal-genome scoring pipelines. In the section *Communication of Risk Scores*, we discuss our views on future directions particularly with respect to improving how genetic findings are presented to people. Finally, in *Ethics and Implications*, we discuss the ethics of providing access to health-related interpretation of DNA data.

DEVELOPMENT AND SETUP

The first challenge in development of personal genomic services is standardization. As the name impute.me implies, all genotype data are processed by imputation of genotype data (Howie et al., 2009; Delaneau et al., 2013). This procedure expands the data available into ungenotyped SNPs and increases overlap with public genome-wide association study (GWAS) summary statistics used to estimate risk. It also expands the SNP overlap between microarray types from the major vendors, such as 23andMe, MyHeritage, and Ancestry.com. Further, we have found that imputation helps in avoiding major errors, for example, strand-flip issues that arise from the dozens of different data formats. Eliminating such problems from further processing is one important step to minimize mis-interpretation of genome analysis. To ensure high standard of reported results, impute.me requires a fully completed imputation for continued analysis.

The second challenge is to estimate PRSs that are accurate and robust to heterogenous data sources. This is particularly important to an application utilized by people from around the world leveraging data from dozens of different vendors and data types. Importantly, PRSs calculated from GWAS of a population of (for example) European ancestry will perform better for individuals of the same ancestry, and the systematic shift (i.e., bias) in risk scores in individuals from other populations is a problem (Curtis, 2018). Because studies of all disease traits are not yet available for all non-European populations, the pragmatic solution has been to include a population-specific normalization


Advanced analysis of your DNA: Beyond one SNP at the time
Number of people with this score
−3
−2
−1
0
1
2
3
Genetic risk score
unknown
Variablity explained

	Your genotype and score information			Information from genome-wide association study				
SNP	Your Geno-type	SNP-score	SNP-score (popu-lation norma-lized)	Risk/ non-risk Allele	Effect Size (β)	P-value	Minor Allele and Freq	Reported Gene
rs4648845	C/T	0.07	0.0033	T/C	0.070	9E-10	C 0.52	PLCH2
rs1498232	C/C	0	-0.041	T/C	0.067	3E-09	C 0.69	intergenic
rs11210892	A/A	0.14	0.049	A/G	0.068	3E-10	A 0.64	KDM4A
rs12129573	C/A	0.075	0.016	A/C	0.075	2E-12	A 0.4	LRRIQ3
rs1702294	T/C	0.12	-0.07	C/T	0.12	3E-19	T 0.21	DPYD
…more SNPs								

FIGURE 1 | Basic pipeline setup from the user point of view. On upload of a genome, data are checked according quality control (QC) parameters that have been developed to handle most types of microarray-based consumer genetics data. The genome is then imputed using 1000 Genomes as reference (*left*). The imputed data are then further subjected to automated analysis scripts from 15 different modules, most of which are based on polygenic risk score calculations. The calculations include 1,859 traits from genome-wide association studies (GWASs) and 634 traits from the UK Biobank, as well as customized modules for height, and drug response. Most polygenic risk scores use GWAS significant single-nucleotide polymorphisms (SNPs) out of necessity, although 20 major diseases are based on LDpred all-SNP scores (*center*). A user can then browse their scores in relation to the population, shown together with a chart displaying how much variability is explained (*right*).

attempting to minimize the systematic shifts of scores for non-European ancestry users. Further, it is computationally and logistically easier to implement PRSs that use only the most (i.e., genome-wide) significant SNPs (often referred to as top SNPs), but the prediction strength is better when more SNPs are included (all-SNP), which, however, is more sensitive to ancestry biases (Lam et al., 2019). The impute.me pipelines calculate PRSs for each trait or disease on the basis of all-SNP-based PRS calculations if full genome-wide summary statistics are available and processed, and top-SNP-based PRS calculations if not.

The third challenge is presentation. For a single rare large-effect variant, such as for the pathogenic variants in the *BRCA* genes conferring very high risk of cancers (odds ratio >10; **Figure 2A**, upper left), presentation focuses on absence versus presence (Maxwell et al., 2016). However, also, low-effect variants, for example, as in pharmacogenetics, impacting statin response, is considered as having potential clinical use (Natarajan et al., 2017; **Figure 2A**, lower right). This difference in effect magnitude is a major challenge in result presentation and understanding, particularly because a firm threshold is difficult to set: In the context of a drug-prescription situation or a question of which of two suspected disease risks is the most likely, it may be useful to know such scores. But in the context of an otherwise healthy individual, genetic risks are only relevant if we are very certain of them, they are serious, and preferably actionable [e.g., BRCA variants (Kalia et al., 2017)]. For this reason, we have made the design choice to avoid the use of lists sorted by risk score. Currently, scores are accessible through either an alphabetically sorted list or in a tree-like setup where genetic scores are reported in a health-context tree (**Figure 2B**). In this, all scores are included, but scores that are less relevant to healthy individuals (i.e., most of them) are buried deeper into the health-context tree. As further discussed in the section *Future Challenges*, there are a lot of remaining challenges to solve in this question.

VALIDATION AND TESTING

To evaluate pipelines on individuals with known disease outcomes, we investigated 242 samples from the CommonMind data set. The CommonMind data set includes patients with schizophrenia (SCZ), bipolar disorder, and controls, from European ancestry and from African ancestry. For each disorder and each ancestry group, the full impute.me pipelines were applied, including imputation and PRS calculation. Additionally, SNP sets corresponding to each of three major DTC companies were extracted and re-calculated. This was done to test the hypothesis that PRS calculation in mixed SNP sets poses particular challenges with regard to missing SNPs. Such sets of genotyped SNPs that are different in each sample are an unavoidable consequence of working with online data uploads.

We found that disease prediction strength, measured as variability explained, corresponded well to theoretical expectations of known SNP heritability (Lee et al., 2017; Li et al., 2017; Wünnemann et al., 2019). Secondly, we found that using all-SNP scores resulted in better prediction than top-SNP scores, which was as expected (Vilhjálmsson et al., 2015). Thirdly, we found that prediction was more accurate in individuals of European ancestry compared with individuals of African ancestry, which is concordant with the PRSs being developed

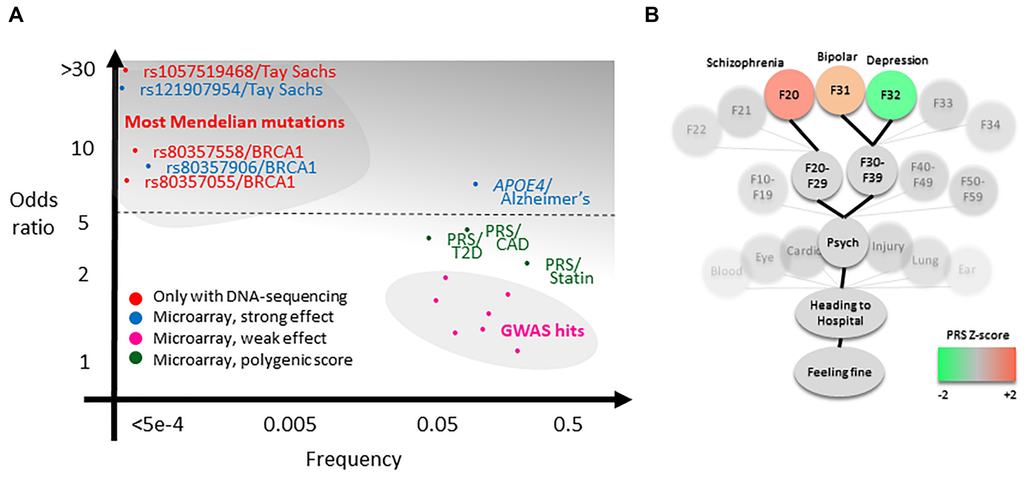


FIGURE 2 | Theoretical background of the analysis pipeline. **(A)** Clinical genetics currently concern high-effect DNA variants that often can only be sequenced (*red*). Additionally, high-effect variants such as APOE4 and a small subset of BRCA1 and BRCA2 pathogenic variants are possible to measure using microarray (*blue* includes several other variants not shown in plot, e.g., Parkinson's variants). There may be an untapped potential for valuable clinical information in polygenic risk scores (PRSs) for common disease (*green*), for example, for type 2 diabetes (T2D), coronary artery disease (CAD), or statin response (Natarajan et al., 2017; Khera et al., 2018; Wünnemann et al., 2019). It is a primary aim of the impute.me project to make this potential available more broadly, balancing the practice of relying on individual genome-wide association study (GWAS) single-nucleotide polymorphisms (SNPs) and/or reporting of SNP genotypes (*pink*). **(B)** The secondary aim is to provide genetic scores in a relevant context, exemplified in the precision medicine module showing the so-called health-context tree. This tree consists of all entries from the international classification of disease [*International Classification of Diseases*, 10th Revision (ICD-10)], linked to all genetic studies. It allows browsing of PRSs in a relevant context. In the example shown, the tree is open on the psychiatry chapter, showing PRSs for schizophrenia (F20), unipolar depression (F32), and bipolar depression (F31). Although these scores have little predictive relevance for a healthy individual, they may be useful in the context of psychiatric evaluation, particularly in the case of more extreme scores.

from European Ancestry GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Hou et al., 2016; Lee et al., 2018). These observations match well with findings from studies of PRSs in much larger data sets. We caution that universally valid estimates of variability explained are better derived from larger studies that can consider the numerous issues such as balancing of cases and controls, realistic sampling conditions, and other inflations of effects. The intention here is to provide a specific test of impute-me pipelines and address DTC data-related questions.

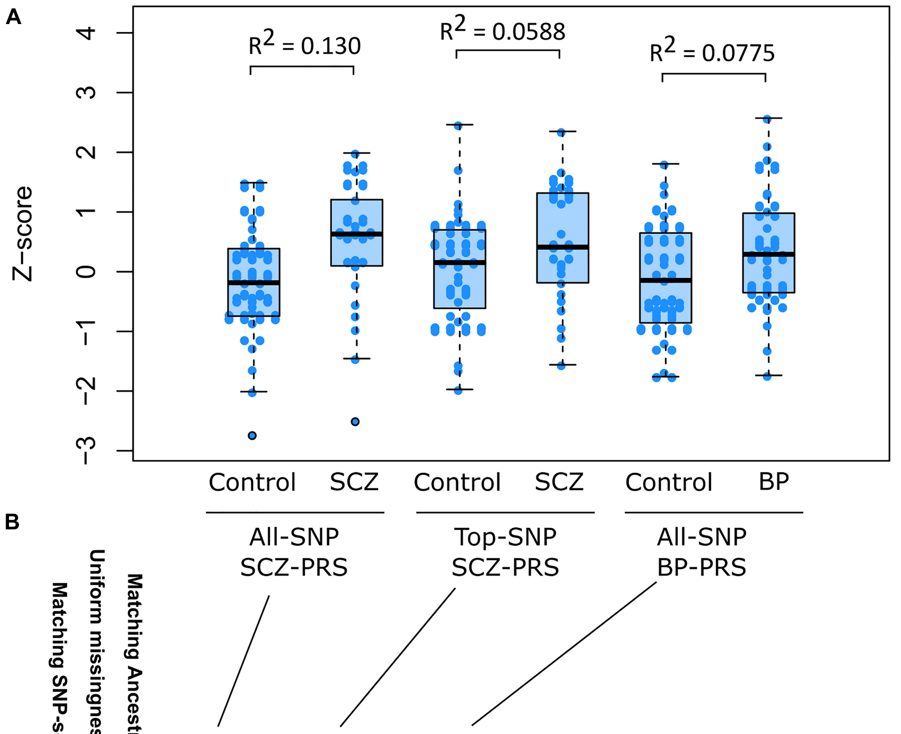
Of importance to this, we found that PRS prediction in mixed samples of non-imputed data causes severe problems. When training PRS algorithms, an SNP set is prespecified. The pipelines evaluated here were trained with HapMap3 as SNP set. Similar choices are made in other published PRSs. However, such SNP sets may not match with the SNPs available in downloadable raw data from DTC vendors. We therefore tested what prediction strength would be possible when using raw data directly from DTC vendors, both in a uniform setting (e.g., "all individuals use 23andMe v4 data") and in a mixed setting (e.g., "individuals have data from different vendors"). We found that in the uniform setting, roughly half the predictive strength remained when using genotype data that are not imputed to match the HapMap3 SNP sets (**Figure 3**, rows 2 and 4). In the mixed setting, virtually no predictive strength remained (**Figure 3**, rows 3 and 6). The mixed setting is the reality that is faced, both for third-party analytical services and for DTC vendors with different chip versions. Imputation is therefore likely to be an essential requirement in such scenarios.

To compare these findings with approaches that look at one SNP at the time, we extracted the SNPedia/Promethease SNPs that were indicated as associated with SCZ (Cariaso and Lennon, 2012). All cases (n = 25) and all controls (n = 39) had at least one risk variant from at least one of the 139 SNPs that indicated SCZ association. When focusing on SNPs that had the SNPedia/Promethease-defined *"magnitude"*-level (*sic.*) at >1.5, we found that 80% of the SCZ cases (20 of 25) had at least one SNPedia/Promethease risk variant. Among the healthy controls, 84% (33 of 39) had at least one such risk variant (p = 0.9 for difference in proportions). In other words, it is not very predictive to know if you have a SCZ SNP. This illustrates the importance of considering more than one SNP at the time.

Finally, we compared pipeline reproducibility using two genome-data files, one obtained from MyHeritage and one from Ancestry.com, but sampled from the same person. After processing through the impute-me pipelines, the correlation between PRS values over 1,468 traits was r = 0.933 between the two samples. Traits that showed discrepancy between the two data files typically were based on only few SNPs, of which one did not meet imputation quality thresholds for one of the data files.

EVALUATION OF USAGE

As of June 2019, a total of 28,651 genomes had been uploaded to impute.me, and a total of 3.1 million analytical queries had been performed (**Figure 4A**). The following additional



#	Matching SNP-set	Uniform missingness	Matching Ancestry	Variability explained (%)			Example
1	+	+	+	13.06	5.88	7.75	SNP-set matching the PRS *(e.g. imputed data)*
2		+	+	4.4±2.3	3.0±2	6.2±0.9	Wrong SNP-set, uniform missingness *(e.g. one non-imputed microarray type with a general PRS)*
3			+	0.01±3.9	0.03±1.6	0.0042±2.9	As #2, but with non-uniform set of microarrays *(e.g. live online setting with mixed format uploads)*
4	+	+		5.52	0.0001	2.24	As #1, but for individuals with African ancestry
5		+		2.5±0.7	0.16±0.17	5±2.4	As #2, but for individuals with African ancestry
6				0.029±3.2	0.001±0.82	0.079±1.0	As #3, but for individuals with African ancestry

FIGURE 3 | Pipeline evaluation using publicly available genotyped cohorts. **(A)** Three scores were calculated in individuals of European ancestry and relevant diagnoses (n_{control} = 39, n_{SCZ} = 25, and n_{BP} = 39): a schizophrenia (SCZ) all-single-nucleotide polymorphism (SNP) score (n_{SNP} = 558,406), an SCZ top-SNP score (n_{SNP} = 93), and a bipolar (BP) all-SNP score (n_{SNP} = 554977). The BP top-SNP score only used five genome-wide significant SNPs and was not tested. The proportion of variance explained (Nagelkerke R^2) is shown above each case–control pair. **(B)** Testing different conditions of ancestry and input SNP sets. Row #1 corresponds to the variability explained after processing through the full impute.me pipeline, that is, the same calculation shown in the plot. Row #2 shows the prediction level when the polygenic risk score (PRS) algorithm uses input samples from only one type of direct-to-consumer (DTC) vendor, but the algorithm has not been trained specifically for that SNP set. Values are given as mean ± SD of three analyses in which SNP sets were all from 23andMe (v4), ancestry-com, or MyHeritage. row #3 shows the prediction when each sample uses different SNP sets, that is, the actual situation when dealing with user-uploaded DTC data online. Values are given as mean ± SD over 100 random drawings of combinations of the 23andMe (v4), Ancestry.com, and MyHeritage sets, in proportions of 55, 30, and 15%, respectively. These proportions correspond to what are observed in live users. Rows #4–6 shows the same as #1–3 but calculated for CommonMind individuals of African ancestry (n_{control} = 47, n_{SCZ} = 39, and n_{BP} = 6). The corresponding AUC values for this figure are 0.693, 0.614, and 0.634 for row #1: for row #2, 0.55 ± 0.12, 0.53 ± 0.084, and 0.62 ± 0.012; and for row #3, 0.58 ± 0.047, 0.55 ± 0.03, and 0.57 ± 0.047. Additionally, an extended version of the figure is available at www.impute.me/prsExplainer, where additional metrics of prediction can be explored interactively.

observations about user behavior may be of use to the genetics research community.

Common and well-known diseases are the most sought after. By overall click count and comparing over several different modules, there is no doubt that users are most interested in common disease types; diseases of the brain, heart, and metabolism are more requested. Interface design may of course play important roles in such choices. For example, the choice to serve disease traits as alphabetically sorted lists is likely to artificially inflate interest in, for example, abdominal aneurysm (**Figure 4B**). However, the larger interest in psychiatry, cardiovascular, and metabolic disorders remains also in the precision medicine module, which is not presented as an alphabetically sorted list (**Figure 4C**). It is possible that greater scientific interest in PRSs in these fields also drives some of these effects, but we cannot explain why other fields where PRSs are actively discussed, such as cancer, are not attracting more attention.

Likewise, it seems that common disease ("complex disease module") is more sought after than rare disease ("rare disease

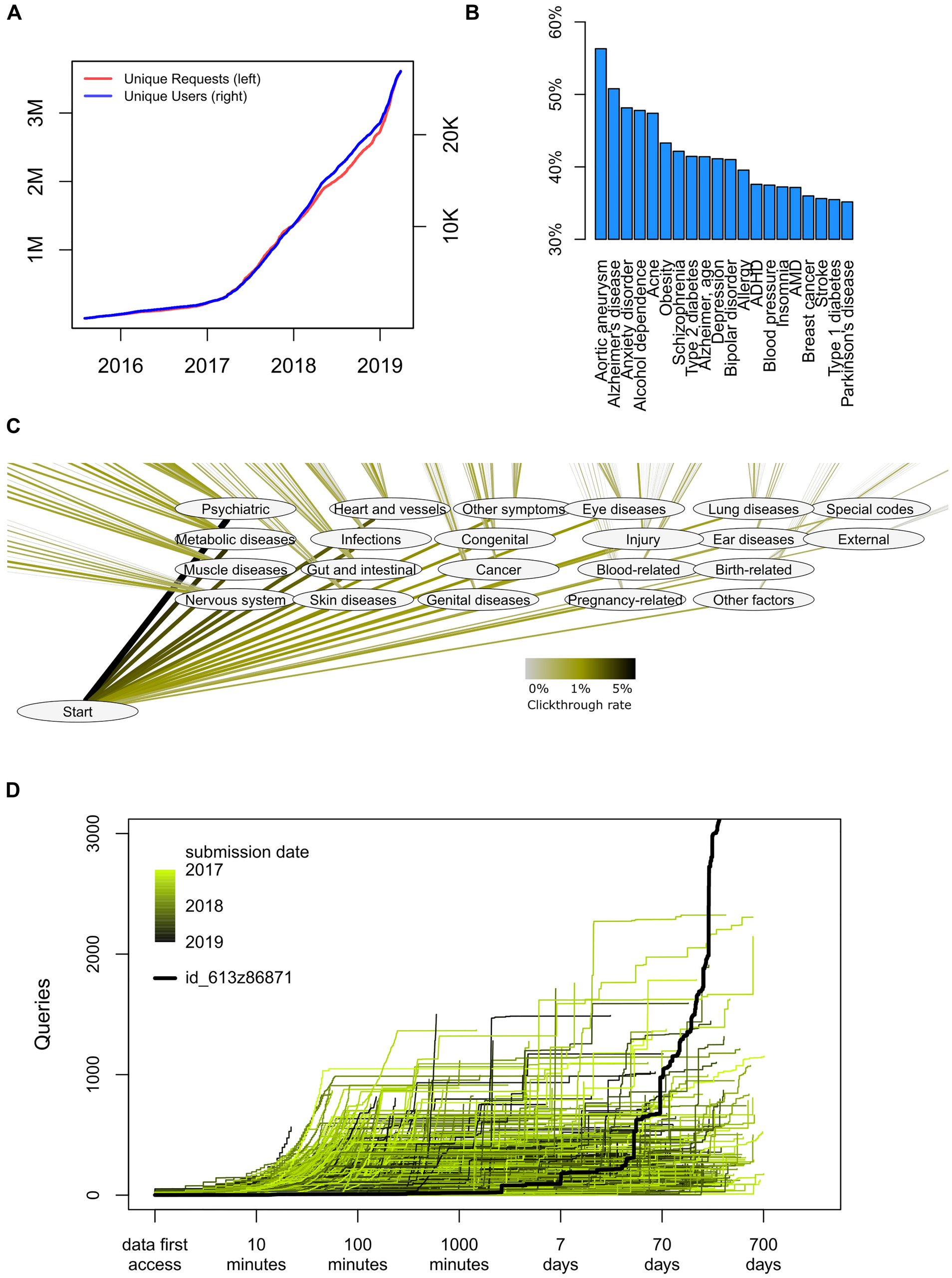


FIGURE 4 | Detailed usage statistics. **(A)** Overall count of unique users and unique analysis requests since August 2015. Each *request* corresponds to a specific analysis, for example, the risk score for a disease, or a view in the *International Classification of Diseases*, 10th Revision (ICD-10)-based map in the precision medicine module. Each *user* corresponds to an uploaded genome with a unique md5sum. There is no check for twins, altered files, or users with data from separate direct-to-consumer (DTC) companies. **(B)** Distribution of user interests in a trait in the *complex disease module*. In this module, each disease entry is presented on an alphabetically sorted list, with aortic aneurysm being the default value. The percentage indicates how many of the users scrolled down and selected this disease at least once (n_{clicks} = 871,855). **(C)** Distribution of interests in a trait in the *precision medicine module*. In this module, each disease entry is presented in the layout of the ICD-10 classification system. The click-through rate reflects how many users pursued information in a given chapter or subchapter, as percentage of total amount of clicks (n_{clicks} = 114,039). **(D)** Analysis of how individuals use the interface over time. For each user, the number of queries is shown as a function of time after they first access their data. As all data are automatically deleted after 2 years, no queries extend beyond 730 days. The color code indicates the submission date. The highlighted black line indicates the publically available permanent test user with ID id_613z86871, which is omitted from all other analyses.

module”); 95% of all users visit the first, whereas only 70% visit the second. Again, interface design and project goals probably play a big role in this—the landing page headers says *Beyond one SNP at the time*, and the rare disease module is found in the navigation bar only below seven other module entries. But it may also illustrate a central communication challenge for the field:

People are more interested in the genetics of common, complex diseases with small effect sizes (**Figure 2A**, lower right) but may interpret the results as if they were for rare diseases with large effect sizes (**Figure 2A**, upper left).

Finally, we have observed that usage of health genetic data surprisingly often is not just a test-and-forget event. When plotting query count as a function of time from first data access, we find an expected pattern of intense browsing the hours and days after first data access (**Figure 4D**). However, many users revisit their data even months and years after first data access, perhaps implying that results are considered and saved and then revisited at a later time in a different context.

COMMUNICATION OF RISK SCORES

Generation of the PRS data presents one set of challenges, but communicating them to people in such a way as to make it both comprehensible and useful presents another (Lipkus and Hollands, 1999; Naik et al., 2012). We believe that this is a crucial unmet need in current genetics research, because presenting PRS data in a way that is useful requires an understanding of people's motivations for accessing them in the first place.

To date, studies of PRSs have focused on providing people with PRS information in relation to specific conditions [e.g., cancer (Bancroft et al., 2014, 2015; Smit et al., 2018; Young et al., 2018)] for which participants have an indicated risk and exploring understanding and reactions. No studies have examined what motivates people to seek out and access their own PRSs for common complex conditions, and little is known about how people understand or respond to the data they receive.

Polygenic risk scores information is inherently probabilistic in nature, which is well known to be difficult for people to understand (Hallowell et al., 1998; Smerecnik et al., 2009), and receiving information about genetic risk is not necessarily benign. When people receive genetic test results that they perceive to reflect high risk for a condition, this can have negative impact on outcomes like self-perception and affect, and in the case of receiving high-risk test results for Alzheimer's disease—can actually impact objective measures of cognitive performance (Wilhelm et al., 2009; Dar-Nimrod et al., 2013; Lineweaver et al., 2014; Lebowitz and Ahn, 2017; Turnwald et al., 2019). Therefore, how information about genetic risk is communicated matters.

The literature suggests that when communicating risk, the most useful and effective strategy is to use absolute risks (Lipkus and Hollands, 1999; Reyna et al., 2009; Naik et al., 2012). In the case of PRSs with modest predictive power, however, this may simply result in restating the population prevalence of a disease for everyone (Janssens, 2019). It is therefore important that the predictive strength is also included in this communication; that is how much the genetic component potentially could alter the absolute risk. The genetic component corresponds to the SNP heritability, and we are therefore exploring how to best include this information (e.g., **Figure 1**, right). Currently, we have registered the SNP heritability for 294 of the reported traits, available as an experimental option called "plot heritability." We believe that a main future direction is to experiment and expand on how to best communicate this to people.

It will therefore be useful to have a constant flow of people that are interested in interpreting their genetics and expose them to various modes of presentation. Some could involve statistically advanced concepts, like the area under the receiver operating characteristic curve (AUC) and SNP heritability, but others may take simpler approaches, such as the explanatory jar model pioneered for talking with families about genetics (Peay and Austin, 2011; Austin, 2019). One may even imagine layered models of increasing complexity. This should be followed up with questionnaires probing the level of understanding and general impact on users, something that is possible using the impute.me platform.

ETHICS AND IMPLICATIONS

Using genetics to maximize the benefits and minimize the harms to individuals and society requires the effective management of the ethical, legal, and social implications of genetics. Researchers have a responsibility to ensure that the technology and the knowledge developed through genetic research are used responsibly, in light of the bioethical principles of beneficence, non-maleficence, justice, and autonomy (Lázaro-Muñoz et al., 2019). Given that for most complex disorders there is currently a lack of data regarding the harms or benefits of accessing PRS information, the fundamental principle in favor of making PRSs available to the public is that of autonomy—in the context of genetic testing, this refers to "the right of persons to make an informed, independent judgment about whether they wish to be tested and then whether they wish to know the details of the outcome of the testing" (Andrews et al., 1994). Accordingly, currently, DTC users can access health information through portals of DTC providers and through third-party applications (Kalokairinou et al., 2018; Tiller and Lacaze, 2018; Ahmed and Shabani, 2019). The problem is that many popular websites do not communicate high-quality genetic knowledge, in part possibly owing to the lack of engagement by the research communities (Badalato et al., 2017). One solution to this problem is to call for regulation and to ban such sites. Alternatively, as we propose here, it is possible to meet user demands and strive to do so as ethically as possible.

To exemplify this, as researchers, we have a choice in whether to provide access to a state-of-the-art PRS for a disease or not. We know that this PRS does not explain everything about the disease, does not account for all the genetic information, and is not part of today's clinical guidelines. However, we also know that users are already accessing information about disease through DTC genetics. These users may get their information from flawed assumptions of SNP effect sizes or from commercial platforms with little interest in explaining the limitations of the score. We argue that the choice that maximizes the potential for benefits to individuals is to provide the score and to provide it in a setting that puts its consequence in perspective.

An example of such perspective is that of giving reports by disease score, and not by individual risk variant as is currently the case in most third-party analytics apps. Many people carry the high-risk allele for a common variant, but fewer people have a high PRS, which is the sum of all such risk variants. An example of this is the 84% frequency of SCZ risk variants in healthy users according to SNPedia, as reported above. This means that for those autonomously seeking information on health genetics data, the use of PRSs has the potential to decrease the level of induced worry in people in comparison with the current levels. Similarly, smart interface design can actively steer people toward browsing results by indication, and away from the pervasive practice of reporting the worst genetic scores for any disease first. This too may serve to reduce induced worry, in alignment with the general approach of testing only on indication to limit false-positive rates. Finally, of course, adaptive warnings based on risk levels, including referral to resources such as findageneticcounselor.com, is something we continuously strive to optimize.

CONCLUSION

In summary, we present impute.me as a fully operational General Data Protection Regulation (GDPR)-compliant genetic analysis engine covering a very broad range of health-related traits, specifically focusing on optimizing possibilities from microarray-based DNA measurements. The challenges, their solutions, and the curation work behind them are highly relevant today in a setting of highly varying quality in interpretation of personal consumer genetics. In the future, we can expect that PRS predictiveness will increase. This will mean a continued and increasing relevance of the platform, even more so as the number of individuals doing genetic testing increases. With a directed push toward responsible use of genetics, this may even prove to be an overall clinical benefit.

METHODS

Data Privacy and Security

On data submission, each personal genome is assigned a nine-digit alphanumeric unique identifier ("uniqueID"). This uniqueID is used as login and identifier throughout all downstream processes because it has no information that is personally linkable, as opposed to, for example, an email address. The uniqueID is initially linked to two types of data: those that can be traced back to individual that submitted the genome and those that cannot. Genomic data, filename of submitted data, and email address are of the first type: genomic data because it can be used with software such as *gedmatch* to trace family patterns, filename because it often contains the name of the submitter (e.g., 23andMe data use full name as standard), and email for obvious reasons. Data of the traceable type are deleted 14 days after processing, which is the period in which users are able to download their full imputed data sets. The exception is email addresses, which are not deleted but instead unlinked from the uniqueID and kept elsewhere for the purpose of follow-up studies. Either way, this means that 14 days after processing, there exists nothing on the servers that can link results (designated with a uniqueID) with the person who submitted the data (any of the three traceable data types). Thus, even if the database is leaked or lost, it is not possible to link the data to an actual person. After 2 years, the remaining non-traceable data, for example, the derived calculations, the risk scores, and the genotypes of SNPs of specific interest, are all completely deleted. All ingoing and outgoing data transfers are encrypted using Transport Layer Security (TLS 1.3). All storage is encrypted using the AES-256 standard.

This means that all data are collected for specified, explicit, and legitimate purposes in a transparent manner and kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. We therefore consider that these measures both provide adequate security and privacy protection and are in accordance with the GDPR.

Preprocessing and Bioinformatics

After submission of data, a comprehensive bioinformatic processing of the genotype data takes place. This is done in order of free computing nodes becoming available, consisting of several support programs; first, a shapeit call is made to phase the data correctly (Delaneau et al., 2013), and then an impute2 call is made with 1000 Genomes version 3 as reference (Howie et al., 2009; 1000 Genomes Project Consortium et al., 2015). Although the pipelines are not guaranteed to handle any format they receive, they currently operate with less than 1% processing failures, meaning uploads that cannot proceed through the full quality control and imputation pipelines. The failures are typically due to file formatting errors, missing chromosomes, or any number of other odd data corruptions that real-world data exchange suffers from.

Several customizations have been made with the goal of minimizing memory footprint and thereby allowing running in a clustered fashion on a series of small cloud computers. This allows for relatively easy scaling of capacity: one simple setup ("hub-only"), where calculations are run on the same computer as the website interface. Another is a hub + node-setup, where a central hub server stores data and shows the website, while a scalable number of node-servers perform all computationally heavy calculations. After preprocessing is finished, two new files are created: a *.gen* file with probabilistic information from imputation calls and a *simple format* file with best guess genotypes, called at a 0.9 impute2 INFO threshold. All further calculations are based on these files. A mail with download links to these two files is returned to the user, along with a JSON-formatted file containing a machine-readable summary of all calculations, as well as links with guidance to obtain more in-depth information on personal DNA interpretation (Folkersen, 2018).

Polygenic Risk Score Calculation

From the preprocessed data, a modular set of trait predictor algorithms is applied. For many of the modules, the calculations are trivial. For example, this could be the reporting of presence and/or absence of a specific genotype, such as ACTN3 and ACE-gene SNPs known to be (weakly) associated with athletic performance. These are included mostly because users expect them to be. For others, we rely heavily on PRSs.

An important distinguishing factor between different PRS algorithms is how risk alleles are selected. A commonly used approach includes variants based on whether they surpass a given p-value threshold in the GWAS, retaining only linkage disequilibrium (LD)-independent variants using LD-based clumping, often with a p-value threshold of genome-wide significance ($p < 5e^{-8}$). Herein, we refer to this approach as the "top-SNP" approach. The top-SNP approach has the advantage that it is simple to explain, is easy to obtain for many GWAS, and has a light computational burden (e.g., Buniello et al., 2019; Lambert et al., 2019; Patron et al., 2019; Watanabe et al., 2019). However, research has repeatedly shown that the inclusion of variants that do not achieve genome-wide significance improves the variance explained by PRSs, with PRSs including all variants often explaining the most variance. PRSs based on GWAS effect sizes that have undergone shrinkage to account for the LD between variants have been shown to explain more variance than PRSs that account for LD via LD-based clumping (Vilhjálmsson et al., 2015). Herein, we refer to this approach as the all-SNP approach. It is more computationally and practically intensive to implement at scale. Consequently, within impute.me, each trait or disease reported shows all-SNP-based PRS calculations if such is available, and top-SNP-based PRS calculations if not.

In the top-SNP calculation mode, the results are scaled such that the mean of a population is zero and the standard deviation (SD) is 1, according to the relevant 1000 Genomes super-population: African, admixed American, East Asian, European, or South Asian.

$$\text{Population-score}_{\text{snp}} = \text{frequency}_{\text{snp}} \times 2 \times \text{beta}_{\text{snp}}$$

$$\text{Zero-centered-score} = \sum \text{Beta}_{\text{snp}} \times \text{Effect-allele-count}_{\text{snp}} \text{-Population-score}_{\text{snp}}$$

$$Z\text{-score} = \text{Zero-centered-score}/\text{Standard-deviation}_{\text{population}}$$

where beta [or log(odds ratio)] is the reported effect size for the SNP effect allele, frequency$_{\text{SNP}}$ is the allele frequency for the effect allele, and the Effect-allele-count$_{\text{SNP}}$ is the allele count from genotype data (0, 1, or 2).

In the all-SNP calculation, the scaling is similar but done empirically, that is, based on previous impute.me users of matching ethnicity. This mode of scaling is also available as an optional functionality in the top-SNP calculations and generally seems to match well with the default 1000 Genomes super-population scaling.

The all-SNP scores were derived using weightings from the LDpred algorithm (Vilhjálmsson et al., 2015). This algorithm adjusts the effect of each SNP allele for those of other SNP alleles in LD with it and also takes into account the likelihood of a given allele to have a true effect according to a user-defined parameter, which here was taken as *wt1*, that is, the full set of SNPs. The algorithm was directed to use hapmap3 SNPs that had a minor allele frequency >0.05, Hardy–Weinberg equilibrium $p > 1e^{-05}$, and genotype yield >0.95, consistent with our expectation that these would be the best imputed SNPs after full pipeline processing.

Pipeline Testing

To test the pipelines described herein, the CommonMind genotypes measured with the microarray of the type H1M were downloaded along with phenotypic information. Each sample was processed through the impute.me pipelines, using the batch upload functionality. Reported ethnicity was compared with pipeline (genotype) assigned ethnicity and found to be concordant.

After pipeline completion, we extracted three PRSs for each sample, corresponding to SCZ all-SNP, SCZ top-SNP, and BP all-SNP. In the github repository for impute.me, these three correspond to the scores labeled *SCZ_2014_PGC_EXCL_DK.EurUnrel.hapmap3.all.ldpred.effects*, *schizophrenia_25056061*, and *BIP_2016_PGC.All.hapmap3.all.ldpred.effects* trait IDs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Hou et al., 2016). These extracted scores formed the basis of the row #1 and #4 calculations in **Figure 3**. The remaining rows were created by subsetting the best guess imputed genotypes into new sets of users, corresponding to each of three major DTC vendors and then re-running the scoring algorithms with either uniform data or mixed data. Uniform data are here defined as all 195 samples having the same set of SNPs available, corresponding to one of three DTC vendors in each run. Mixed data are defined as samples having different sets of SNPs available, a set corresponding to actual distributions of customers from different DTC vendors, with distributions redrawn 100 times. We estimated the predictive ability of the PRSs using Nagelkerke's R^2 and AUC.

Usage Evaluation

A log data freeze was performed on June 8, 2019 by making a copy of all usage log files and then removing the uniqueID of each user. This was done to prevent it from being linked with the genetic data of that user. The exception was the publicly available permanent test user with ID id_613z86871, which was lifted out before analysis and is not included in other summary statistics. Generally, a user corresponds to an uploaded genome with a unique md5sum. Click-through rates were calculated as fraction of users that performed any query in the module in question; for example, the precision medicine module was only launched in September 2018 and, therefore, only counts clicks from people who have used it. Plots were generated using base-R version 3.4.2 and cytoscape version 3.71.

URLS

Code repository: https://github.com/lassefolkersen/impute-me Web resource: https://www.impute.me/

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: CommonMind data doi: 10.7303/syn275 9792.

ETHICS STATEMENT

The studies involving genotypes of human participants were reviewed and approved by the CommonMind Consortium. This data is generated from postmortem human brain specimens originating from tissue collections at the Mount Sinai NIH Brain Bank and Tissue Repository, University of Pennsylvania Brain Bank of Psychiatric illnesses and Alzheimer's Disease Core Center, The University of Pittsburgh NIH NeuroBioBank Brain and Tissue Repository, and the NIMH Human Brain Collection Core. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LF coded the code. All authors contributed to interpretation, drafting the work, critical revision for important intellectual content, and final approval of the manuscript.

FUNDING

CL and OP were supported by UK Medical Research Council grant N015746 and by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. CommonMind is supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd., and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH- 075916, P50M096891, P50MH084053S1, R37MH057881, AG02219, AG05138, MH06692, R01MH110921, R01MH109677, R01MH109897, U01MH103392, and contract HHSN271201300031C through IRP NIMH. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

In addition to the crucial resources cited in the text, we wish to thank the CommonMind Consortium for availability of testing data. We also wish to thank Ron Nudel for discussions on the ethics section.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Ahmed, E., and Shabani, M. (2019). DNA data marketplace: an analysis of the ethical concerns regarding the participation of the individuals. *Front Genet.* 10:1107. doi: 10.3389/fgene.2019.01107

Andrews, L. B., Fullarton, J. E., Holtzman, N. A., and Motulsky, A. G. (1994). *Assessing Genetic Risks - Implications for Health and Social Policy*. Washington, DC: National Academies Press.

Austin, J. C. (2019). Evidence-based genetic counseling for psychiatric disorders: a road map. *Cold Spring Harb. Perspect. Med.* 9:a036608. doi: 10.1101/cshperspect.a036608

Badalato, L., Kalokairinou, L., and Borry, P. (2017). Third party interpretation of raw genetic data: an ethical exploration. *Eur. J. Hum. Genet.* 25, 1189–1194. doi: 10.1038/ejhg.2017.126

Baig, S. S., Strong, M., Rosser, E., Taverner, N. V., Glew, R., Miedzybrodzka, Z., et al. (2016). UK Huntington's disease prediction consortium, quarrell OW. 22 years of predictive testing for Huntington's disease: the experience of the UK Huntington's prediction consortium. *Eur. J. Hum. Genet.* 24, 1396–1402. doi: 10.1038/ejhg.2016.36

Bancroft, E. K., Castro, E., Ardern-Jones, A., Moynihan, C., Page, E., Taylor, N., et al. (2014). It's all very well reading the letters in the genome, but it's a long way to being able to write: men's interpretations of undergoing genetic profiling to determine future risk of prostate cancer. *Fam. Cancer* 13, 625–635. doi: 10.1007/s10689-014-9734-3

Bancroft, E. K., Castro, E., Bancroft, G. A., Ardern-Jones, A., Moynihan, C., Page, E., et al. (2015). The psychological impact of undergoing genetic-risk profiling in men with a family history of prostate cancer. *Psychooncology* 24, 1492–1499. doi: 10.1002/pon.3814

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Byrjalsen, A., Stoltze, U., Wadt, K., Hjalgrim, L. L., Gerdes, A. M., Schmiegelow, K., et al. (2018). Pediatric cancer families' participation in whole-genome sequencing research in Denmark: parent perspectives. *Eur. J. Cancer Care* 27:e12877. doi: 10.1111/ecc.12877

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

Curtis, D. (2018). Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* 28, 85–89. doi: 10.1097/YPG.0000000000000206

Dar-Nimrod, I., Zuckerman, M., and Duberstein, P. R. (2013). The effects of learning about one's own genetic susceptibility to alcoholism: a randomized experiment. *Genet. Med.* 15, 132–138. doi: 10.1038/gim.2012.111

Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* 93, 687–696. doi: 10.1016/j.ajhg.2013.09.002

Folkersen, L. (2018). *Understand Your DNA - A Guide*. Singapore: World Scientific Publishing.

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

Hallowell, N., Statham, H., and Murton, F. (1998). Women's understanding of their risk of developing breast/ovarian cancer before and after genetic counseling. *J. Genet. Couns.* 7, 345–364. doi: 10.1023/A:1022072017436

Hou, L., Bergen, S. E., Akula, N., Song, J., Hultman, C. M., Landén, M., et al. (2016). Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum. Mol. Genet.* 25, 3383–3394. doi: 10.1093/hmg/ddw181

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Janssens, A. C. J. W. (2019). Proprietary algorithms for polygenic risk: protecting scientific innovation or hiding the lack of it? *Genes* 10:E448.

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, update (ACMG SF v2.0): a policy statement of the american college of medical genetics and genomics. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Kalokairinou, L., Howard, H. C., Slokenberga, S., Fisher, E., Flatscher-Thöni, M., Hartlev, M., et al. (2018). Legislation of direct-to-consumer genetic testing in Europe: a fragmented regulatory landscape. *J. Commun. Genet.* 9, 117–132. doi: 10.1007/s12687-017-0344-2

Kaye, J. (2008). The regulation of direct-to-consumer genetic tests. *Hum. Mol. Genet.* 17, R180–R183.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Lam, M., Chen, C. Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., et al. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* 51, 1670–1678. doi: 10.1038/s41588-019-0512-x

Lambert, S., Gil, L., Jupp, S., Chapman, M., Parkinson, H., Danesh, J., et al. (2019). *The Polygenic Score (PGS) Catalog: An Open Database To Enable Reproducibility And Systematic Evaluation*. Available at: www.pgscatalog.org (accessed November 2019).

Lázaro-Muñoz, G., Sabatello, M., Huckins, L., Peay, H., Degenhardt, F., Meiser, B., et al. (2019). ISPG Ethics Committee. International society of psychiatric genetics ethics committee: issues facing us. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 180, 543–554. doi: 10.1002/ajmg.b.32736

Lebowitz, M. S., and Ahn, W. K. (2017). Testing positive for a genetic predisposition to depression magnifies retrospective memory for depressive symptoms. *J. Consult. Clin. Psychol.* 85, 1052–1063. doi: 10.1037/ccp0000254

Lee, S. H., Clark, S., and van der Werf, J. H. J. (2018). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLoS One* 12:e0189775. doi: 10.1371/journal.pone.0189775

Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4:e1000231. doi: 10.1371/journal.pgen.1000231

Lee, S. H., Weerasinghe, W. M., Wray, N. R., Goddard, M. E., and van der Werf, J. H. (2017). Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci. Rep.* 7:42091. doi: 10.1038/srep42091

Lewis, C. M., and Vassos, E. (2017). Prospects for using risk scores in polygenic medicine. *Genome Med.* 9:96. doi: 10.1186/s13073-017-0489-y

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* 49, 1576–1583.

Lineweaver, T. T., Bondi, M. W., Galasko, D., and Salmon, D. P. (2014). Effect of knowledge of APOE genotype on subjective and objective memory performance in healthy older adults. *Am. J. Psychiatry* 171, 201–208. doi: 10.1176/appi.ajp.2013.12121590

Lipkus, I. M., and Hollands, J. G. (1999). The visual communication of risk. *J. Natl. Cancer Inst.* 25, 149–163.

Maxwell, K. N., Domchek, S. M., Nathanson, K. L., and Robson, M. E. (2016). Population frequency of germline BRCA1/2 mutations. *J. Clin. Oncol.* 34, 4183–4185. doi: 10.1200/jco.2016.67.0554

Naik, G., Ahmed, H., and Edwards, A. G. (2012). Communicating risk to patients and the public. *Br. J. Gen. Pract.* 62, 213–216.

Natarajan, P., Young, R., Stitziel, N. O., Padmanabhan, S., Baber, U., Mehran, R., et al. (2017). Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* 135, 2091–2101. doi: 10.1161/circulationaha.116.024436

Patron, J., Serra-Cayuela, A., Han, B., Li, C., and Wishart, D. S. (2019). Assessing the performance of genome-wide association studies for predicting disease risk. *PLoS One* 14:e0220215. doi: 10.1371/journal.pone.0220215

Peay, H. L., and Austin, J. (2011). *How To Talk With Families About Genetics And Psychiatric Illness*. New York, NY: W. W. Norton & Company.

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol Bull.* 135, 943–973. doi: 10.1037/a0017327

Schizophrenia Working Group of the Psychiatric Genomics Consortium, (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi: 10.1038/nature13595

Smerecnik, C. M., Mesters, I., Verweij, E., de Vries, N. K., and de Vries, H. (2009). A systematic review of the impact of genetic counseling on risk perception accuracy. *J. Genet. Couns.* 18, 217–228. doi: 10.1007/s10897-008-9210-z

Smit, A. K., Newson, A. J., Best, M., Badcock, C. A., Butow, P. N., Kirk, J., et al. (2018). Distress, uncertainty, and positive experiences associated with receiving information on personal genomic risk of melanoma. *Eur. J. Hum. Genet.* 26, 1094–1100. doi: 10.1038/s41431-018-0145-z

Tiller, J., and Lacaze, P. (2018). Regulation of internet-based genetic testing: challenges for australia and other jurisdictions. *Front. Public Health* 6:24. doi: 10.3389/fgene.2019.00024

Turnwald, B. P., Goyer, J. P., Boles, D. Z., Silder, A., Delp, S. L., and Crum, A. J. (2019). Learning one's genetic risk changes physiology independent of actual genetic risk. *Nat. Hum. Behav.* 3, 48–56. doi: 10.1038/s41562-018-0483-4

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592.

Watanabe, K., Stringer, S., Frei, O., Umiæeviæ Mirkov, M., de Leeuw, C., Polderman, T. J. C., et al. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* 51, 1339–1348. doi: 10.1038/s41588-019-0481-0

Wilhelm, K., Meiser, B., Mitchell, P. B., Finch, A. W., Siegel, J. E., Parker, G., et al. (2009). Issues concerning feedback about genetic testing and risk of depression. *Br. J. Psychiatry* 194, 404–410. doi: 10.1192/bjp.bp.107.047514

Wünnemann, F., Sin Lo, K., Langford-Avelar, A., Busseuil, D., Dubé, M. P., Tardif, J. C., et al. (2019). Validation of genome-wide polygenic risk scores for coronary artery disease in french canadians. *Circ. Genom. Precis. Med.* 12:e002481.

Young, M. A., Forrest, L. E., Rasmussen, V. M., James, P., Mitchell, G., Sawyer, S. D., et al. (2018). Making sense of SNPs: women's understanding and experiences of receiving a personalized profile of their breast cancer risks. *J. Genet. Couns.* 27, 702–708. doi: 10.1007/s10897-017-0162-z

Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflict of Interest: The authors declare that the voluntary donations received by impute.me go to a registered company, from where all of it is used to pay for server-costs. The company is a Danish-law IVS company with ID 37918806, financially audited under Danish tax law.

Copyright © 2020 Folkersen, Pain, Ingason, Werge, Lewis and Austin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

MINI REVIEW
published: 22 July 2020
doi: 10.3389/fgene.2020.00753

A Review on the Challenges in Indian Genomics Research for Variant Identification and Interpretation

Sandhya Kiran Pemmasani, Rasika Raman, Rajkishore Mohapatra, Mathukumalli Vidyasagar and Anuradha Acharya*

Research and Development Division, Mapmygenome India Limited, Hyderabad, India

Today, genomic data holds great potential to improve healthcare strategies across various dimensions – be it disease prevention, enhanced diagnosis, or optimized treatment. The biggest hurdle faced by the medical and research community in India is the lack of genotype-phenotype correlations for Indians at a population-wide and an individual level. This leads to inefficient translation of genomic information during clinical decision making. Population-wide sequencing projects for Indian genomes help overcome hurdles and enable us to unearth and validate the genetic markers for different health conditions. Machine learning algorithms are essential to analyze huge amounts of genotype data in synergy with gene expression, demographic, clinical, and pathological data. Predictive models developed through these algorithms help in classifying the individuals into different risk groups, so that preventive measures and personalized therapies can be designed. They also help in identifying the impact of each genetic marker with the associated condition, from a clinical perspective. In India, genome sequencing technologies have now become more accessible to the general population. However, information on variants associated with several major diseases is not available in publicly-accessible databases. Creating a centralized database of variants facilitates early detection and mitigation of health risks in individuals. In this article, we discuss the challenges faced by genetic researchers and genomic testing facilities in India, in terms of dearth of public databases, people with knowledge on machine learning algorithms, computational resources and awareness in the medical community in interpreting genetic variants. Potential solutions to enhance genomic research in India, are also discussed.

Keywords: clinical genomics, variant classification, Indian genomics research, Indian genomic databases, machine learning

OPEN ACCESS

Edited by:
Gustavo Glusman,
Institute for Systems Biology (ISB),
United States

Reviewed by:
Ignacio Macpherson,
International University of Catalonia,
Spain
Stylianos Chatzichronis,
National and Kapodistrian University
of Athens, Greece

****Correspondence:***
Sandhya Kiran Pemmasani
drsandhyakiran@mapmygenome.in

Specialty section:
This article was submitted to
ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

Received: *23 December 2019*
Accepted: *24 June 2020*
Published: *22 July 2020*

Citation:
Pemmasani SK, Raman R, Mohapatra R, Vidyasagar M and Acharya A (2020) A Review on the Challenges in Indian Genomics Research for Variant Identification and Interpretation. Front. Genet. 11:753. doi: 10.3389/fgene.2020.00753

INTRODUCTION

Dynamic migration history, ethnic and genetic diversity and a high degree of consanguinity contribute to the complex and heterogeneous nature of the Indian population. There are many known genetic diseases affecting different population subgroups and insufficient scientific resources to diagnose and treat them (Aggarwal and Phadke, 2015; GUaRDIAN Consortium et al., 2019). Large-scale genetic studies in Indian patients are required to study disease-causing mutations and to develop personalized treatment methods. Another important aspect is accurate analysis and interpretation of genetic data. While tried and tested statistical methods work fairly well for biomarker discovery, advanced solutions like machine learning algorithms bring a promise of genomics driven clinical solutions. In this article, we discuss the current scope of Indian genomics in healthcare, challenges in scientific resources and data analysis, and solutions to enhance genomic medicine in India.

CURRENT STATE OF GENETIC TESTING

Genetic testing in India has evolved in leaps and bounds in the past decade. Currently, there exist DNA-based tests that address multiple concerns in healthcare, from disease prevention to molecular diagnosis (Kar and Sivamani, 2016). In the case of preventive healthcare, genetic tests estimate the lifetime risk of disease, predisposition to biological traits and health parameters (Mohan et al., 2011). They also analyze a person's response to drugs in terms of efficacy and risk for adverse reactions. These tests are primarily used as screening tools for establishing an effective strategy to reduce disease risk, delay, or avoid symptoms and manage existing conditions. Diagnostic genetic tests, on the other hand, help in identification of the molecular cause of the disease. These tests are used to confirm known or suspected diagnosis, carrier status determination, identification of at-risk genetic relatives, optimize treatments, and clinical decisions (Gupta et al., 2017; Aravind et al., 2019; Uttarilli et al., 2019). There are different types of diagnostic genetic tests currently available in India such as single-gene and multigene testing, exome, and genome sequencing, carrier and newborn screening (Puri et al., 2017; Singh et al., 2018). Other types of tests include those which assess reproductive risk, such as prenatal testing and preimplantation genetic diagnosis (Dada et al., 2008).

Challenges and Limitations

Understanding the need of the patient is the key for determining the right genetic test. The biggest hurdles faced by clinicians are genetic data interpretation, finding genetic links for complex conditions, and lack of actionable genetic information. In certain cases of complex conditions, such as cancer, an array of genetic tests might be ordered to determine the genetic cause (Prabhash et al., 2019). However, no findings may come to light, thus posing a challenge for the patient and the clinician. The accuracy and precision of genetic tests lies in the translation of genetic findings into clinical outcomes. In the absence of information on genotype-phenotype correlations, genetic test results might be inconclusive.

GENETIC DATA INTERPRETATION FOR INDIAN PATIENTS

Genetic diagnosis via clinical sequencing (e.g., genome-, exome-, single-, or multi-gene) is the front-line test recommended for many inherited diseases (Verma et al., 2018; Ganapathy et al., 2019). Establishing the genetic cause of disease is vital for patient care and treatment, and hence clinical findings must be reported with high precision and accuracy (Singh et al., 2016). All clinical reporting protocols are required to adhere to standards set by American College of Medical Genetics (ACMG), for proper classification of variants and subsequent disclosure to patient/clinician (Richards et al., 2015). As per ACMG guidelines, in order to differentiate benign and pathogenic variants, a detailed study of the variant's clinical significance is required. This includes multiple criteria such as variant frequency, location in or near the gene, mechanism of said gene, effect of variant on protein domain or function, hotspot, or nearby mutations if any, etc. Apart from these, evidence of the variant having caused the disease in patients with similar clinical phenotype is essential to establish pathogenicity.

Challenges and Limitations

Currently, there is a dearth of publicly available resources that provide an extensive list of clinically significant variants in Indian patients, for several genetic diseases. In the absence of published literature for a particular variant, which clearly is not benign, the variant gets classified as a variant of uncertain significance (VOUS). Interpreting VOUS is often challenging as they are not actionable, yet hold potential for establishing pathogenicity. For accurate classification and high-precision reporting of genetic variants, it is vital that geneticists and scientists have access to information on the complete spectrum of variants and mutations in Indian patients (Rajasimha et al., 2014; Genomics and other Omics tools for Enabling Medical Decision, 2019). Only the most relevant mutations are listed in databases like OMIM, which use selection criteria such as frequency, phenotype, significance, disease mechanism, and inheritance, etc.

Case Study: Retinitis Pigmentosa

Retinitis Pigmentosa (RP) represents a very large group of eye disorders, with different clinical features, and symptoms. RP can be inherited in an autosomal dominant, autosomal recessive, or *X*-linked manner. Genetic diagnosis of RP helps in establishing genetic cause of disease, screening in at-risk family members and clinical management. But there are limited studies which report population specific mutations in Indian RP patients (**Table 1**).

Let us examine a case study of a 42 year old male reported with personal medical history of RP, who had been diagnosed at 16 years of age. His sister and two paternal cousins were also affected with RP. Married to a non-consanguineous partner, there were no clinical conditions in his children – 16 year old son and 12 year old daughter. Exome sequencing was done at Mapmygenome (Mapmygenome, 2020) to identify disease causing gene mutations associated with RP. Exome analysis revealed a heterozygous missense variant in exon 4 of the *NR2E3* gene. The observed variant is not reported as a variant in the normal samples of 1000 Genomes database and has a minor allele frequency of 0.018% in the gnomAD database. The variant is conserved across the species and *in silico* prediction by Mutation taster was found to be damaging. Another missense variant in this gene, Pro152Ser, has previously been reported with "retinitis pigmentosa 37," and "retinitis pigmentosa (recessive)" in clinvar (Clinvar, 2019) as VOUS. In Indian Genetic Disease Database

TABLE 1 | Information on summary of mutation studies in Indian RP patients.

Number of patients	Significant findings	Gene	Method	Variant information available	References
14 Families (autosomal recessive RP) and 100 cases (sporadic RP)	12 Novel mutations	EYS	Whole exome sequencing of proband followed by targeted analysis in family members	Yes	Di et al., 2016
1 Family and 100 cases (sporadic)	4 Novel mutations	CRB1		Yes	Yang et al., 2016
1 Family (autosomal recessive RP)	1 Novel mutation	MERTK		Yes	Bhatia et al., 2018b
1 Family (non-syndromic autosomal dominant RP)	1 Novel mutation	PRPF31	Targeted sequencing	Yes	Bhatia et al., 2018a
1 Family (autosomal recessive RP)	1 Novel mutation	MERTK	Microarray	Yes	Srilekha et al., 2015
2 Families (autosomal recessive RP) and 100 cases (sporadic RP)	1 Novel mutation	FAM161A	Whole exome sequencing	Yes	Zhou et al., 2015
2 Families (autosomal recessive RP)	1 Novel mutation	NR2E3	Microarray	Yes	Kannabiran et al., 2012
101 Cases (48 isolated cases and 53 autosomal dominant RP cases)	2 Novel mutations	PRPF31	Capillary sequencing	Yes	Gandra et al., 2008

(IGDD), which is the first patient based genetic disease database of India (Pradhan et al., 2011) only seven genes have been mapped for RP. No variant was reported from the *NRE23* gene in IGDD. There are no other databases which have mutation information from Indian RP patients. Since there is no functional or published study of *NR2E3* mutations in Indian RP patients, Mapmygenome had classified the variant in exon 4 of the gene as a VOUS. There is insufficient evidence to establish this variant's pathogenicity.

AVAILABLE RESOURCES WITH GENOTYPE-PHENOTYPE ASSOCIATIONS IN INDIAN POPULATION

Databases which host information on gene variants and associated diseases help genome analysts to make clinically significant and medically actionable inferences. However, most of the publicly available Indian databases are incomplete. This can be attributed to legal, ethical, financial, or administrative procedures due to which a lot of key parameters do not get recorded. Some of the Indian-specific databases, along with their scope and utility, have been discussed below.

Index-dB

A database of exonic variants from normal individuals of Indian sub-continent (Ahmed et al., 2019). It is a user-friendly database with a querying feature and a browser to search for the variants. But the current version is based only on 109 individuals and is still under development.

TMC-SNPdB

Contains variants generated from exome data of normal samples derived from tongue, gall bladder, and cervical cancer patients of Indian origin (Upadhyay et al., 2016). The major limitation of the database is not only the sample size of 62, but also the way the variants were processed. The COSMIC database was used to filter out somatic variants, because of which some novel Indian variants might have got filtered out.

SAGE

A repository of genetic variants derived through an integration of six datasets comprising 1213 South Asian genomes and exomes (Judith et al., 2018). It contains more than 154 million variants, out of which 69 million are novel variants. Though this a comprehensive database of South Asians, it should be enriched with region or ethnicity specific datasets within South Asia.

Indian Genetic Disease Database (IGDD)

A curated database of variants associated with diseases prevalent in Indian population (Pradhan et al., 2011). Diseases were categorized into different therapeutic areas. The current version of the database covers 104 diseases with a total of ∼3500 patients. Further enrichment is required to cover more diseases in the population.

Indian Genome Variation Database (IGVdB)

This was started as a consortium activity in 2003, with the goal to create a variation database of Indian population (Indian Genome Variation Consortium, 2005; Narang et al., 2010). However, this database does not contain disease-variant associations, which are helpful in interpreting the data obtained from genetic tests.

GWAS Central – India

A genotype-phenotype association database with summary level findings from genetic association studies (Indian GWAS, 2010). Lack of regular updates and absence of extensive data points for genetic diagnosis, make this database a less effective tool for clinicians, or bioinformaticians, thereby limiting its clinical utility.

Indian SNP Data

Contains genotype data of 871,771 SNPs, obtained from 15 Dravidian trios, and 13 Indo-European trios (Indian SNP, 2020). Browser and query features are not available for this database. Files can be downloaded for academic and research purposes only. Although it was initially developed as a reference panel for Indians, it has limited data and the work is still in progress.

Genotype/Phenotype DB

This database contains genotype and phenotype data of Indian population along with their demographic details (CCMB, 2020). Browser and query features are not available for this database. Commercial organizations are strictly prohibited from using the data.

Indigen Project

This is an initiative from Council of Scientific and Industrial Research (CSIR) for whole genome sequencing of 1000 Indian genomes, across diverse ethnic groups, with the goal to enable clinical applications in rare genetic diseases. This is an initiative which is yet to see fruition and is yet to be publicly available for the scientific community (IndiGen, 2020).

The above databases have not been presented in a way that allows the user to understand the pathogenicity of variants. Genomics companies like Mapmygenome (Mapmygenome, 2020), do not have access to most of such databases. A centralized database curated from Indian patients, for different diseases, would help in precise reporting and clinical decision making.

Publicly available data and results generated from genome wide association studies (GWAS) can also be utilized in interpreting the variants and in identifying new variants. There are case-control association studies done on Indian population, for majorly occurring diseases – Type 2 Diabetes, cardiovascular diseases and cancers (Chauhan et al., 2010; Nagrani et al., 2017; Bellary et al., 2019). Polygenic Risk Scores (PRS) developed from GWAS act as prognostic indicators in preventive healthcare. However, reliability of the results depends on the algorithm used and the data available.

MACHINE LEARNING (ML) ALGORITHMS IN INDIAN GENOMICS

With the availability of diverse data types – gene expression, SNP genotypes, demographics, heath history, laboratory findings, and images etc. – machine learning algorithms have become the obvious choice for accurate prediction of disease risk and personalized treatment. They can learn patterns underlying complex data and build models that can be used for prediction purposes. Numerous machine learning methods, such as support vector machines, random forests, and Bayesian networks, are being used successfully in genomics research and applications (Libbrecht and Noble, 2015; Xu and Jackson, 2019). Now, deep learning algorithms, a subcategory of machine learning, have emerged as the most successful algorithms for combining clinical data with genomics (Ching et al., 2018; Zou et al., 2019). They use artificial neural networks to progressively extract novel features from input data and learn from the features (Eraslan et al., 2019).

Deep learning and machine learning algorithms, which come under the umbrella term Artificial Intelligence (AI), are being used in clinical practice through numerous commercial applications involving clinical and genomics data. A well known personal genomics company, 23andme (2020) uses machine learning algorithms in disease risk prediction. IBM's Watson for Oncology (IBM Watson for Oncology, 2020) helps clinicians in identifying most appropriate treatment options based on information collated from medical records, medical journals, genomic journals, and relevant guidelines. Many startups are increasingly using the combination of machine learning algorithms and genomics in creating tools and processes that enhance the healthcare systems. For example, Freenome (2019), Benevolent AI (2020), Cambridge Cancer Genomics (2020), and DeepGenomics (2020) use AI in predicting disease risk, response to therapy and in developing personalized treatment regimens. In India, very few organizations use machine learning algorithms in clinical genomics, with the reasons being lack of awareness and lack of expertise in research and application of AI. Some of the Indian pharmaceutical and genomic organizations that are using AI include Innoplexus (2019), Lantern Pharma (2019), Manipal Group of Hospitals (2019), TCS Innovation Labs (2019), BioXcel Therapeutics Inc. (2020), Mapmygenome (2020), OncoStem (2020), and PierianDx (2020).

Challenges and Limitations

Main technical challenges in the application of ML algorithms are data curation and data pre-processing (Ngiam and Khor, 2019). Different hospitals and laboratories adopt different terminologies to record a disease or a health condition and use different reference ranges. In India, Electronic Health Records Standards were released by the Ministry of Health and Family Welfare in 2016. But sharing of data between the hospitals through a common platform is still a work in progress.

Data sets used in training the machine learning algorithms should clearly represent the target data for which risk predictions are made. For example, genetic algorithms trained on data from North Indians might make less accurate predictions when applied on South Indians. Comprehensive and robust clinical data sets that represent the ethnic differences among the people of India are still unavailable. To facilitate sharing of biological data across various research organizations in India, especially high-throughput data generated by sequencing and microarrays, and to create National Biological Data Centre, Ministry of Science and Technology has released zero draft on Biological data storage, access and sharing policy of India in July 2019 (Department of Science and Technology, 2020). But it is still in its nascent stage. A standard procedure for normalizing the raw data must be developed to maintain uniformity across the research groups.

Lack of understanding among clinicians and patients about the machine learning algorithms and their predictions make them considered as black box algorithms (Vayena et al., 2018). Data scientists should explain the general logic behind the algorithm-based decisions. Doctors and patients should understand the risk associated with such decisions. Clear communication between data scientists, doctors and patients is required to maintain ethical standards in clinical applications.

Solutions to Overcome the Challenges

NITI Aayog, a policy think tank of the government of India, made several recommendations to address the challenges and to harness the power of AI in India (National Strategy for AI, 2018). They include – establishing Centres of Research Excellence (COREs), increasing R&D resources, supporting Ph.D. researchers, establishing common supercomputing facilities, and creating an ecosystem for development and application of AI. Encouraging institute-industry partnerships, creating investment funds for AI startups and reskilling the existing workforce have also been discussed in detail. Other research agencies like Itihaasa (2018) made similar recommendations.

Institutional review boards, ethical review committees and scientific societies should come up with best practices for application of ML in clinical genomics. Government should start a regulatory body in lines similar to the United States Food and Drug Administration (FDA) to enforce best practices. Data sets used in training the algorithms, variables considered in building the models and accuracy of the predictions should be scrutinized. Updating the models by retraining the algorithms and checking the efficiency of the models should be done in coordination with the clinicians. The Government of India should take initiatives to train clinicians in understanding machine learning algorithms. Certification programs run through premier institutes would encourage the people to take up such courses.

ETHICAL AND LEGAL CONCERNS IN DATA SHARING

Genomic data is sensitive in nature and public sharing of such data brings a fair share of ethical and legal concerns with it. Given the increasing number of direct-to-consumer tests that are available, there is a need to streamline certain processes. The collection, storage and usage of genetic data must enable meaningful outcomes for personalized medicine. Data security and privacy remains one of the major concerns reported by users. The “Personal Genomes: Accessing, Sharing and Interpretation” conference held in the United Kingdom, in April 2019 (Genetics Society, 2019) addressed several conundrums which hinder sharing of genetic and medical data, for the creation and maintenance of genomic databases. There is also a growing segment of users who are open to sharing their de-identified data (Kim et al., 2015; Rubin and Glusman, 2019). They share their data for getting updates on their health reports, for providing social good or for financial compensation (Hendricks-Sturrup and Lu, 2020). In India, with the release of Personal Data Protection Bill 2019 (The Personal Data Protection Bill, 2019) certain principles were laid down on collection and usage of personal data. Informed consent, data minimization and storing a copy of data within India are some of the essential requirements under the bill.

The benefits of sharing genomic data in the scientific community are far too many to ignore. Collaborative efforts between sequencing facilities, data scientists, clinics, and healthcare providers must be directed toward building a healthy ecosystem for data sharing. De-identification of the genetic information as well as medical records is essential. Wright et al. (2019) proposes a system wherein genetic variant details and their associated conditions can be shared in online databases, without requiring explicit consent from patients. However, detailed clinical information and case study at a deeper level will require consent from the doctors and their patients. For the Indian scenario, a robust system for data sharing is required. This system must be regulated by measures which protect the patients’ interests as well. Policy makers and leaders must come together to develop a framework that allows more variant databases to become publicly accessible, without breach of privacy.

ROLE OF CLINICIANS IN INDIAN GENOMICS

Clinicians play a very important role in facilitating genomics-driven healthcare. From the time a patient visits the clinic to the time of treatment, there are several stages that require the clinician to relay information related to testing procedures and their possible outcomes. The clinician holds a key responsibility of comprehending the implications of genetic findings and making the necessary correlations for treatment and management. Hence, it is imperative that the clinician is well versed with different genetic mechanisms, inheritance, gene-gene, and gene-environment interaction mechanisms, variants and their pathogenicity. In the clinic, staff must be trained to perform timely reviews of clinical and family history and identify cases which warrant genetic testing. For the current generation of clinicians, training on genetic diseases, testing methodologies, clinical variant interpretation and application in medicine, must be included as part of their continuing education. Policy makers such as Medical Council of India and Board of Education play an important role in training clinicians on utilizing genomics in their practice (Scheuner et al., 2008; Aggarwal and Phadke, 2015).

CONCLUSION

Given the broad spectrum of genetic diseases and their burden on the Indian population, it is essential for genomic researchers to tap Indian genetic data for disease prevention, timely diagnosis, and treatment. Studies show that there are novel mutations in Indian patients, for different phenotypes. Hence, genome analysts need to refer to Indian-specific databases for meaningful translation of genomics data into clinical reporting. Current challenges can be met by united efforts from government health agencies and genetic research institutes by executing large scale sequencing projects, accompanied by detailed documentation on patients' clinical features and family history. Obtaining informed consent from the patients must be mandatory, to protect their interests including concerns about data privacy and safety. The patients must be educated about protocols such as de-identification, data security and research objectives.

Novel variants must be made available in a centralized database for analysts to refer to, and draw inferences from. Such a database would vastly improve the diagnostic accuracy of genetic diseases. Indian genomics will also greatly benefit by the development of machine learning algorithms for analyzing health trends in the Indian population. Additionally, clinicians from all walks of medicine must be equipped with technical knowledge on medical genetics and its clinical application, for enhanced patient care.

AUTHOR CONTRIBUTIONS

SP and RR have contributed conception and design of the study. SP, RR, and RM wrote sections of the manuscript. MV and AA have supervised and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

23andme (2020). *23andme*. Available online at: https://www.23andme.com/ (accessed November, 2019).

Aggarwal, S., and Phadke, S. R. (2015). Medical genetics and genomic medicine in India: current status and opportunities ahead. *Mol. Genet. Genomic Med.* 3, 160–171. doi: 10.1002/mgg3.150

Ahmed, P. H., Viswanath, V., More, R. P., Viswanath, B., Jain, S., Rao, M. S., et al. (2019). INDEX-db: the indian exome reference database (Phase I). *J. Comput.* 26, 225–234. doi: 10.1089/cmb.2018.0199

Aravind, S., Ashley, B., Mannan, A., Ganapathy, A., Ramesh, K., Ramachandran, A., et al. (2019). Targeted sequencing of the DMD locus: a comprehensive diagnostic tool for all mutations. *Indian J. Med. Res.* 150, 282–289. doi: 10.4103/ijmr.IJMR_290_18

Bellary, K., Dwarkanath, K. M., Nagalla, B., and Mohini, T. A. (2019). Genetic variants of chromosome 9p21.3 region associated with coronary artery disease and premature coronary artery disease in an Asian Indian population. *Indian Heart J.* 71, 263–271. doi: 10.1016/j.ihj.2019.04.005

Benevolent AI (2020). *Benevolent AI*. Available online at: https://benevolent.ai/ (accessed November, 2019).

Bhatia, S., Goyal, S., Singh, I., Singh, D., and Vanita, V. (2018a). A novel mutation in the PRPF31 in a North Indian adRP family with incomplete penetrance. *Doc. Ophthalmol.* 137, 103–119. doi: 10.1007/s10633-018-9654-x

Bhatia, S., Kaur, N., Singh, I., and Vanita, V. (2018b). A novel mutation in MERTK for rod-cone dystrophy in a North Indian family. *Can. J. Ophthalmol.* 54, 40–50. doi: 10.1016/j.jcjo.2018.02.008

BioXcel Therapeutics Inc. (2020). *BioXcel Therapeutics Inc*. Available online at: https://www.bioxceltherapeutics.com/ (accessed November, 2019).

Cambridge Cancer Genomics (2020). *Cambridge Cancer Genomics*. Available online at: https://www.ccg.ai/ (accessed November, 2019).

CCMB (2020). *Genotype/Phenotype dB*. Available online at: https://www.ccmb.res.in/bic/database_pagelink.php?page=genotype (accessed November, 2019).

Chauhan, G., Spurgeon, C. J., Tabassum, R., Bhaskar, S., Kulkarni, S. R., Mahajan, A., et al. (2010). Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians. *Diabetes Metab. Res. Rev* 59, 2068–2074. doi: 10.2337/db09-1386

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15:20170387. doi: 10.1098/rsif.2017.0387

Clinvar (2019). *Clinvar*. Available online at: https://www.ncbi.nlm.nih.gov/clinvar (accessed November, 2019).

Dada, R., Kumar, R., Shamsi, M. B., Tanwar, M., Pathak, D., Venkatesh, S., et al. (2008). Genetic screening in couples experiencing recurrent assisted procreation failure. *Indian J. Biochem. Biophys.* 45, 116–120.

DeepGenomics (2020). *DeepGenomics*. Available online at: https://www.deepgenomics.com/ (accessed November, 2019).

Department of Science and Technology (2020). *Department of Science, and Technology*. Available online at: https://dst.gov.in/ (accessed November, 2019).

Di, Y., Huang, L., Sundaresan, P., Li, S., Kim, R., Ballav Saikia, B., et al. (2016). Whole-exome sequencing analysis identifies mutations in the eys gene in retinitis pigmentosa in the indian population. *Sci. Re.* 6:19432. doi: 10.1038/srep19432

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403. doi: 10.1038/s41576-019-0122-6

Freenome (2019). *Freenome*. Available online at: https://www.freenome.com/ (accessed November, 2019).

Ganapathy, A., Mishra, A., Soni, M. R., Kumar, P., Sadagopan, M., Kanthi, A. V., et al. (2019). Multi-gene testing in neurological disorders showed an improved diagnostic yield: data from over 1000 Indian patients. *J. Neurol.* 266, 1919–1926. doi: 10.1007/s00415-019-09358-1

Gandra, M., Anandula, V., Authiappan, V., Sundaramurthy, S., Raman, R., Bhattacharya, S., et al. (2008). Retinitis pigmentosa: mutation analysis of RHO, PRPF31, RP1, and IMPDH1 genes in patients from India. *Mol. Vis.* 14, 1105–1113.

Genetics Society (2019). *Genetics Society*. Available online at: https://genetics.org.uk/events/personal-genomes-accessing-sharing-and-interpretation/ (accessed November, 2019).

Genomics and other Omics tools for Enabling Medical Decision (2019). *Genomics and other Omics tools for Enabling Medical Decision*. Available online at: http://gomed.igib.in/ (accessed November, 2019).

GUaRDIAN Consortium, Sivasubbu, S., and Scaria, V. (2019). Genomics of rare genetic diseases-experiences from India. *Hum. Genomics* 14:52. doi: 10.1186/s40246-019-0215

Gupta, S., Chaurasia, A., Pathak, E., Mishra, R., Chaudhry, V. N., Chaudhry, P., et al. (2017). Whole exome sequencing unveils a frameshift mutation in CNGB3 for cone dystrophy: a case report of an Indian family. *Medicine* 96:e7490. doi: 10.1097/MD.0000000000007490

Hendricks-Sturrup, R., and Lu, C. (2020). What motivates the sharing of consumer-generated genomic information? *SAGE Open Med.* 8:205031212091540. doi: 10.1177/2050312120915400

IBM Watson for Oncology (2020). *IBM Watson for Oncology*. Available online at: https://www.ibm.com/watson-health/oncology-and-genomics (accessed November, 2019).

Indian Genome Variation Consortium (2005). The Indian genome variation database (IGVdb): a project overview. *Hum. Genet.* 118, 1–11. doi: 10.1007/s00439-005-0009-9

Indian GWAS (2010). *Centra*. Available online at: https://vigeyegpms.in/gpmsv2/gwascentralindia/.

Indian SNP (2020). *Database*. Available online at: https://www.ccmb.res.in/bic/database_pagelink.php?page=snpdata (accessed November, 2019).

IndiGen (2020). *Project*. Available online at: https://indigen.igib.in/ (accessed November, 2019).

Innoplexus (2019). *Innoplexus*. Available online at: https://www.innoplexus.com/ (accessed November, 2019).

Itihaasa (2018). *Itihaasa*. Available online at: http://www.itihaasa.com/pdf/itihaasa_AI_Research_Report.pdf (accessed November, 2019).

Judith, M. H., Shamsudheen, K. V., Ankit, V., Anop, S. R., Rijith, J., Rowmika, R., et al. (2018). SAGE: a comprehensive resource of genetic variants integrating South Asian whole genomes and exomes. *Database* 2018:bay080. doi: 10.1093/database/bay080

Kannabiran, C., Singh, H., Sahini, N., Jalali, S., and Mohan, G. (2012). Mutations in TULP1, NR2E3, and MFRP genes in Indian families with autosomal recessive retinitis pigmentosa. *Mol. Vis.* 18, 1165–1174.

Kar, B., and Sivamani, S. (2016). Directory of genetic test services and counselling centres in India. *Int J Hum Genet.* 16, 148–157. doi: 10.1080/09723757.2016.11886292

Kim, K. K., Joseph, J. G., and Ohno-Machado, L. (2015). Comparison of consumers' views on electronic data sharing for healthcare and research. *J. Am. Med. Inform. Assoc.* 22, 821–830. doi: 10.1093/jamia/ocv014

Lantern Pharma (2019). *Lantern Pharma*. Available online at: https://www.lanternpharma.com/ (accessed November, 2019).

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Manipal Group of Hospitals (2019). *Collaboration with IBM's Watson for Oncology*. Available online at: https://www.manipalhospitals.com/ (accessed November, 2019).

Mapmygenome (2020). *Mapmygenome*. Available online at: https://mapmygenome.in/ (accessed November, 2019).

Mohan, V., Goldhaber-Fiebert, J. D., Radha, V., and Gokulakrishnan, K. (2011). Screening with OGTT alone or in combination with the Indian diabetes risk score or genotyping of TCF7L2 to detect undiagnosed type 2 diabetes in Asian Indians. *Indian J. Med. Res.* 133, 294–299.

Nagrani, R., Mhatre, S., Rajaraman, P., Chatterjee, N., Akbari, M. R., Boffetta, P., et al. (2017). Association of genome-wide association study (GWAS) identified SNPs and risk of breast cancer in an indian population. *Sci. Rep.* 7:40963. doi: 10.1038/srep40963

Narang, A., Roy, R. D., Chaurasia, A., Mukhopadhyay, A., and Mukerji, M. (2010). Indian genome variation consortium. Das D. IGVBrowser–a genomic variation resource from diverse Indian populations. *Database* 2010:baq022. doi: 10.1093/database/baq022

National Strategy for AI (2018). *National Strategy for AI*. Available online at: https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf (accessed November, 2019).

Ngiam, K. Y., and Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 20, e262–e273. doi: 10.1016/s1470-2045(19)30149-4

OncoStem (2020). *OncoStem*. Available online at: https://www.oncostem.com/ (accessed November, 2019).

PierianDx (2020). *PierianDx*. Available online at: https://www.pieriandx.com/ (accessed November, 2019).

Prabhash, K., Advani, S. H., Batra, U., Biswas, B., Chougule, A., Ghosh, M., et al. (2019). Biomarkers in non-small cell lung cancers: indian consensus guidelines for molecular testing. *Adv. Ther.* 36, 766–785. doi: 10.1007/s12325-019-00903-y

Pradhan, S., Sengupta, M., Dutta, A., Bhattacharyya, K., Bag, S., Dutta, C., et al. (2011). Indian genetic disease database. *Nucleic Acids Res.* 39(Suppl. 1), D933–D938. doi: 10.1093/nar/gkq1025

Puri, R. D., Tuteja, M., and Verma, I. C. (2017). genetic approach to diagnosis of intellectual disability. *Indian J. Pediatr.* 83, 1141–1149. doi: 10.1007/s12098-016-2205-0

Rajasimha, H. K., Shirol, P. B., Ramamoorthy, P., Hegde, M., Barde, S., Chandru, V., et al. (2014). Organization for rare diseases India (ORDI) - addressing the challenges and opportunities for the Indian rare diseases' community. *Genet. Res.* 96:e009. doi: 10.1017/S0016672314000111

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Rubin, I. R., and Glusman, G. (2019). Opportunities and challenges in interpreting and sharing personal genomes. *Genes* 10:643. doi: 10.3390/genes10090643

Scheuner, M. T., Sieverding, P., and Shekelle, P. G. (2008). Delivery of genomic medicine for common chronic adult diseases: a systematic review. *JAMA* 299, 1320–1334. doi: 10.1001/jama.299.11.1320

Singh, B., Mandal, K., Lallar, M., Narayanan, D. L., Mishra, S., Gampbhir, P. S., et al. (2018). Next generation sequencing in diagnosis of MLPA Negative cases presenting as duchenne/ becker muscular dystrophies. *Indian J. Pediatr.* 85, 309–310. doi: 10.1007/s12098-017-2455-5

Singh, J., Mishra, A., Pandian, A. J., Mallipatna, A. C., Khetan, V., Sripriya, S., et al. (2016). Next-generation sequencing-based method shows increased mutation detection sensitivity in an Indian retinoblastoma cohort. *Mol. Vis.* 22, 1036–1047.

Srilekha, S., Arokiasamy, T., Srikrupa, N. N., Umashankar, V., Meenakshi, S., Sen, P., et al. (2015). Homozygosity mapping in leber congenital amaurosis and autosomal recessive retinitis pigmentosa in south indian families. *PLoS One* 10:e0131679. doi: 10.1371/journal.pone.0131679

TCS Innovation Labs (2019). *TCS Innovation Labs*. Available online at: https://www.tcs.com/reimagining-drug-safety-powered-by-genomics-information-integration-and-emerging-technologies (accessed November, 2019).

The Personal Data Protection Bill (2019). *The Per. sonal Data Protection Bill*. Available online at: https://www.prsindia.org/billtrack/personal-data-protection-bill-2019 (accessed November, 2019).

Upadhyay, P., Gardi, N., Desai, S., Sahoo, B., Singh, A., Togar, T., et al. (2016). TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. *Database* 2016:baw104. doi: 10.1093/database/baw104

Uttarilli, A., Shah, H., Bhavani, G. S., Upadhyai, P., Shukla, A., and Girisha, K. M. (2019). Phenotyping and genotyping of skeletal dysplasias: evolution of a center and a decade of experience in India. *Bone* 120, 204–211. doi: 10.1016/j.bone.2018.10.026

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS Med.* 15:e1002689. doi: 10.1371/journal.pmed.1002689

Verma, I. C., Paliwal, P., and Singh, K. (2018). Genetic Testing in Pediatric Ophthalmology. *Indian J. Pediatr.* 85, 228–236. doi: 10.1007/s12098-017-2453-7

Wright, C. F., Ware, J. S., Lucassen, A. M., Hall, A., Middleton, A., Rahman, N., et al. (2019). Genomic variant sharing: a position statement. *Wellcome Open Res.* 4:22. doi: 10.12688/wellcomeopenres.15090.2

Xu, C., and Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol.* 20:76. doi: 10.1186/s13059-019-1689-0

Yang, Y., Yang, Y., Huang, L., Zhai, Y., Li, J., Jiang, Z., et al. (2016). Whole exome sequencing identified novel CRB1 mutations in Chinese and Indian populations with autosomal recessive retinitis pigmentosa. *Sci. Rep.* 6:33681. doi: 10.1038/srep33681

Zhou, Y., Saikia, B., Jiang, Z., Zhu, X., Liu, Y., Huang, L., et al. (2015). Whole-exome sequencing reveals a novel frameshift mutation in the FAM161A gene causing autosomal recessive retinitis pigmentosa in the Indian population. *J. Hum. Genet.* 60:625. doi: 10.1038/jhg.2015.92

Zou, J., Huss, M., Abid, A., Pejman, M., Ali, T., and Amalio, T. (2019). A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18. doi: 10.1038/s41588-018-0295-5

Conflict of Interest: SP, RR, RM, and AA are employed by the company Mapmygenome India Limited. MV is a member of Scientific Advisory Board at Mapmygenome India Limited.

Copyright © 2020 Pemmasani, Raman, Mohapatra, Vidyasagar and Acharya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

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/

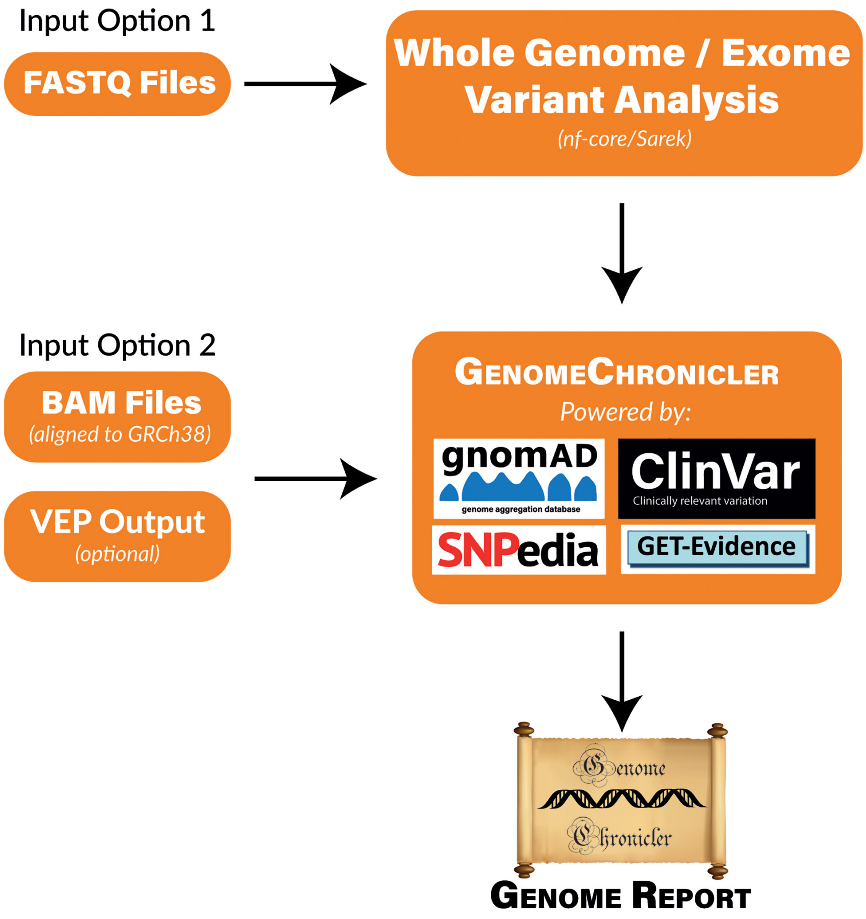


FIGURE 1 | Flow Diagram of GenomeChronicler processing pipeline, illustrating the multiple entry points for the pipeline, resources integrated by default and generated outcomes. Either entry point of the pipeline can be run locally in a single machine, as a Nextflow workflow or in the Cloud. All source 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

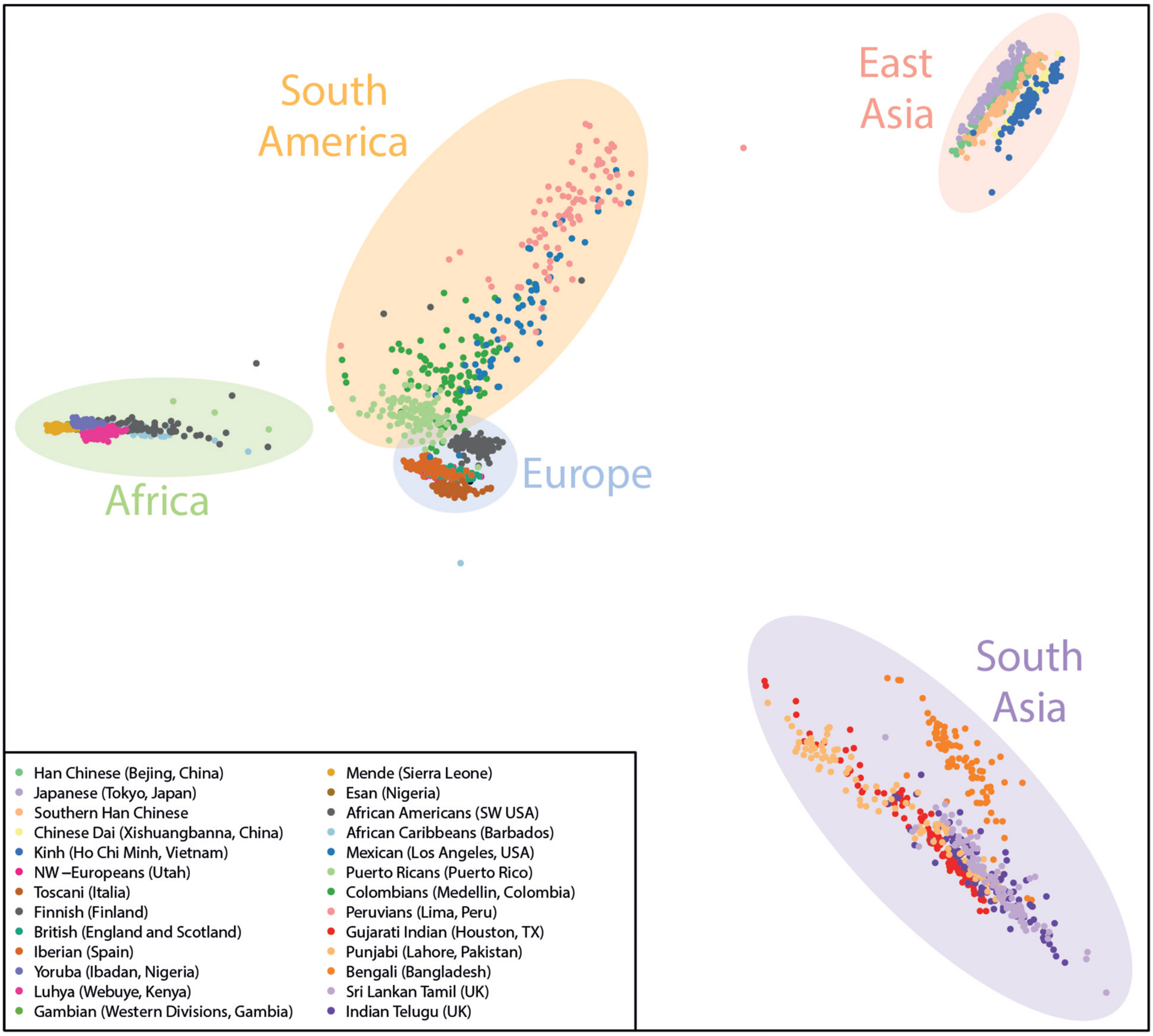


FIGURE 2 | Example Ancestry PCA plot containing the current reference data from the 1000 genomes project used by GenomeChronicler, with shaded areas broadly illustrating the origin of the populations represented.

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.

Copyright © 2020 Guerra-Assunção, Conde, Moghul, Webster, Ecker, Chervova, Chatzipantsiou, Prieto, Beck and Herrero. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2021
doi: 10.3389/fgene.2021.535123

Whole Genome Interpretation for a Family of Five

Manuel Corpas[1,2,3]*, *Karyn Megy*[1,4], *Vanisha Mistry*[5], *Antonio Metastasio*[1,6] and *Edmund Lehmann*[1]

[1] Cambridge Precision Medicine Limited, ideaSpace, University of Cambridge Biomedical Innovation Hub, Cambridge, United Kingdom, [2] Institute of Continuing Education Madingley Hall Madingley, University of Cambridge, Cambridge, United Kingdom, [3] Facultad de Ciencias de la Salud, Universidad Internacional de La Rioja, Madrid, Spain, [4] Department of Haematology, University of Cambridge & National Health Service (NHS) Blood and Transplant, Cambridge, United Kingdom, [5] Fabric Genomics, Oakland, CA, United States, [6] Camden and Islington NHS Foundation Trust, London, United Kingdom

Although best practices have emerged on how to analyse and interpret personal genomes, the utility of whole genome screening remains underdeveloped. A large amount of information can be gathered from various types of analyses via whole genome sequencing including pathogenicity screening, genetic risk scoring, fitness, nutrition, and pharmacogenomic analysis. We recognize different levels of confidence when assessing the validity of genetic markers and apply rigorous standards for evaluation of phenotype associations. We illustrate the application of this approach on a family of five. By applying analyses of whole genomes from different methodological perspectives, we are able to build a more comprehensive picture to assist decision making in preventative healthcare and well-being management. Our interpretation and reporting outputs provide input for a clinician to develop a healthcare plan for the individual, based on genetic and other healthcare data.

Keywords: whole genome sequencing, personal genomics, interpretation, precision medicine, genetic risk score, pharmacogenomics, nutrigenomics

OPEN ACCESS

Edited by:
Mehdi Pirooznia,
National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Bastian Greshake Tzovaras,
Centre de Recherches Interdisciplinaires (CRI), France
Michal Korostynski,
Institute of Pharmacology PAS, Poland

Correspondence:
Manuel Corpas
m.corpas@cpm.onl

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

Received: *14 February 2020*
Accepted: *15 February 2021*
Published: *08 March 2021*

Citation:
Corpas M, Megy K, Mistry V, Metastasio A and Lehmann E (2021) Whole Genome Interpretation for a Family of Five. Front. Genet. 12:535123. doi: 10.3389/fgene.2021.535123

INTRODUCTION

A great deal of literature has been generated over the past decade defining best practices for clinical interpretation of personal genomes (Nykamp et al., 2017; Biesecker et al., 2018; Brandt et al., 2019; Machini et al., 2019). Some additional approaches involve the simultaneous analysis of parents and child, for example in the case of pediatric diagnosis for children with rare diseases (Wright et al., 2018). Other studies have used family genomes to assign the precise chromosomal position of variants (Roach et al., 2011). To our knowledge, however, the use of genome analysis for screening and disease prevention remains underdeveloped. To address this shortcoming, our current study sheds light on two areas. First, we provide a comprehensive whole genome analysis of pathogenicity screening, genetic risk, pharmacogenomic, fitness, and nutrition trait analysis. Second, we discuss the joint interpretation of these results within the perspective of a family of five for whom we have deep phenotypic knowledge, allowing us to find "true positive" predictions based on the family observations.

In the past, we have performed assessment of personal genome analysis for the same set of family individuals using direct to consumer data and crowdsourcing methods (Glusman et al., 2012; Corpas et al., 2015). We were limited by the amount of data available at the time (DNA chip or exome) as well as a lack of reference data sources and analysis platforms to help with the interpretation that have appeared in recent years [e.g., gnomAD (Karczewski et al., 2019), ClinVar (Landrum et al., 2014)]. In this new iteration, we perform whole genome sequencing analysis for the same family of five and expand from our previous research to encompass a more comprehensive set of analyses and individual genomic data following published standard practice for interpretation of results as much as possible. Following standard practice is not always possible given that authoritative guidelines for interpretation of variants (Richards et al., 2015) are mostly applied to pathogenicity screening, rather than preventative healthcare using personal genomes. We were indeed able to perform a pathogenicity screening for the five members of the family quintet. For four genomes we also report genetic risk scores for 49 phenotypes using published Genome Wide Association Study (GWAS) markers (see **Supplementary Table 1**). We make a distinction between our genetic marker score notation and polygenic risk scores in the literature (Khera et al., 2018; Georgi et al., 2019; Meisner et al., 2019; Palmer, 2019; Torkamani and Topol, 2019) as we only use markers reported above a certain threshold of probability (as defined by GWAS studies).

To date, genome analysis of pathogenicity screening, genetic risk scoring for cardiovascular disease and some pharmacogenomics characterization has been performed by the MedSeq project for 100 individuals (Machini et al., 2019). Compared to this study, we offer novel perspectives on several fronts: (1) Our genetic risk analysis encompasses mental, metabolic, and autoimmune diseases, in addition to only cardiovascular being done in the MedSeq Project. (2) We include a systematic curation of known nutrition and fitness markers following newly developed guidelines to evaluate the scientific validity of gene x lifestyle interventions (Grimaldi et al., 2017). (3) Our deep phenotype and clinical knowledge of analyzed participants, helps us interpret and report results in a familial context within a wellness and prevention point of view. In addition, this work provides a proof-of-principle approach about an application of genetic risk scores within a family-oriented preventative healthcare and well-being case, recognizing that we are studying only one family and therefore this represents only an illustration of our proposed methodology for comprehensive whole genome analysis. Whenever possible, we use established guidelines from the American College for Medical Genetics and Genomics (ACMG), Food and Drug Administration (FDA), the Clinical Pharmacogenetics Implementation Consortium (CPIC), and other specialized organizations. We also discuss how different whole genome analysis methods can be integrated into more actionable outcomes for the individual and his or her relatives.

METHODS

Ethical Framework

This project builds on prior work (Glusman et al., 2012; Corpas et al., 2015). We started as an open source project in 2010 using the data available from direct to consumer providers. As the project evolved and exome sequencing was performed, a consent form was created and signed for the collection of samples, analysis, and publishing of results. This form identified participants as voluntary donors of their genetic data to the public domain and educated participants, making them aware of the potential discomforts and risks that doing this research might bring.

Here we base our analysis on the whole genome rather than the exome. To facilitate this work, further collection of samples has been performed in order to sequence and analyse whole personal genomes for this family. All participants underwent a new consent process and signed a consent form accepting the terms and conditions of this work as well as the potential consequences of performing such analysis. When developing the consent framework, we drew on the Personal Genome Project UK (PGP-UK Consortium, 2018) as an example of a rigorous approach to informed consent. As a result, the consent process developed for this work included the following elements: (a) participants underwent extensive training on the risks of genetic analysis including the risks of publishing personal genetic data; (b) participants completed an exam to demonstrate their comprehension of the risks and protocols associated with participating in genetic analysis which may be published and (c) participants were judged truly capable of giving informed consent. Consent forms were signed by all family individuals or their next to kin (in the case of a deceased member). This ethical framework has been independently assessed and approved by the Ethics Committee of Universidad Internacional de La Rioja (code PI:029/2020).

Family Dataset

We selected this family dataset for two reasons: (1) We have performed and published in the past decade two studies describing state of the art personal genomics analysis for a family of related individuals using array chip data and Illumina exome data (Glusman et al., 2012; Corpas et al., 2015). (2) The accumulated genetic studies and follow up of the disease and lifestyle history of the family through their continuous research have afforded us a deep knowledge of their phenotypes and disease history. **Figure 1** shows the family pedigree. In it we have individuals PT00010A (Aunt), who is the sister of PT00008A (Mother). PT00007A (Father) is Mother's spouse and both have two children (PT00009A and PT00002A; Daughter and Son). From here onwards, and for simplicity, we refer to family members as (Aunt, Father, Mother, Daughter, Son). All individuals of the family except Aunt had their DNA sequenced from saliva, whereas Aunt's DNA was sequenced from hair (see **Supplementary Materials** for details). This is because at the time of sample collection Aunt was already deceased (see next section for phenotypic details). Thirty-six hairs were retrieved

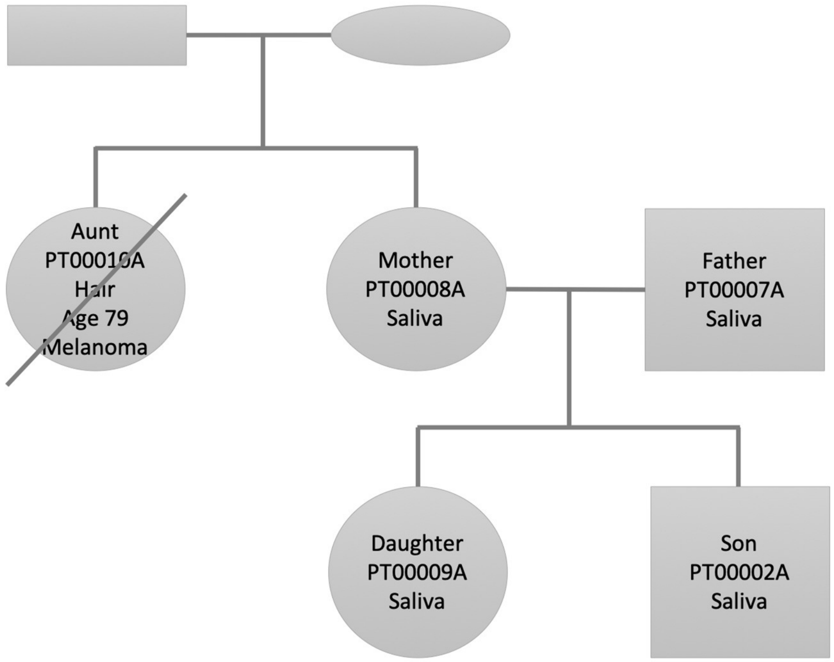


FIGURE 1 | Family pedigree showing the relationship, gender (square: male, circle: female), and sample used for whole genome sequencing (saliva/hair). The crossed circle indicates a deceased individual.

from a personal comb only she used and her DNA extracted from hair roots using a different protocol described in the **Supplementary Materials**.

When we analyzed the variant output of all samples, we benchmarked against Fabric Genomics Clinical Grade Scoring Rules (http://help.fabricgenomics.com/hc/en-us/articles/206433937-Appendix-4-Clinical-Grade-Scoring-Rules; accessed 7/January/2020), where Clinical Grade is a measure of a variant file's overall quality and fitness for clinical interpretation. The hair sample failed the criteria for clinical-grade coverage, genotype quality, homozygous/heterozygous ratio, and transition/transversion ratio (**Table 1**).

We performed a further analysis of quality of variants by counting those that pass the default standard filters of quality for interpretation given our analysis software (**Table 2**; see **Supplementary Materials**). For Aunt, we eliminated all variants below the threshold of QUAL < 20. The performance of the variant count and the level of coverage was sufficient to include Aunt in pathogenicity screening, but not sufficient for participation in the rest of the analysis.

Family History of Lifestyle and Disease

We conducted research into the family disease and lifestyle history. This research consisted of face-to-face interviews with all family members, during which they were asked about past illnesses, hospitalizations, reasons of death for past relatives and any ongoing condition that they think might related to the phenotypes and traits we analyse in this study. At the time of our last interview (October 2020), Mother and Father are in their mid-eighties, a similar age Aunt would be, had she not passed due to metastasised melanoma at age 79. Daughter is in her late fifties and Son mid-forties. All members of the family have been diagnosed obese or overweight at some point in their adulthood years. Childhood obesity was present in both Son and Daughter. Mother had a benign breast tumor removed in her early forties. She has also suffered from a history of low blood pressure and was diagnosed with chronic inflammation of her colon in her sixties, suffering from lower abdominal pain ever since. Father has a history of high blood pressure and heart problems. He has recently been diagnosed with atrial fibrillation. He displays difficulty breathing at moderate exertion levels and has been taking anticoagulants to prevent thromboembolism as a consequence of his atrial fibrillation, with some episodes of adverse drug reactions to warfarin. He is suspected to be lactose intolerant. In addition to her metastasised melanoma, before Aunt's passing she suffered from several episodes of venous thromboembolism, treated with anticoagulants (warfarin). There is no history in the family of diabetes or Parkinson's disease, although the father of both Mother and Aunt was diagnosed with Alzheimer's disease in his mid-eighties. Apart from Aunt's melanoma, there is no history of any other malignancy known to the family, no major mental health episodes or alcohol dependence diagnosed to date. All family members except Mother reported being light smokers for a period of their lives, all having quit more than a decade ago except daughter who still smokes several cigarettes a day.

Pathogenicity Screening

All single nucleotide variants and indels were filtered according to three different gene panels: (1) genes present in the

TABLE 1 | Statistics for clinical grade measures of the quality of the variant file.

Sample ID	Coverage	Genotype quality	Homozygous/ heterozygous ratio	Transition/ transversion ratio
PT00002A (Son)	43.0*	94.3	0.51*	2.81*
PT00007A (Father)	25.0	95.9*	0.51*	2.81*
PT00008A (Mother)	24.0	95.7*	0.51*	2.79*
PT00009A (Daughter)	29.0	97.8*	0.48	2.79*
PT00010A (Aunt)	2.0	4.7	0.11	1.06

Star-marked values (*) indicates the quality is of clinical standards and no-star values that it is below clinical standards (see Fabric Genomics Clinical Grade Scoring Rules [http://help.fabricgenomics.com/hc/en-us/articles/206433937-Appendix-4-Clinical-Grade-Scoring-Rules]). Coverage in values with a star indicates that the median coverage of coding variants exceeds 40. Genotype quality with a starred value: more than 95% of the coding variants have a quality above 40. Starred homozygous/heterozygous ratio: the ratio for the coding variants is between 0.5 and 0.61. Starred transition/transversion ratio: The ratio for the coding variants is between 2.71 and 3.08. None of the quality measures for clinical grade sampling was met by Aunt, whereas clinical grade quality measures are reached by other individuals.

TABLE 2 | The total number of variants for all saliva samples and the total number of coding variants for each family member.

Sample ID	Total number of variants	Total number of coding variants
PT00002A (Son)	4,956,742	27,286
PT00007A (Father)	4,650,536	27,504
PT00008A (Mather)	4,695,886	27,329
PT00009A (Daughter)	4,812,818	27,400
PT00010A (Aunt)	970,018	16,182

OMIM morbid list (Amberger et al., 2015), (2) ACMG 59 genes (Kalia et al., 2017), and (3) a Hereditary Cancer panel of 52 genes (https://info.fabricgenomics.com/ace; accessed 10/February/2020). All three panels required pathogenic or likely pathogenic alleles matching ClinVar (Landrum et al., 2014) evidence. The variant prioritization was based on their ClinVar evidence, their frequency in gnomAD (Karczewski et al., 2019), the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and their predicted variant effect (i.e., loss of function, non-synonymous or other). Our selection of frequency threshold is based on the gnomAD database (https://gnomad.broadinstitute.org/faq) criteria of common variant sites, defined as frequency >0.01. For each of the variants that passed the filtering, we classified them following the guidelines proposed by the ACMG (Richards et al., 2015) into 5 categories; from most to least pathogenic these categories are: pathogenic, likely pathogenic, uncertain significance, likely benign, benign. Relevant scientific literature as well as a number of algorithms were also used to assess each prioritized variant [i.e., OMICIA (Coonrod et al., 2013), VAAST (Hu et al., 2013), VVP (Flygare et al., 2018), and CADD (Rentzsch et al., 2019)].

Genetic Risk Scores

Genetic risk scores, also called genetic predisposition scores, aim to quantify the cumulative effects of a number of variants affecting multiple genes, which may individually confer only small risk susceptibility. Genetic risk scores are not diagnostic, as a high-risk score does not necessarily mean that a person will develop a condition, and a low score does not mean that they will not develop it. Nevertheless, genetic risk scores may be pointers for further exploration when looking for potential preventative interventions, particularly for multigenic conditions like diabetes type 2, hypertension or many mental illnesses. They can be useful when other independent sources of risk information are also concordant [e.g., genotype/phenotype additional knowledge (Fahed et al., 2020), family history, imaging data]. A database of 4,688 published GWAS SNPs was generated encompassing 49 common diseases (we call these common diseases "phenotypes" from now onwards; **Supplementary Table 1**), their risk alleles and weighted contributions (odds ratio or beta scores). These phenotypes were selected according to GWAS Catalog criteria (https://www.ebi.ac.uk/gwas) as having studies including a primary GWAS analysis, defined as array-based genotyping and analysis of 100,000+ pre-QC SNPs selected to tag variation across the genome and without regard to gene content. Individual SNP-trait associations were collected with a statistical significance (SNP-trait p-value $<1.0 \text{ x } 10^{-5}$) in the overall (initial GWAS and replication) population. To create genetic risk scores, each collected SNP marker was required to possess (a) the risk allele and (b) the measurement or effect size that this risk allele confers to the individual that carries this mutation. A genetic risk score was calculated as the sum of the weights of all the phenotype's risk alleles observed in the individual divided by the total number of alleles reported for that phenotype. We used the final (Phase 3) dataset of the 1000 Genomes Project containing data for 2,504 individuals from 26 populations to calculate their genetic risk scores for each of the 49 phenotypes. The 1000 Genomes Project individuals became our background distribution of genetic risks against which to measure how far from the mean each of the family participant lies. We required that the identified GWAS SNPs are also present in the 1000 Genomes Project since individuals from the 1000 Genomes Project were used as a background population to which compare the participant's score. In order to control for potential differences in results due to the ethnic diversity of the background population, we also performed the analysis using a background population of only the 503 European (CEU) participants in the 1000 Genomes Project, given that all family members are of European origin.

We plotted the genetic risk score of each family member to establish whether he or she lies on the higher tail of the distribution of scores in relation to the calculated risk scores of 1000 Genomes Project individuals. In order to evaluate whether a member of the family had a reportable genetic risk, we applied a two standard deviations (2SD) threshold from the mean genetic risk score of the background 1000 Genomes individual distribution for a particular phenotype (equivalent to the top 5 percentile normal distribution of a predicted risk). We use a threshold of 2SD to give confidence that results are not attributable to chance. Furthermore, scores from both 1000 Genome individuals and family members are calculated independently. Multiple testing correction is not performed since the objective here is to identify family members in the extreme risk tail of the 1000 Genomes background distribution of calculated scores. For completeness, we also noted those phenotypes for which a greater than one standard deviation (1SD) from the mean background genetic risk is reached by the tested family individual.

Pharmacogenomics

We analyzed three well-known genes influencing pharmacogenomic responses, all of them forming part of the Cytochrome P450 family: *CYP2D6*, *CYP2C9*, *CYP2C19*. In order to extract relevant pharmacogenomic data, we rely on Food and Drug Administration (FDA) and European Medicines Agency (EMA) guidance sourced via the PharmGKB database (https://www.pharmgkb.org). We also take guidance from the Clinical Pharmacogenetics Implementation Consortium (CPIC; https://cpicpgx.org), the Association for Molecular Pathology and College of American Pathologists (https://www.amp.org), and the American College of Medical Genetics and Genomics (https://www.acmg.net). In order to perform the genotyping of pharmacological genes we allow the extraction of non-variant positions if required.

The testing of specific positions within a gene provides an accurate representation of the metaboliser status of an individual for that particular gene. For instance, if we were to assume that an individual has a nucleotide change 1846G>A in *CYP2D6*, this polymorphism is determinant for allele *4. If there are no other variants, it is presumed that the other allele this person harbors is a wild haplotype, being denoted as *4/*1 [see (Nofziger et al., 2020) for more detail]. Once both haplotypes for an individual are identified, a lookup table is referenced where the pharmacological effect of the observed haplotypes are indexed.

Pharmacological analysis also depends on whether analyzed genes have their copy number altered. We performed a consensus-based algorithm prediction using the short-read sequencing data for Father, Mother, Daughter and Son (**Supplementary Materials**). This analysis did not yield significant evidence for presence of copy number alterations in all Cytochrome P450 genes analyzed here.

Fitness and Nutrition

Besides pathogenicity screening, genetic risk scoring, and pharmacogenomics, there is further useful information that can be extracted from whole genomes using genotyping. In particular, we identify two areas of interest that provide further information about a person's genetic load: fitness and nutrition. We recognize that these areas of genetic analysis are less developed than pathogenicity screening, and so we add rigor to the analysis by first evaluating the quality of supporting evidence before testing for the presence of variants in the family. To evaluate the scientific validity and evidence for genotype-based dietary or fitness advice, we first performed a literature search to identify an initial list of potential genetic markers, and then adopted the proposed recommendations of Grimaldi et al. (2017) for specific gene x interactions and their relation to a health outcome. This framework allows us to establish levels of confidence for each of the SNPs or groups of SNPs we analyse for fitness and nutrition, according to a set of peer-reviewed guidelines. These guidelines differentiate four levels of scientific evidence assessment:

- "Convincing": gene x interaction is based on at least 3 studies with high subject numbers, showing the relation and mechanistic knowledge.
- "Probable" is based on several studies showing the relation and/or some mechanistic understanding.
- "Possible": based on a few studies showing the relation.
- "Not demonstrated" is any level of evidence below the established above.

The levels of assessment above rely on the following criteria:

- "Study quality rating": either A, B, C or D, based on whether a study is (a) interventional or observational; (b) prospective or retrospective, (c) whether it is randomized, placebo controlled and blinded; (d) the number of subjects with effect alleles (where possible); (e) the effect magnitude; (f) *P*-values, false discovery rate and multiple testing and; (g) replications in other populations and meta-analyses.
- "Type of gene x interaction": direct phenotype, intermediate phenotype, or indirect phenotype.
- "Nature of genetic variant": causal, in linkage disequilibrium with functional variant or associated but unknown function.
- "Biological plausibility," rated as high, medium, low, or unknown, based on our critical assessment of current understanding of the physiological effect of identified SNPs.

Our initial selection and classification of genotyping markers for fitness and nutrition are shown in **Tables 3**, **4**. We then assess each marker according to the above criteria of scientific evidence, carrying forwards for analysis in the family participants those classified as convincing (fitness $n = 2$; nutrition $n = 13$) and probable (fitness $n = 5$; nutrition $n = 1$).

Having made the selection of relevant SNPs according to the above framework, we proceed to analyse the family. The trait analysis is performed as follows. First, a list of all the positions of the SNPs to be tested is created. All those positions are queried in the VCF files for each of the family members and all observed alleles are then recorded. The observed alleles are then interpreted via lookup tables collected from the scientific literature.

An exception to the above approach concerns the phenotype susceptibility to VO_2max trainability, where we use a specific study. To calculate the VO_2Max trainability genetic score, we follow the methodology outlined in Bouchard et al. (2011), that identifies SNPs associated with improvements in VO_2Max. This study provides a panel of 21 SNPs that accounted for 49% of the variance in VO_2Max trainability.

RESULTS

Pathogenicity Screening

Figure 2 shows a summary of the pedigree and filtered variants found listed within each individual. For Son, when searching for

TABLE 3 | Summary of fitness trait analysis candidates assessed according to the scientific validity score as proposed by Grimaldi et al. (2017).

Category	Trait	RSID	Gene	Study quality rating	Type of gene x trait interaction	Nature of genetic variant	Biological plausibility	Number of independent studies	Total number subjects studied	Knowledge of biological mechanism involved	Scientific validity score	References
Fitness	VO_2max	21 SNPs	Multiple	A	Direct phenotype	Causal	High	35	>1000	Medium	Convincing	Rice et al., 2012; Ghosh et al., 2013; Williams et al., 2017
Fitness	Muscle performance	rs1815739	ACTN3	A	Indirect phenotype	Causal	High	24	>1000	High	Convincing	Kikuchi et al., 2014, 2015, 2017b; Schadock et al., 2015; Yvert et al., 2015, 2016; Baumert et al., 2016; Itaka et al., 2016; Min et al., 2016; Del Coso et al., 2017, 2019a,b; Galeandro et al., 2017; Houweling et al., 2018; Zhang et al., 2019; Baltazar-Martins et al., 2020; Calvano Küchler et al., 2020; Murtagh et al., 2020; Płóciennik et al., 2020
Fitness	Caffeine sensitivity/Increased exercise performance with caffeine	rs762551	CYP1A2	C	Direct phenotype	Causal	High	7	250	High	Probable	Pataky et al., 2016; Salinero et al., 2017; Guest et al., 2018; Puente et al., 2018; Carswell et al., 2020; Grgic et al., 2020; Muñoz et al., 2020
Fitness	Endurance	rs4253778	PPARA	B	Indirect phenotype	Causal	Medium	6	3267	High	Probable	Ahmetov et al., 2009; Ahmetov and Fedotovskaya, 2015; Lopez-Leon et al., 2016; Petr et al., 2019; Johansen et al., 2020; Murtagh et al., 2020
Fitness	Lactate blood levels	rs1049434	MCT1	B	Direct phenotype	Causal	High	4	2048	High	Probable	Cupeiro et al., 2012; Fedotovskaya et al., 2014; Ben-Zaken et al., 2015; Kikuchi et al., 2017a
Fitness	Osmotic balance by water support	rs1049305	AQP1	B	Indirect phenotype	Causal	High	3	2613	Medium	Probable	Saunders et al., 2015; Rivera and Fahey, 2019; Rivera et al., 2020
Fitness	Performance	rs12594956	NRF-2	C	Indirect phenotype	Causal	High	4	1598	Medium	Probable	He et al., 2007; Eynon et al., 2010, 2013; Peplonska et al., 2017
Fitness	Glucose transportation and lipid and glucose oxidation	rs8192678	PPARGC1A	C	Indirect phenotype	Causal	High	5	409	Medium	Possible	Petr et al., 2018
Fitness	Endurance	rs12722	COL5A1	C	Indirect phenotype	Causal	High	3	952	Medium	Possible	O'Connell et al., 2013; Bertuzzi et al., 2014; Murtagh et al., 2020
Fitness	Elite endurance	rs4994	ADRB3	D	Indirect phenotype	Causal	High	2	453	Low	Not demonstrated	Gómez-Gallego et al., 2010; Santiago et al., 2011

A total of 10 fitness traits were identified for their gene x interaction assessment. We classified as "Convincing" those traits whose gene x interaction is based on at least 3 studies with high subject numbers, showing the relation and mechanistic knowledge. A trait classified as "Probable" is based on several studies showing the relation and/or some mechanistic understanding. A trait is deemed "Possible" if based on a few studies showing the relation. "Not demonstrated" are those traits for which any level of evidence is below the established criteria above.

TABLE 4 | Summary of nutrition trait analysis candidates assessed according to the scientific validity score as proposed by Grimaldi et al. (2017).

Category	Trait	RSID	Gene	Study quality rating	Type of gene x trait interaction	Nature of genetic variant	Biological plausibility	Number of independent studies	Total number subjects studied	Knowledge of biological mechanism involved	Scientific validity score	References
Nutrition	Homocystine levels	rs1801133	MTHFR	A	Direct phenotype	Causal	High	70	>100000	High	Convincing	Boccia et al., 2008, 2009; Clarke et al., 2011; Liew and Gupta, 2015
Mental health / Nutrition	Alzheimer's	rs429358, rs7412	APOE	A	Direct phenotype	Causal	High	146	>100000	High	Convincing	Martins et al., 2006; Zhang et al., 2015; Rasmussen et al., 2018
Nutrition	Alcohol dependence	rs1229984	ADH1B	A	Direct phenotype	Causal	High	59	>100000	High	Convincing	Jorgenson et al., 2017; Katsarou et al., 2017; Masaoka et al., 2017; Wolf et al., 2017; Hubacek et al., 2018; Justice et al., 2018; Polimanti and Gelernter, 2018; Walters et al., 2018; Yokoyama et al., 2018, 2019, 2020a,b,c; Howe et al., 2019; Johnson et al., 2019; Lai et al., 2019; Sun et al., 2019; Szentkereszty-Kovács et al., 2019; Thompson et al., 2020
Nutrition	Greater total body adiposity	rs9939609	FTO	A	Direct phenotype	Causal	High	25	>100000	Medium	Convincing	Bollepalli et al., 2010; Dedoussis et al., 2011; Mangge et al., 2011; Dwivedi et al., 2012; Lauria et al., 2012; Meng et al., 2014; Zhang et al., 2014; Zhao et al., 2014a; Qi et al., 2015a; Quan et al., 2015; Duicu et al., 2016; García-Solís et al., 2016; Livingstone et al., 2016; Bordoni et al., 2017; Almeida et al., 2018; Ferreira Todendi et al., 2019; Ranzenhofer et al., 2019; Todendi et al., 2020
Nutrition	Vitamin D Metabolism	rs4588	GC	B	Direct phenotype	Causal	High	21	>100000	High	Convincing	Robien et al., 2013; Nissen et al., 2014, 2015; Pekkinen et al., 2014; Braithwaite et al., 2015; Madden et al., 2015; Touvier et al., 2015; Petersen et al., 2017; Yao et al., 2017; Chuaychoo et al., 2018; Karuwanarint et al., 2018; Al-Daghri et al., 2019; Bahrami et al., 2019; Enlund-Cerullo et al., 2019; Mehramiz et al., 2019; Rahimi et al., 2019; Zhou et al., 2019; Gibbs et al., 2020a,b

(Continued)

TABLE 4 | Continued

Category	Trait	RSID	Gene	Study quality rating	Type of gene x trait interaction	Nature of genetic variant	Biological plausibility	Number of independent studies	Total number subjects studied	Knowledge of biological mechanism involved	Scientific validity score	References
Nutrition	Vitamin B12 level	rs602662	FUT2	A	Direct phenotype	Causal	High	6	>9000	High	Convincing	Hazra et al., 2009; Tanaka et al., 2009; Tanwar et al., 2013; Allin et al., 2017; Nongmaithem et al., 2017; Zhao and Schooling, 2017
Nutrition	Vitamin C level	rs33972313	SLC23A1	A	Direct phenotype	Causal	High	12	>100000	High	Convincing	Timpson et al., 2010; Duell et al., 2013; Amir Shaghaghi et al., 2014; Kobylecki et al., 2015, 2018; Wade et al., 2015; Ravindran et al., 2019
Nutrition	Vitamin E level	rs964184	BUD13/ZNF259	B	Direct phenotype	Causal	High	4	>10000	High	Convincing	Major et al., 2011, 2012, 2014; Wang and Xu, 2019
Nutrition	Iron Overload /Hemochromatosis	rs1800562	HFE	B	Direct phenotype	Causal	High	4	>5000	High	Convincing	McLaren et al., 2011; Katsarou et al., 2016; Barton et al., 2018; Wilman et al., 2019
Nutrition	Saturated fat / risk of T2D	rs1137101	LEPR	C	Indirect phenotype	Causal	High	12	>10000	Medium	Convincing	Domínguez-Reyes et al., 2015; Yang et al., 2016
Nutrition	Polyunsaturated Fatty Acids	rs174547	FADS1	C	Direct phenotype	Causal	High	11	3713	Medium	Convincing	Huang et al., 2017; Ching et al., 2019; Wang et al., 2020
Nutrition	Lactose persistence	rs4988235	MCM6-LCT	A	Direct phenotype	Causal	High	>10	>100000	High	Convincing	Baffour-Awuah et al., 2015
Nutrition	Celiac disease	rs2187668	HLA-DQA1	A	Direct phenotype	Causal	High	Many	7249	High	Convincing	van Heel et al., 2007; Hunt et al., 2008
Nutrition	Saturated fat	rs5082	APOA2	B	Direct phenotype	Causal	High	3	2856	Medium	Probable	Yabuta et al., 2016; Moran et al., 2019; Amengual et al., 2020; Graßmann et al., 2020
Nutrition	Vitamin A level	rs6564851	BCO1	C	Direct phenotype	Causal	High	4	328	Medium	Possible	Delgado-Lista et al., 2007; Smith et al., 2013; Noorshahi et al., 2016
Nutrition	Total Carbohydrates	rs7578326	IRS1	B	Indirect phenotype	Causal	High	2	~2000	Medium	Possible	Zheng et al., 2013; Mahmutovic et al., 2019
Nutrition	Total Carbohydrates	rs2943641	IRS1	B	Indirect phenotype	Causal	High	2	~2000	Medium	Possible	Zheng et al., 2013; Mahmutovic et al., 2019
Nutrition	Sugar	rs7903146	TCF7L2	A	Indirect phenotype	Causal	High	2	26905	Medium	Possible	Hindy et al., 2012, 2016
Nutrition	Alcohol metabolism	rs698	ADH1C	C	Direct phenotype	Causal	High	Many	>100000	High	Possible	Bierut et al., 2010; Martínez et al., 2010; Olfson and Bierut, 2012; Kranzler et al., 2019
Nutrition	Sweet Foods / Sweet Tooth	rs838133	FGF21	A	Direct phenotype	Causal	Medium	1	6514	High	Not demonstrated	Søberg et al., 2017

(Continued)

TABLE 4 | Continued

Category	Trait	RSID	Gene	Study quality rating	Type of gene x trait interaction	Nature of genetic variant	Biological plausibility	Number of independent studies	Total number subjects studied	Knowledge of biological mechanism involved	Scientific validity score	References
Nutrition	Vitamin B6 level	rs4654748	ALPL	B	Direct phenotype	Causal	High	1	∼3000	High	Not demonstrated	Tanaka et al., 2009
Nutrition	Total Carbohydrates	rs2241201	MMAB	C	Indirect phenotype	Causal	High	1	920	Low	Not demonstrated	Junyent et al., 2009
Nutrition	Fiber	rs4457053	ZBED3	B	Indirect phenotype	Causal	High	1	26905	Medium	Not demonstrated	Hindy et al., 2016
Nutrition	Fiber	rs10923931	NOTCH2	B	Indirect phenotype	Causal	High	1	26905	Medium	Not demonstrated	Hindy et al., 2016
Nutrition	Sugar	rs12255372	TCF7L2	B	Indirect phenotype	Causal	High	2	26979	Medium	Not demonstrated	Hindy et al., 2016; López-Ortiz et al., 2016
Nutrition	Total fat	rs324420	FAAH	C	Direct phenotype	Causal	High	5	>5000	Medium	Not demonstrated	Jensen et al., 2007; de Luis et al., 2011; Knoll et al., 2012; Balsevich et al., 2018; Doris et al., 2019
Nutrition	Saturated fat	rs12449157	GFOD2	D	Direct phenotype	Causal	High	1	41	Medium	Not demonstrated	Guevara-Cruz et al., 2013
Nutrition	Omega-3 Fatty Acids	rs17300539	ADIPOQ	C	Direct phenotype	Causal	High	1	310	Medium	Not demonstrated	Alsaleh et al., 2013
Nutrition	Saturated Fatty Acids	rs1800629	TNF	C	Indirect phenotype	Causal	High	2	472	Medium	Not demonstrated	Cormier et al., 2016; Oki et al., 2017
Nutrition	Protein	rs12785878	DHCR7	D	Indirect phenotype	Causal	High	1	732	Medium	Not demonstrated	Qi et al., 2015b
Nutrition	Calcium	rs2228570	VDR	C	Direct phenotype	Causal	High	3	>5000	Medium	Not demonstrated	Jenab et al., 2009; Slattery et al., 2010; Zhou et al., 2015
Nutrition	Zinc	rs73924411	SLC30A3	D	Direct phenotype	Causal	High	2	350	Low	Not demonstrated	da Rocha et al., 2014a,b

A total of 32 nutrition traits were identified for their gene x interaction assessment. We classified as "Convincing" those traits whose gene x interaction is based on at least 3 studies with high subject numbers, showing the relation and mechanistic knowledge. A trait classified as "Probable" is based on several studies showing the relation and/or some mechanistic understanding. A trait is deemed "Possible" if based on a few studies showing the relation. "Not demonstrated" are those traits for which any level of evidence is below the established criteria above.

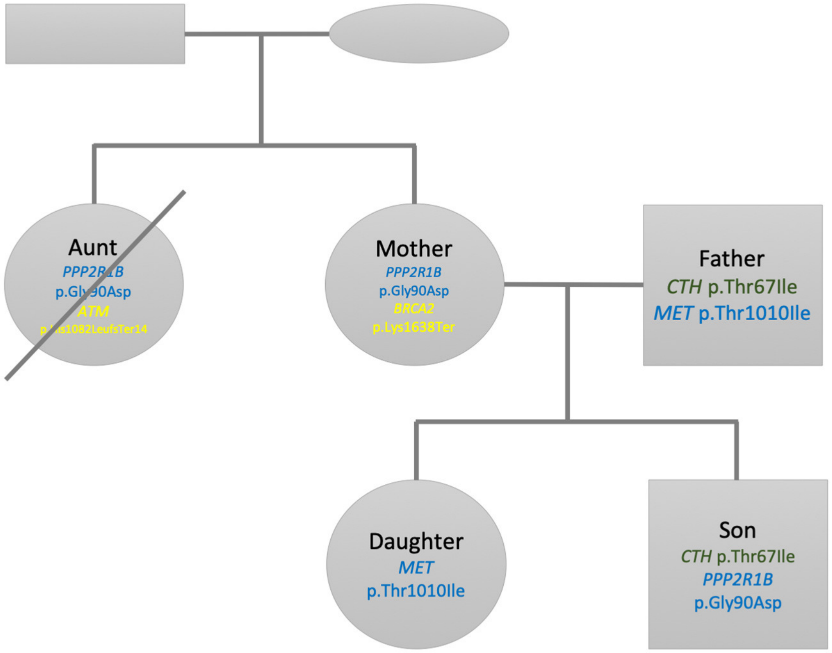


FIGURE 2 | Family pedigree showing the relationship, gender (square: male, circle: female), and variants found listed within each individual. Green variants are inferred as benign, blue are variants of unknown significance and yellow pathogenic variants according to ACMG scoring. The crossed circle indicates a deceased individual.

pathogenic or likely pathogenic mutations within a panel of 4,100 OMIM morbid genes, we found that only two mutations passed our prioritization and filtering criteria (see Methods section). No other mutations passed the threshold criteria within the ACMG59 and Hereditary Cancer panels. The first mutation is a heterozygous missense change of C → T; c.200C>T; p.Thr67Ile within the CTH gene. This change has been associated to cystathioninuria, a disorder observed in 1 out of 20,000 individuals (ORPHANET, Pavan et al., 2017). However, the ExAC (Lek et al., 2016) frequency in non-Finnish European is (∼1 in 100), so much higher than the prevalence of the disorder. We also observe that this missense variant is not excessively constrained: its missense z-score is −0.127428 (excessively constrained genes are those with a missense z-score $>$ 3.09, corresponding to a p-value $<$ 0.001). Multiple lines of computational evidence suggest no impact on the gene. Our current assessment is that this variant is benign according to the ACMG scoring and inferred classification and is therefore not considered any further. The second mutation for Son corresponds to chr11:111764842 (rs1805076) producing C → T; c.269G>A; p.Gly90Asp in PPP2R1B. This gene encodes a regulatory subunit of protein phosphatase 2. Protein phosphatase 2 is one of the four major Ser/Thr phosphatases, and it is implicated in the negative control of cell growth and division. While ClinVar evidence suggests a matching allele to cause lung cancer, the computational and other sources of evidence are inconclusive, hence we infer this variant to be of uncertain significance (VUS).

For Father two variants pass the filters. The first variant is selected from the 4100 genes OMIM panel and corresponds to the heterozygous missense change of C → T; c.200C>T; p.Thr67Ile within the *CTH* gene, which is the same one Son has. We set the same classification as above and conclude it to be a benign variant as well. The second variant is located in the *MET* gene, part or our Hereditary Cancer panel, on chr7:116771936 (rs56391007) and produces C → T; c.3029C>T; p.Thr1010Ile. This gene encodes a member of the receptor tyrosine kinase family of proteins and the product of the proto-oncogene *MET*. The *MET* gene is associated with autosomal dominant hereditary papillary renal cell carcinoma (Takahashi et al., 2002). According to ORPHANET (Pavan et al., 2017), the prevalence of this cancer is less than 1 in 1,500,000, while the allele frequency of this variant is 1 in ∼89 in non-Finnish European, higher than expected for the disorder. There are seven pathogenic and 15 likely pathogenic ClinVar missense variants in this gene while there are two benign and 25 likely benign ClinVar missense variants in this gene. Evidence thus indicates that missense variants are not a common mechanism of disease. In addition, there are multiple sources that point to this mutation being both pathogenic and likely benign. We conclude that this variant is of uncertain significance.

The first selected mutation for Mother comes from the OMIM disease gene panel and is the same as the one previously reported for Son corresponding to chr11:111764842 (rs1805076) producing C → T; c.269G>A; p.Gly90Asp in *PPP2R1B*. We apply the same criteria as above, inferring this variant effect being of uncertain significance. The second variant selected for interpretation in Mother corresponds to a heterozygous stop gained mutation in chr13:32339267 (rs886040553) producing the change A → T; c.4912A>T; p.Lys1638Ter in the *BRCA2* gene. The gene is included both in the ACMG 59 and our Hereditary Cancer gene panel. The impact of this mutation is stop-gained in a splice site. ClinVar evidence contains several entries in its classification history all matching the allele and all pathogenic. This variant is absent from gnomAD or the 1000 genomes project. We thus classify this entry as pathogenic according to the ACMG scoring and inferred classification and as associated to hereditary breast and ovarian cancer syndrome. However, since the gene is autosomal recessive, both alleles need to be affected in order to cause the disorder.

For Daughter we did not find any mutation in the panels OMIM disease genes and ACMG 59 after our filtering criteria. One mutation passes our criteria for the Hereditary Cancer panel, corresponding to the variant located in chr7:116771936 (rs56391007) producing C → T; c.3029C>T; p.Thr1010Ile, located in the *MET* gene, identified for Father as well. As above, we conclude that this variant is of uncertain significance.

We prioritize the filtered variants from Aunt's variant file. After performing the selection of variants above the VCF threshold quality of 20, we identify three variants. Among these three variants, we decided to discard a homozygous variant corresponding to chr1:25563142 (rs121908326) in the *LDLRAP1* gene, as this one did not qualify to have a sufficient genotype quality (minimum acceptable threshold = 40). Among the remaining two variants, the first one corresponds to chr11:111764842 (rs1805076), producing C → T; c.269G>A; p.Gly90Asp in *PPP2R1B*, already observed in Son and Mother. We apply the same criteria as above, inferring this variant effect being of uncertain significance. The remaining heterozygous variant, chr11:108272812 (rs587776549) in the *ATM* gene, produces a frameshift mutation CATC → CTGATc.3245_3247delinsTGAT p.His1082LeufsTer14. The ATM protein plays a critical role assisting cells in recognizing damaged or broken DNA strands, enabling them to repair broken strands and work on maintenance of the stability of the cell's genetic information. The mutation identified here has been described as pathogenic, involved in ataxia-telangiectasia syndrome and has also been linked to a hereditary cancer predisposition (Laake et al., 2000). Mutations to the *ATM* gene have a 20% to 30% lifetime risk of lymphoid, gastric, breast, central nervous system and skin, including melanoma (Choi et al., 2016). We conclude this variant as being pathogenic.

Given the limitations of the VCF file produced for Aunt, for the reminder of the results section, we are only able to perform further analyses for Father, Mother, Daughter and Son; analyses which include genetic risk scores, pharmacogenomics and nutrition/fitness traits.

Genetic Risk Scores

From our initial list of 49 GWAS phenotypes (**Supplementary Table 1**), we identified members of the family who have a risk score (or predisposition) of one or two standard deviations (SD) from the average risk score of the 1000 genomes population for the same condition. **Table 5** shows the phenotypes whose genetic score from the initial list is more than 2SD (yellow) and those with more than 1SD (green). The GWAS studies from which the genetic risk SNPs originated are sourced in the "Reference Studies" column.

First, we find that no one phenotype in yellow (>2SD) occurs in isolation. There is either another member of the family in yellow or green (>1SD). This occurs for the predicted higher risks of ulcerative colitis and nicotine withdrawal symptoms. Second, we observe that related phenotypes with high risk overlap (although not always) in the same family individual. For instance, ulcerative colitis is a subtype of inflammatory bowel disease (Ronald et al., 2006; Wu et al., 2009a; Uhlig and Muise, 2017). We find that Mother and Daughter have risks overlapping both phenotypes whereas Son only ulcerative colitis. Father has a risk for three of the four phenotypes for the mental health category. For the disorders where we observe this overlapping phenotype risk, there is scientific literature (Ronald et al., 2006; Wu et al., 2009a; Uhlig and Muise, 2017) supporting this pattern. Third, there are high risk phenotypes in a parent also observed in their offspring. For instance, ulcerative colitis is not present in Father but it is predicted high risk in Mother together with both children displaying inherited risk.

If we describe results according to categories, for general health, we thus find that ulcerative colitis constitutes a phenotype where the risk of the condition is shared among several family members (Mother, Daughter, Son; **Supplementary Figures 1A,B**). The high risk for ulcerative colitis also overlaps with the higher than average (>1SD) risk of inflammatory bowel disease in two same family members (Mother, Daughter). For cancer there is a more elevated than normal (>1SD) predicted risk of breast cancer among Mother and Daughter, with some isolated moderately high (>1SD) predicted risks of bladder carcinoma for Daughter, glaucoma for Son and prostate cancer for Father. For mental health, Father has higher than normal (>1SD) predicted risk on bipolar disorder, depression, and posttraumatic disorder. Mental health diseases share similar markers thus influencing the greater number of potentially deleterious yet related phenotypes observed in Father. For dependence and withdrawal symptoms phenotypes we find that the paternal line has a higher than average alcohol (>1SD) dependence predicted risk whereas the maternal line passes on to Son a high predicted risk (>2SD) of nicotine withdrawal symptoms. Phenotypes in green (>1SD) that are not shared with other family members are not considered any further. Graphical representation of the ulcerative colitis results using both the whole 1000 Genomes background population (2,504 individuals) and only the Europeans (503) may be found in **Supplementary Figure 1**.

Pharmacogenomics

We analyzed the metaboliser status of three cytochrome P450 genes (*CYP2C9*, *CYP2C19*, *CYP2D6)* affecting pharmacological responses in Father, Mother, Daughter, and Son. We also look at some pharmacology-related SNPs in additional genes.

CYP2C9 is responsible for the metabolic clearance of up to 15–20% of all drugs undergoing Phase 1 metabolisation, including warfarin, phenytoin, and oral hypoglycaemics (source: Get to

TABLE 5 | Phenotypes for family members with <1SD genetic risk score.

Category	Phenotype	Father	Mother	Daughter	Son	Reference Studies
General health	Inflammatory bowel disease		🟩	🟩		Liu et al., 2015
	Ulcerative colitis		🟨	🟩	🟨	Berndt et al., 2013
	Obesity	🟩			🟩	Berndt et al., 2013
Lipids	Triglycerides			🟩		Willer et al., 2013
Cancer	Bladder carcinoma			🟩		Kiemeney et al., 2008, 2010; Wu et al., 2009b; Rothman et al., 2010; Rafnar et al., 2011, 2014; Figueroa et al., 2014; Matsuda et al., 2015; Wang et al., 2016
	Breast cancer		🟩	🟩		Howard et al., 2018
	Glaucoma				🟩	Choquet et al., 2018
	Prostate cancer	🟩				Schumacher et al., 2018
Mental health	Bipolar disorder	🟩				Ferreira et al., 2008; Smith et al., 2009; Cichon et al., 2011; Psychiatric and Consortium Bipolar Disorder Working Group, 2011; Mühleisen et al., 2014; Hou et al., 2016; Ikeda et al., 2018
	Depression	🟩				Howard et al., 2018
	Posttraumatic stress disorder	🟩				Nievergelt et al., 2015; Stein et al., 2016
	Schizophrenia				🟩	Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014
Dependence /withdrawal	Alcohol dependence	🟩		🟩	🟩	Gelernter et al., 2014; Mbarek et al., 2015
	Nicotine withdrawal		🟨		🟨	Hällfors et al., 2019
Fitness	Heart rate recovery	🟩				Ramírez et al., 2018
Nutrition	Caffeine metabolism				🟩	Cornelis et al., 2016
Appearance	Male pattern baldness	🟩				Pirastu et al., 2017

We then calculate their average score and standard deviation (SD). For each family participant we calculate their genetic risk score in the same way as individuals from the 1000 Genomes Project and mark as yellow if his/her risk score is >2SD of the 1000 Genomes Project average and green if his/her risk score is >1SD of the average in 1000 Genomes Project individuals. Both Son and Mother have a >2SD risk score for nicotine withdrawal symptoms. For ulcerative colitis, Mother and Son have a >2SD score (yellow) and Daughter a >1SD score (green). The increased risk of inflammatory bowel disease for both Mother and Daughter overlaps with their predicted ulcerative colitis susceptibility. Father has a >2SD risk of autism spectrum disorder. Autism spectrum disorder genetic risk appears also for the two children (>1SD; Daughter, Son). The alcohol dependence genetic risk score reflects another paternal line predicted predisposition inherited by both children. Since no one in the family to date has been predicted to suffer from autism spectrum disorder or alcohol dependence, it is not possible to confirm this result. A slightly increased risk of obesity in Father and Son is also predicted by this multigenic risk score calculation.

Know an Enzyme: CYP2C9[1]). Some of the more potent *CYP2C9* inhibitors include amiodarone, fluorouracil, metronidazole, and sulphaphenazole. Dangerous drug-drug interaction can arise when an inhibitor is added to a therapeutic regime that includes drugs with a low therapeutic index, such as s-warfarin. Inducers, such as rifampicin, can substantially increase *CYP2C9* activity (source: Get to Know an Enzyme: CYP2C9). For *CYP2C9*, Father, Mother and Son have a predicted metaboliser status of intermediate (*1/*2). For Daughter, the predicted metaboliser status for *CYP2C9* is poor (*2/*2).

Warfarin is an anticoagulant used in the prevention and treatment of venous thrombosis, pulmonary embolism, and the complications associated with atrial fibrillation and/or cardiac valve replacement (Dean, 2018). Warfarin metabolism is influenced by genetic polymorphisms in *CYP2C9* and *VKORC1* (Biss et al., 2012). Carriers of the common allelic variants (*2 or *3) of the *CYP2C9* are associated with a lower warfarin dose requirement accompanied by a greater tendency to experience haemorrhagic complications. In addition, adults with *VKORC1* (rs9923231) CC alleles require higher warfarin doses than TC or TT. Based on these alleles, we found that Son and Father have a *1/*2 *CYP2C9* variant and a TT for rs9923231. Mother has a *1/*2 *CYP2C9* variant and a CT for rs9923231. Daughter has a *2/*2 *CYP2C9* variant and a TT for rs9923231. This makes Son, Father and Mother intermediate metabolisers and Daughter a poor metaboliser of warfarin.

CYP2C19 is a liver enzyme that acts on at least 10% of drugs in current clinical use (source: Genetics Home Reference[2]; see references), most notably the antiplatelet treatment clopidogrel (Plavix) but also drugs that treat pain associated with ulcers, such as omeprazole, antiseizure drugs such as mephenytoin, the antimalarial proguanil, and the anxiolytic diazepam. For this gene we found Son, Mother and Daughter to be predicted normal metabolisers (*1/*1) whereas Father is predicted an intermediate metaboliser (*17/*4A).

For *CYP2D6*, the final cytochrome we analyse here, we are able to estimate the metabolism and elimination of approximately 25% of clinically used drugs including the opiate codeine (Wang et al., 2009). *CYP2D6* is highly polymorphic in the

[1] Get to Know an Enzyme: CYP2C9 Pharmacy Times. Available at: https://www.pharmacytimes.com/publications/issue/2008/2008-03/2008-03-8462 [accessed October 15, 2019].

[2] Genetics Home Reference CYP2C19 gene. Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/CYP2C19 [accessed October 15, 2019].

human population, with marked inter-racial variation observed. Individuals are identified as ultra-rapid (UM), extensive (EM), intermediate (IM) or poor metaboliser (PM), according to the number of functional alleles.

For members of this family we find that there is considerable variation in the alleles detected. For Son and Father, we find them to be predicted extensive metabolisers (*2/*41 and *1/*2, respectively). Mother has the following star alleles *10/*2/*41/*4 [activity score: 0.5–1 (Gaedigk et al., 2018)] which make her predicted range between an intermediate and extensive metaboliser. Daughter has *10/*4/*20, which makes her a poor or intermediate predicted metaboliser (activity score 0–0.5).

rs12979860 is a SNP near the *IL28B* gene, encoding interferon-lambda-3 (IFN-lambda-3). This SNP influences hepatitis C treatment-induced viral clearance. It is associated with an approximately twofold change in response to pegylated interferon-alpha (PEG-IFN-alpha) plus ribavirin (RBV) treatment, both among patients of European ancestry ($p = 1.06$ x 10e-25). Research indicates that the virus was eradicated in ∼80% of CC patients, compared to only about 25% of those with TT, while CT response was intermediate (Elkader and Sproule, 2005). We found that Son and Father carry a CC genotype, whereas Mother and Daughter carry a CT genotype.

Fitness Trait Analysis

Filtering

First, we performed a filtering of the SNPs associated with fitness traits in order to determine which of them should be applied to our family cohort. The full results of our filtering can be found in **Table 3**. From a total of 10 markers initially selected for genotyping, we classified two as “Convincing” (VO$_2$max, and rs1815739 for *ACTN3*), five as “Probable,” two as “Possible” and one “Not demonstrated.” As an example of the application of this framework, in **Table 3**, the best studied SNP marker is rs1815739 for *ACTN3*. We identified 24 studies for rs1815739 in fitness, most of which suggested a significant decrease in muscle performance by the effect allele (also known as X allele). Based on these studies, we classify the biological plausibility of this marker as high. Our scientific evidence assessment for rs1815739 is “Convincing.”. For “increased performance with caffeine,” we assess existing scientific evidence as “Probable” because despite finding 7 studies, the total number of participants summed by all seven studies is only 250. Within those, there is also one study not showing significant differences in performance with coffee intervention (Pataky et al., 2016; Salinero et al., 2017; Guest et al., 2018; Puente et al., 2018; Carswell et al., 2020; Grgic et al., 2020; Muñoz et al., 2020).

For family trait analysis, we only apply those markers that are either classified as “Convincing” or “Probable”. In the next section we describe in detail our selected fitness analysis results.

Fitness Trait Analysis Performed on the Family Cohort

Table 6 summarizes the fitness traits analyzed for 4 family members. Concerning VO$_2$Max trainability, training response markers within the 21 SNP panel show Son scoring 13/21 favorable alleles, Father and Mother 16/21 favorable alleles and Daughter scores 15/21 favorable alleles. This contrasts with ≥19 of these alleles associated with elite athletes (Bouchard et al., 2011; Rice et al., 2012; Ghosh et al., 2013; Williams et al., 2017).

The *ACTN3* R577X (rs1815739) C>T base substitution results in the transformation of an arginine amino acid (R) to a premature stop codon (X). X allele homozygotes are deficient in the alpha-actinin-3 protein, which is associated with a lower fast-twitch fiber percentage and potentially increased injury risk (Yang et al., 2003; Massidda et al., 2019). We found that Father, Mother and Daughter have a CT genotype (XR); whereas Son, harbors a homozygote X allele genotype (XX).

A polymorphism in the *CYP1A2* gene (rs762551; AA genotype) has been associated with improved exercise performance when combined with caffeine intake, with no effect for those with the AC genotype and diminished performance in those with the CC genotype (Guest et al., 2018). We found that most family members (Son, Father, and Daughter) had an AA genotype for this SNP, whereas Mother had a CA genotype.

The role of the peroxisome proliferator activated receptor alpha (*PPARA*) gene intron 7 G/C polymorphism (rs4253778) is also tested in the family. Athletes with high ability in endurance sports have a higher frequency of the G allele (Lopez-Leon et al., 2016). We found that Son and Mother did not have any of the G allele, whereas Father and Daughter had a G allele each.

For the *MCT1* gene’s rs1049434, we find all family members to have the TT genotype, associated with lower lactate levels (Cupeiro et al., 2012; Fedotovskaya et al., 2014; Ben-Zaken et al., 2015; Kikuchi et al., 2017a). For the *AQP1* gene, which is associated with osmotic balance and fluid loss when exercising, possession of the C allele has been associated with faster cardiorespiratory endurance (Rivera and Fahey, 2019). For this gene, we found C (favorable) alleles in Son, Father, and Daughter, while no C alleles were found in Mother. Finally, for rs12594956 in *NRF-2*, we find that the genotypes observed (CA/CC) in all family members are not associated with the effect allele (He et al., 2007; Eynon et al., 2010, 2013; Peplonska et al., 2017).

Nutrition Trait Analysis

Our analysis includes markers involved in the metabolism of main components of diet: carbohydrates, fats, and proteins. We also look at metabolization of essential nutritional components such as vitamins, minerals, and specific dietary substances like lactose, whose metabolism is strongly linked to a genetic marker according to our suggested framework. **Table 7** provides a summary of the nutrition markers explained in this section.

Filtering

We performed a filtering of the SNPs associated with nutrition traits to select SNPS to be applied to our family cohort. The full results of our filtering can be found in **Table 5**. From a total of 32 markers initially selected for analysis, we classified 13 as “Convincing,” one as “Probable,” five as “Possible”, and 12 “Not demonstrated.” An example of convincing scientific evidence for nutrition interventions in **Table 4** includes *MTHFR* (rs1801133). This SNP is said to affect homocysteine concentrations, which are influenced by dietary folate (Boccia et al., 2008, 2009; Clarke et al., 2011; Liew and Gupta, 2015). A large number of studies ($n = 70$)

TABLE 6 | Summary of fitness trait analysis for 4 family members.

Trait	RSID	Gene	Scientific validity score	Father	Mother	Daughter	Son
VO_2Max	21	Multiple	Convincing	16/21	16/21	15/21	13/21
Muscle performance	rs1815739	*ACTN3*	Convincing	XR	XR	XR	XX
Caffeine sensitivity/Increased exercise performance with caffeine	rs762551	*CYP1A2*	Probable	AA	CA	AA	AA
Endurance	rs4253778	*PPARA*	Probable	GC	CC	GC	CC
Lactate blood levels	rs1049434	*MCT1*	Probable	TT	TT	TT	TT
Osmotic balance by water support	rs1049305	*AQP1*	Probable	GC	GG	GC	GC
Performance	rs12594956	NRF-2	Probable	CA	CC	CC	CA

The table shows the observed genotype or (for VO_2max) favorable alleles for those traits whose scientific evidence assessment was judged as "Convincing" or "Probable".

TABLE 7 | Summary of nutrition trait analysis for the 4 family members.

Trait	RSID	Gene	Scientific validity score	Father	Mother	Daughter	Son
Homocystine levels	rs1801133	*MTHFR*	Convincing	GA	GA	AA	GG
Vitamin B12 level	rs602662	*FUT2*	Convincing	GA	GG	GA	GG
Vitamin C level	rs33972313	*SLC23A1*	Convincing	CC	CC	CC	CC
Vitamin D Metabolism	rs4588	*GC*	Convincing	GT	GT	GT	GG
Vitamin E level	rs964184	*BUD13 / ZNF259*	Convincing	CC	GG	GC	GC
Greater total body adiposity	rs9939609	*FTO*	Convincing	AA	TT	TA	TA
Saturated fat	rs5082	*APOA2*	Probable	AA	GA	AA	AA
Polyunsaturated Fatty Acids	rs174547	*FADS1*	Convincing	TT	TT	TT	TT
Saturated fat/risk of T2D	rs1137101	*LEPR*	Convincing	AG	AG	GG	AG
Iron Overload /Hemochromatosis	rs1800562	*HFE*	Convincing	GG	GG	GG	GG
Celiac disease	rs2187668	*HLA-DQA1*	Convincing	CC	CC	CC	CC
Lactose persistence	rs4988235	*MCM6-LCT*	Convincing	GG	AA	GA	GA
Alzheimer's	rs429358, rs7412	*APOE*	Convincing	ε3/ε3	ε3/ε3	ε3/ε3	ε3/ε3
Alcohol dependence	rs1229984	*ADH1B*	Convincing	CC	TC	CC	CC

The table follows the scientific validity score presented in Methods and the observed genotype for the specific trait. We only analyse those traits for which there is a convincing or probable genotype x diet intervention scientific evidence. From among these, here we only show those with a predicted effect, except in the case of Alzheimer's and Coronary Artery Disease.

have been performed to date about this interaction, including randomized trials. We evaluate this interaction as having a high biological plausibility. An example of nutrition marker we classify as possible is *BCO1* (rs6564851). According to our research (**Table 4**), there are 4 studies with a number of total subjects analyzed of 328 (Yabuta et al., 2016; Moran et al., 2019; Amengual et al., 2020; Graßmann et al., 2020). Our judgement of the underlying knowledge of the biological mechanism involved is medium and there are some cases where a potential intervention may not have the desired effect. We do not include this marker in our subsequent analyses.

Same as in the fitness category, for family trait analysis we only apply those markers that are either classified as "Convincing" or "Probable." In the next section we describe in detail our selected nutrition trait analysis results.

Nutrition Trait Analysis Performed on the Family Cohort

The B vitamins contribute to DNA synthesis and methylation, with homocysteine as a by-product of their metabolism associated with coronary heart disease, stroke, and neurological disease (Tanaka et al., 2009). "A" alleles in the rs1801133 SNP within the *MTHFR* gene have been associated with higher homocysteine levels and reduced folic acid processing (Tanaka et al., 2009). We note that Father and Mother have one A allele whereas Daughter has the two A ("detrimental") alleles. Son has the two G alleles genotype. Next, the presence of the A allele in rs602662 SNP in *FUT2*, has been associated with higher B12 concentrations (Tanaka et al., 2009). We found the presence of an A allele in Father and Daughter, and no A allele presence in the other individuals. With regards to circulating concentrations of vitamin C (L-ascorbic acid), a variation at rs33972313 (*SLC23A1* gene) has been associated with a reduction in circulating concentrations of L-ascorbic acid (Timpson et al., 2010). None of the family members have the predicted detrimental allele. With regards to vitamin D, rs4588 was genotyped. Son was found homozygous for the major allele (GG) and the rest of the family heterozygous for the minor allele (GT). The effect allele for higher a-tocopherol concentration in plasma (G) is found in both alleles in Mother (GG) and one allele in Son and Daughter (Major et al., 2011, 2012, 2014; Wang and Xu, 2019).

With regards to dietary fat, we analyse a number of SNPs in genes involved in nutrition: *FTO*, *APOA2*, *FADS1*, and *LEPR*. With regards to rs9939609 *FTO* variant alleles (homozygous = AA and heterozygous = AT), both Son and Daughter are heterozygous for the risk allele and Father is homozygous for the risk allele. Each additional copy of the rs9939609 A allele has been associated with a BMI increase of a mean of 0.10 *Z*-score units, equivalent to ∼0.4 kg/m^2 (Sonestedt et al., 2009; Tanofsky-Kraff et al., 2009; Zhao et al., 2014b). For the observed genotypes in *APOA2* and *FADS1*, there is no associated effect (Yabuta et al., 2016; Huang et al., 2017; Ching et al., 2019; Moran et al., 2019; Amengual et al., 2020; Graßmann et al., 2020; Wang et al., 2020). For *LEPR*, a study found that rs1137101 AG and GG carriers with a high fat total intake had 3.0 times higher risk of obesity and 4.1 times higher risk of high cholesterol levels than those with a low intake of total fat (Domínguez-Reyes et al., 2015). All family members are carriers of the risk allele (G) of rs1137101.

The HFE protein interacts with other proteins on the cell surface to detect the amount of iron in the body (Katsarou et al., 2019). For rs1800562, a SNP in *HFE*, an A allele was not observed in any of the individuals analyzed here. This A allele causes ∼85% of all cases of hemochromatosis (Katsarou et al., 2019). The rs2187668 SNP's CC alleles in all family members have not been associated with Celiac disease (van Heel et al., 2007; Hunt et al., 2008). *MCM6-LCT* regulates lactose persistence. According to a recent study (Mattar et al., 2012), both genotypes of rs4988235 GA and AA were associated with the lactase-persistence phenotype, indicating that the presence of one single lactase-persistence allele in the heterozygous state has a dominant effect, rendering the person a lactose digester, whereas the genotype CC, when the lactase-persistence allele T is absent, is consistent with lactose maldigestion. Father's genotype was found to be CC associated with an increased likelihood of being lactose intolerant.

DISCUSSION

Our main objective is to provide insight into the current development status of personal genomics, using whole genome sequencing, illustrated by a use case of a family of five. To that end, we provide pathogenicity screening, genetic risk scoring, pharmacogenomics and fitness and nutrition trait analysis of the family. This approach is tailored for the situation where knowledge of the disease and lifestyle history of the family is used to "validate" some of the findings. A main limitation of this approach is the *post-hoc* reasoning that only allows to find *true positive* predictions based on the family observations. In contrast, those risks and phenotypes that are not reflected in the family so far can neither be confirmed nor rejected as it is unclear whether those predictions are "wrong" or whether the conditions have not had their time of onset yet. In addition, there are other limitations stemming from the different methodologies and resources used for analysis and interpretation, which we summarize in **Table 8**.

For instance, short read whole genome sequencing provides a limited capacity for detecting copy number and structural variants, which are particularly relevant for Pharmacogenomic analysis. To mitigate this shortcoming, we run prediction algorithms (**Supplementary Materials**) and find no significant prediction of copy number changes in pharmacologically important genes.

For pathogenicity screening, current standards and literature focus on genes (e.g., American College of Medical Genetics and Genomics), and therefore pathogenicity screening does not typically cover intergenic regions. Knowledge bases used for variant annotation may contain inconsistent or incomplete information, and therefore we only report variants where there is consensus among both literature, database and bioinformatic algorithm prediction, within a set of established guidelines. Moreover, while the field of genetics is evolving constantly, it is also a well-known limitation that many variants are currently classified as unknown significance. We do not report variants of unknown significance, but ensure that we use databases that remain current so that we can deploy the latest variant research in the analysis.

Concerning trait analysis in fitness and nutrition, we set a framework for selection and validation of fitness and nutrition markers to mitigate the limitations specific to phenotypes in these areas (smaller study sizes, weaker phenotype – genotype relationships). Application of this framework results in a reduction in the number of markers we were able to test in our family members. Although this filtering has restricted the number of resulting inferences, it has increased the robustness of the analysis.

Finally, the genetic risk analysis we provide here has not been tested in an independent population, and as such serves as an illustration of a potential approach and a template for further work.

Patterns of Inheritance in Pathogenicity Screening

When screening for pathogenicity we find that Son and Father have C → T; c.200C>T; p.Thr67Ile within the *CTH* gene. Father and Daughter share the mutation C → T; c.3029C>T; p.Thr1010Ile, located in the *MET* gene. Son and Mother share C → T; c.269G>A; p.Gly90Asp in *PPP2R1B*. All of these mutations are not deemed reportable due to the unknown significance nature of the inferences. The reportable variant is the A → T; c.4912A>T; p.Lys1638Ter in the *BRCA2* gene for Mother in a recessive context. Mother had a benign breast tumor removed in her forties but it was never analyzed. Therefore, it is not possible to ascertain whether her *BRCA2* gene mutation had any role in her benign tumor formation. Fortunately, this mutation is heterozygous and Father does not carry a known pathogenic mutation in this gene. Both children did not inherit Mother's pathogenic *BRCA2* mutation and therefore are unable to pass it on to their offspring.

Aunt passed away in 2013, aged 79, due to a metastasised melanoma. For this participant, we transform our screening into a quasi-diagnostic setting given that we would like to identify a potential genetic cause for her demise. We were able to retrieve 36 hairs 4 years after her death from one of her combs. The DNA was carefully handled (see **Supplementary Materials**).

TABLE 8 | List of known limitations of the methodologies we have performed for our analysis and the countermeasures we have adopted to contain them.

Methodology	Limitations	Countermeasures
Short read whole genome sequencing	• There may be errors in the variant calls • The whole genome is not wholly sequenceable • Structural and copy number variants are challenging to identify	• We performed a quality filter for each variant • Assume regions not sequenced to be gene deserts or unable to provide useful functional annotation • We run a consensus set of algorithms for prediction of copy number regions
Genome screening	• Screened only regions covered by genes and nearby regions • Incomplete, inconsistent annotations • Use of knowledge databases with conflicting results	• Selected those genes curated by OMIM where there are known mutations • Assumed that the vast majority of pathogenic mutations occur within or near coding regions • Employed a third-party protocol to interpret pathogenicity (Fabric Protocol; see Methods) • Inference only by overlapping evidence in OMIM and ClinVar, supplemented by literature search, computational algorithms and allele frequency information from established international datasets (e.g., gnomAD) • Classification of pathogenicity performed by two independent experts
Genetic risk scores	• GWAS only capture highly significant markers, missing less strongly associated markers with the trait • There may be different studies for a trait and there are challenges when integrating them into a single genetic risk score • GWAS is overwhelmingly European • Genetic risk scores may capture only a small amount of genetic risk	• We choose those studies that are of greatest number of participants, preferably from recognizable consortia, to allow the greatest possible number of markers when defining a contribution to susceptibility • We make use of the curation effort of the GWAS Catalog to select studies and markers • We compare GWAS scores with a background population (1000 Genomes Project) and check that our family participants are matched with the same background population when looking for significant differences with the average risk score • We report genetic risks only for patients whose risk is in the extreme tail of risk prediction
Pharmacogenomics	• There is a large amount of variation in pharmacological genes, not all of which can be detected • There may be cases where it is unclear the metaboliser status of a patient • Short read sequencing has limited ability to assess Copy Number Variants and therefore functional duplication or deletions of genes may be missed	• We strictly follow FDA, CPIC, ACMG guidelines when assigning metaboliser status • We make sure that when the metaboliser status is unclear we provide a range of possible eligible options • We run a consensus approach prediction algorithm (**Supplementary Materials**) to mitigate the risk that we might have failed to detect deletions or duplications within pharmacogenomic genes that may alter their functionality
Fitness	• Small sample sizes; perhaps not so much funding available as for global health conditions • Skewed populations (e.g., mostly European background) • Results often rely on self reporting of adherence to an exercise regimen • Focus in some studies on elite athletes, not necessarily generalisable to the wider population • Traits difficult to phenotype; sensors may only allow indirect measurement (e.g., VO_2max)	• Adopted an establised framework for trait analysis, so as to exclude studies with a weaker evidentiary basis • Systematically reviewed and assessed the literature choosing only those markers where there is ample evidence of their effect • No inferences of phenotype made based only on fitness marker predictions
Nutrition	• Small sample sizes; perhaps not so much funding available as for global health conditions • Skewed populations (e.g., mostly European background) • Difficult to replicate results; experimental design would use extreme fitness traits (e.g., athletes, which would contribute to difficult replication)	• Adopted an established framework for trait analysis to so as to exclude studies with a weaker evidentiary basis • Systematically reviewed and assessed the literature choosing only those markers where there is ample evidence of their effect • No inferences of phenotype made based only on fitness marker predictions
Validity of inference	• We can only confidently assign true positives	• Performed an in-depth query of the disease and lifestyle history of the family, in order to maximize our ability to confirm positive results • We use overlapping information about family members to explain predictions

We were able to assess pathogenicity among those variants that passed our strict quality filters. A heterozygous frameshift mutation, chr11:108272812 (rs587776549) in the *ATM* gene was identified. Recently, the Pan-Cancer Analysis of Whole Genomes Consortium confirmed that many cancer driver mutations are two-hit inactivation events (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), with 17% of patients having rare germline protein-truncating variants (PTVs) in cancer-predisposition genes, DNA-damage response genes and somatic driver genes. Biallelic inactivation due to somatic alteration on top of a germline PTV was observed in 4.5% of patients overall, with 81% of these affecting known cancer-predisposition genes (such as *ATM*). We thus hypothesize that the loss of function of one copy of the *ATM* gene could have contributed to her melanoma. Although Aunt's genome only provides information about her germline genetics and not the actual somatic mutations that led to the disease that ended her life, a more targeted cancer therapy (than the general chemotherapy she was administrated with) targeting defects in the DNA repair caused by *ATM* was already available while she was still alive (Kelley et al., 2014) and was never used.

Genetic Risk Scores

We observe there is a conserved family risk of ulcerative colitis, running in Son, Mother, and Daughter. Ulcerative colitis is a long-term condition that results in inflammation and ulcers of the colon and rectum. It has also been found that both Mother and Daughter have a >1SD risk of inflammatory bowel disease, of which ulcerative colitis is a type. The primary symptoms of active disease are abdominal pain and diarrhea mixed with blood. Mother has reported suffering from a recurrent abdominal pain associated with inflammation of her colon. Her symptoms have appeared intermittently but are more recurrent in older age, affecting her quality of life. Given that ulcerative colitis begins most commonly between the ages of 15 and 25 with a second peak of onset in the 6th decade of life, Mother's reported symptoms are concordant with her ulcerative colitis / bowel disease susceptibility. We also note that according to Sen and Stark (2019), *CYP2D6*4* polymorphisms may be risk factors for ulcerative colitis. Both Mother and Daughter display *CYP2D6*4*.

Pharmacological Management

We have noted that for warfarin, the genotyping analysis has shown that members of the family are either intermediate or poor metabolisers. According to FDA guidance (Dean, 2012), Daughter requires 20% of the standard initial recommended dose and would take a more prolonged time to achieve the maximum anticoagulant effect. Son, Father, and Mother require a 60% of the standard initial recommended dose. This information is particularly relevant to Father, who was recently diagnosed with atrial fibrillation. Atrial fibrillation is a heart condition that causes an irregular and often abnormally fast heart rate. People with atrial fibrillation who have a high or moderate risk of having a stroke are usually prescribed warfarin. This was the case of Father, who was recommended to take warfarin to stop the risk of blood clotting. It has been reported by Father, that as soon as he started taking warfarin, he began to experience sores in legs, changes in the skin color, and severe pain in his lower half of the body. We note that his predicted response to warfarin is concordant with warfarin sensitivity (Vu and Gooderham, 2017). Hence, knowledge of this genetic predisposition would have been helpful to the clinician when making an initial prescription.

We also note that for both Mother and Daughter their predicted metaboliser status for *CYP2D6* is either intermediate or poor. This has important implications in the specific dosage required by these individuals to receive the appropriate effects for pain relievers such as codeine and tramadol (Smith et al., 2019). So far there is some anecdotal evidence that Mother and Daughter are not able to cope well opiates, but nothing that was confirmed medically. Of note, the three most susceptible individuals to ulcerative colitis, Mother, Son, and Daughter are predicted to be normal metabolisers of drugs that treat pain associated with ulcers, such as omeprazole (Dickinson, 1994).

Fitness

We performed an investigation of the literature to identify candidate fitness gene x interactions and their relation to a health outcome (see Methods section). Several limitations were noted throughout these studies, including the robustness of significance for identified variants, small sample sizes, limited cohorts focused primarily on Caucasian populations, and minimal baseline data (Williams et al., 2017). These factors are combined with differences in exercise training programs, diet and other environmental gene expression mediators between studies. As a result, we are able to classify as "Convincing" (2 of 10 candidates) or "Probable" (5 of 10). Overall, we found that fitness studies were made with a smaller sample size compared to nutrition. For instance, *ACTN3*'s rs1815739, one of the most studied fitness-related SNPs, there are >1000 participants studied in total, which would put this SNP among the lowest sample sizes if included in nutrition markers, where we found seven markers with >100,000 study participants.

For all family members, the predicted genetic VO_2max trainability was predicted average or less than average, contrasting with their lower predicted levels of blood lactate accumulation. With the exception of Mother, the family harbor variants in *AQP1* alleles associated with endurance and fluid balance. Their genotype also predicts a predisposition to improved exercise performance if done with caffeine [with the exception of Mother; (Guest et al., 2018)]. Son is unique in the family in having the XX genotype for rs1815739 in *ACTN3*. Deficiency in α-actinin-3 can be accompanied by higher body fatness, lower muscle strength and higher muscle flexibility and range of motion (Yang et al., 2003; Massidda et al., 2019). A study suggested that recreational marathon runners who have the ACTN3 XX genotype could benefit from personalized strength training to improve their performance more than their counterparts with other *ACTN3* genotypes (Del Coso et al., 2019b).

Nutrition

Compared to fitness studies, we found a greater number of candidate nutrition phenotypes passed our filtering (14 out of 32 initially selected phenotypes; **Table 5**). Larger sample sizes and a greater number of studies with concordant results were the main reasons for a larger number of nutrition phenotypes passing our filtering. As with fitness, we choose to analyse those traits whose scientific validity score is convincing or probable and report those that are likely to display pointers for further action or deemed reportable given the family disease and lifestyle history.

Congruent with the general lower likelihood of predicted alcohol dependence by rs1229984 in *ADH1B*, there is no history of alcohol addiction in the family. For all members of the family except Mother, there is a predicted increase in total body adiposity as suggested by *FTO* rs9939609. With regards to vitamin-related traits, all family members with the exception of Son are predicted to be less likely to respond to vitamin D supplements. For vitamin B12 levels, Mother and Son are predicted to harbor lower B12 levels and higher for Father and Daughter. All family members are predicted lower serum L-ascorbic acid.

For homocysteine levels, we found that all family members except Son are predicted to be higher. Reduction of plasma homocysteine levels has been observed with supplementation of vitamin B12 and folic acid (Boccia et al., 2008; Clarke et al., 2011; Liew and Gupta, 2015). Several other studies have observed associations between lower circulating vitamin B12 levels and adverse metabolic health profiles, with insulin resistance, cardiovascular disorders, and adiposity as important features (Hazra et al., 2009; Tanwar et al., 2013; Allin et al., 2017; Nongmaithem et al., 2017; Zhao and Schooling, 2017).

With regards to lactose consumption, we were able to confirm that Father, suspected to be lactose intolerant, has the lactose intolerant genotype.

Negative Findings

When performing a screening study in an individual, for some variants which can confer significant disease risk, it is important to report not just positive findings, but also negative ones if there is family history of the disease. The ApoE2, E3, and E4 isoforms, which are encoded by the ε2, ε3, and ε4 alleles of the *APOE* gene, respectively, differ from one another at amino acid residues 112 and/or 158. There is a significant association between the ε4 allele of *APOE* and Alzheimer's disease. *APOE* ε4 increases the risk of Alzheimer's disease and lowers the age of disease onset in a gene-dose-dependent manner (Liu et al., 2013). A small proportion of apo ε2 homozygotes, develop type III hyperlipoproteinemia, a highly atherogenic form disorder of lipoprotein metabolism characterized by the accumulation of remnant particles derived from the incomplete catabolism of triglyceride-rich lipoproteins (März et al., 2000). All the family members whose genomes were analyzed for this study exhibit the wild type ε3/ε3, meaning that no association to Alzheimer's disease is conferred. This is also further supported by analysis we performed for family members using genetic risks of Alzheimer's disease (**Supplementary Table 1**). The fact that there is history of Alzheimer's disease in the maternal line, makes it interesting to ascertain whether genetic risk for this disease is present in the family. It was thus of special interest for the family to research this trait, with the positive outcome that all family members display the less risky ε3/ε3 alleles. We were also in search of negative findings for classified pathogenic mutations that fall within any of the ACMG 59 genes and were able to find only one positive finding for Mother in *BRCA2*. Our variant analysis did not find any other mutation within the 59 genes. As always, the fact that no mutation was found does not necessarily mean a particular disease might not develop.

Integration of Results

Part of the novelty of the present study revolves around the integration of genetic screening, genetic risk scores and trait analysis. A further layer of integration is constituted by the familial context our participant dataset provides. As stated in Methods, each family individual is tested independently for each of the genetic screening panels, genetic score phenotypes and trait analysis. Although the overlap between each of these methods can only be partial, we now explore the degree of consistency and support that each of the results conveys.

For obesity, the family history indicates a persistent tendency toward this phenotype. At the level of genotyping, the rs9939609 *FTO* marker analysis, shown to be the most contributing to obesity (Sonestedt et al., 2009; Tanofsky-Kraff et al., 2009; Zhao et al., 2014b), yields all genotyped individuals except Mother to carry the risk allele. We acknowledge that the specific contribution of this SNP can only be small. When integrating this genotypic result with GWAS-based genetic risk score, we observe that the obesity risk slightly increased (>1SD) for Father and Son.

Both analysis of specific SNPs in the APOE gene (see **Table 7**), and genetic risk scores do not suggest an increased risk for all family members of Alzheimer's disease. This is also congruent with the observed disease history of the analyzed family, where both parents are highly advanced in years of age (mid-eighties) and no signs for the disease have been observed yet. This does not rule out the possibility that any member of the family could develop Alzheimer's disease at any point in the future. It does rule out, however, both parents having developed early onset Alzheimer's disease. For alcohol dependency, there has not been observed any tendency of addictive behavior in the family. The rs1229984 *ADH1B* marker supports this phenotype. However, the >1SD predicted genetic risk for alcohol dependency in Father, Daughter and Son does not. A way to reconciling this result is that the genetic risk is moderately higher than average and therefore it does not strongly rule out the possibility of a false positive or random fluctuation, since the observed genetic risk for alcohol dependence may be the result of fluctuations in the score that are not significant. Another integration of different analysis sources is the pathogenic heterozygous variant for *BRCA2* p.Lys1638Ter observed in Mother and her >1SD genetic risk observed for breast cancer. An interpretation of this finding is that she only carries one defective allele for the gene, increasing her risk but not high enough to make it to our >2SD average score threshold.

With regards to integrating the results of similar (yet independent) tests performed in different individuals of the family, we note the coincidence of phenotypic history of irritable colon of Mother with her >2SD increased risk of ulcerative colitis. As mentioned earlier, this >2SD phenotype risk is also observed in Son and not so strongly in Daughter (>1SD), suggesting a pointer for preventative action on the part of Son.

Communication and Attitudes Regarding Actionable Results

Results for members of the family were communicated either in person or via phone call. For Son, his results have had an impact in his training exercise program, which has a lot more stretching and warming up, with less emphasis on speed and more on building up his endurance. The predicted ulcerative colitis risk for three members of the family was communicated to Mother, who is already displaying some symptoms of the disease, and to Son, who is already taking steps to bring these results forward to his general practitioner as part of his future health management plans. Father's possible explanation of his adverse reaction to warfarin has also been discussed and he has currently discontinued taking the medicine, having discussed it first with his cardiologist. The communication of Aunt's result of her mutation in the *ATM* gene has not led to any concrete actions by her partner.

The family has been exposed to genetic testing for a decade (Corpas, 2012), and as such were generally comfortable knowing results of genetic analysis. Even so, attention was paid in particular to make sure they were aware of the ramifications of knowing their results of tests related to more serious and harder to treat conditions (such as ulcerative colitis).

As was the case when the same family was analyzed with direct-to-consumer genotyping methods (Corpas, 2012), the tendency to discuss "whose genome is best" is a recurrent pattern that could affect other families when communicating genetic test results. We stress the importance of discussing such results with qualified professionals such as genetic counselors.

Compared to the communication of the results in 2012, we also note the change in attitude toward sharing of personal genomic data. Individuals are less keen to share their genetic data now, arguing that their perceptions regarding the privacy of their data have been changed by their increased awareness of the importance of protecting individual's personal data.

CONCLUSION

By looking at the genome from various methodological angles and applying distinct analytical frameworks as appropriate, we were able to build a "genetic story" of each individual. We built this story in part through having whole genomes as the basis for the analysis. The approach is applied here for a family, but we believe it is also valid for individuals. Our most notable findings for the family were around susceptibility to ulcerative colitis, and in the areas of fitness, nutrition, and pharmacogenomics.

Concerning ulcerative colitis, when analyzing genetic risk scores, we noted that the recurrent intestinal pain Mother has been affected from for years is concordant with her substantially increased risk of suffering from ulcerative colitis. Moreover, this high risk is predicted in three out of the five members of the family, two of them overlapping with increased risk of inflammatory bowel disease, ulcerative colitis being one type of this disease. We report this susceptibility to ulcerative colitis/inflammatory bowel disease as a potential lead for preventative intervention in at least one family member (Son) who is currently asymptomatic.

We observed some associations for fitness and nutrition variants which passed our quality control framework and as such we believe are valuable for relevant nutritional and exercise science specialists to help the family in making plans in those areas.

We were also able to hypothesize a genetic contribution to the development of melanoma leading to the passing of Aunt. A pathogenic heterozygous germline mutation was reported in her *ATM* gene. This gene has been described as being involved in DNA repair and the information gathered here could have been exploited for targeted cancer therapy if caught on time.

Concordance between an adverse reaction to warfarin and a prediction for low dosing requirement was observed in Father, which he has already acted on, in consultation with his cardiologist. There were also informative results for Mother and Daughter regarding their likely metaboliser status for certain drugs. While not relevant to them at the moment, this information could be shared with their physician in the event that these drugs become necessary in the future, with the hope of reducing trial and error in prescribing and so cutting down the possibility of adverse reactions.

We believe that, taken together, these results represent relevant information which the family can use, when working with the appropriate healthcare professionals, to proactively promote their health and well-being. Any one element of the analysis would not allow this genetic "story" to be compellingly told, but when all them are put together, the narrative becomes more actionable, increasing the applicability of whole genome screening to pre-emptive healthcare and well-being management.

DATA AVAILABILITY STATEMENT

The sources of genetic markers used in this study are included in **Table 3** and **Table 4**. The sources for genetic risk score creation are included in **Supplementary Table 1**. The gene panel utilised for cancer screening is in the Supplementary materials. The family genome variation data for this manuscript is not publicly available because they were not consented for open access. Request to access the family genome variation data should be directed to Manuel Corpas (m.corpas@cpm.onl).

ETHICS STATEMENT

We confirm that written informed consent was obtained from the individuals or appropriate next to kin individual for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

MC and EL conceived the experiments and performed the analysis. MC wrote the paper with contributions from all authors. All authors read and approved the paper.

ACKNOWLEDGMENTS

We are grateful to Mahsa Shabani for helpful guidance on the development of an ethical framework for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.535123/full#supplementary-material

REFERENCES

Ahmetov, I. I., and Fedotovskaya, O. N. (2015). Current progress in sports genomics. *Adv. Clin. Chem.* 70, 247–314. doi: 10.1016/bs.acc.2015.03.003

Ahmetov, I. I., Williams, A. G., Popov, D. V., Lyubaeva, E. V., Hakimullina, A. M., Fedotovskaya, O. N., et al. (2009). The combined impact of metabolic gene polymorphisms on elite endurance athlete status and related phenotypes. *Hum. Genet.* 126, 751–761. doi: 10.1007/s00439-009-0728-4

Al-Daghri, N. M., Mohammed, A. K., Bukhari, I., Rikli, M., Abdi, S., Ansari, M. G. A., et al. (2019). Efficacy of vitamin D supplementation according to vitamin D-binding protein polymorphisms. *Nutrition* 63–64, 148–154. doi: 10.1016/j.nut.2019.02.003

Allin, K. H., Friedrich, N., Pietzner, M., Grarup, N., Thuesen, B. H., Linneberg, A., et al. (2017). Genetic determinants of serum vitamin B12 and their relation to body mass index. *Eur. J. Epidemiol.* 32, 125–134. doi: 10.1007/s10654-016-0215-x

Almeida, S. M., Furtado, J. M., Mascarenhas, P., Ferraz, M. E., Ferreira, J. C., Monteiro, M. P., et al. (2018). Association between LEPR, FTO, MC4R, and PPARG-2 polymorphisms with obesity traits and metabolic phenotypes in school-aged children. *Endocrine* 60, 466–478. doi: 10.1007/s12020-018-1587-3

Alsaleh, A., Crepostnaia, D., Maniou, Z., Lewis, F. J., Hall, W. L., Sanders, T. A. B., et al. (2013). Adiponectin gene variant interacts with fish oil supplementation to influence serum adiponectin in older individuals. *J. Nutr.* 143, 1021–1027. doi: 10.3945/jn.112.172585

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Amengual, J., Coronel, J., Marques, C., Aradillas-García, C., Morales, J. M. V., Andrade, F. C. D., et al. (2020). β-Carotene oxygenase 1 activity modulates circulating cholesterol concentrations in mice and humans. *J. Nutr.* 150, 2023–2030. doi: 10.1093/jn/nxaa143

Amir Shaghaghi, M., Bernstein, C. N., Serrano León, A., El-Gabalawy, H., and Eck, P. (2014). Polymorphisms in the sodium-dependent ascorbate transporter gene SLC23A1 are associated with susceptibility to Crohn disease. *Am. J. Clin. Nutr.* 99, 378–383. doi: 10.3945/ajcn.113.068015

Baffour-Awuah, N. Y., Fleet, S., Montgomery, R. K., Baker, S. S., Butler, J. L., Campbell, C., et al. (2015). Functional significance of single nucleotide polymorphisms in the lactase gene in diverse US patients and evidence for a novel lactase persistence allele at−13909 in those of European ancestry. *J. Pediatr. Gastroenterol. Nutr.* 60, 182–191. doi: 10.1097/MPG.0000000000000595

Bahrami, A., Khayyatzadeh, S. S., Jaberi, N., Tayefi, M., Mohammadi, F., Ferns, G. A., et al. (2019). Common polymorphisms in genes related to vitamin D metabolism affect the response of cognitive abilities to vitamin D supplementation. *J. Mol. Neurosci.* 69, 150–156. doi: 10.1007/s12031-019-01344-6

Balsevich, G., Sticht, M., Bowles, N. P., Singh, A., Lee, T. T. Y., Li, Z., et al. (2018). Role for fatty acid amide hydrolase (FAAH) in the leptin-mediated effects on feeding and energy balance. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7605–7610. doi: 10.1073/pnas.1802251115

Baltazar-Martins, G., Gutiérrez-Hellín, J., Aguilar-Navarro, M., Ruiz-Moreno, C., Moreno-Pérez, V., López-Samanes, Á., et al. (2020). Effect of genotype on sports performance, exercise-induced muscle damage, and injury epidemiology. *Sports* 8:99. doi: 10.3390/sports8070099

Barton, J. C., McLaren, C. E., Chen, W.-P., Ramm, G. A., Anderson, G. J., Powell, L. W., et al. (2018). Cirrhosis in hemochromatosis: independent risk factors in 368 HFE p.C282Y homozygotes. *Ann. Hepatol.* 17, 871–879. doi: 10.5604/01.3001.0012.3169

Baumert, P., Lake, M. J., Stewart, C. E., Drust, B., and Erskine, R. M. (2016). Genetic variation and exercise-induced muscle damage: implications for athletic performance, injury and ageing. *Eur. J. Appl. Physiol.* 116, 1595–1625. doi: 10.1007/s00421-016-3411-1

Ben-Zaken, S., Eliakim, A., Nemet, D., Rabinovich, M., Kassem, E., and Meckel, Y. (2015). Differences in MCT1 A1470T polymorphism prevalence between runners and swimmers. *Scand. J. Med. Sci. Sports* 25, 365–371. doi: 10.1111/sms.12226

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* 45, 501–512. doi: 10.1038/ng.2606

Bertuzzi, R., Pasqua, L. A., Bueno, S., Lima-Silva, A. E., Matsuda, M., Marquezini, M., et al. (2014). Is the COL5A1 rs12722 gene polymorphism associated with running economy? *PLoS ONE* 9:e106581. doi: 10.1371/journal.pone.0106581

Bierut, L. J., Agrawal, A., Bucholz, K. K., Doheny, K. F., Laurie, C., Pugh, E., et al. (2010). A genome-wide association study of alcohol dependence. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5082–5087. doi: 10.1073/pnas.0911109107

Biesecker, L. G., the ClinGen Sequence Variant Interpretation Working Group, and Harrison, S. M. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet. Med.* 20, 1687–1688. doi: 10.1038/gim.2018.42

Biss, T. T., Avery, P. J., Brandão, L. R., Chalmers, E. A., Williams, M. D., Grainger, J. D., et al. (2012). VKORC1 and CYP2C9 genotype and patient characteristics explain a large proportion of the variability in warfarin dose requirement among children. *Blood* 119, 868–873. doi: 10.1182/blood-2011-08-372722

Boccia, S., Boffetta, P., Brennan, P., Ricciardi, G., Gianfagna, F., Matsuo, K., et al. (2009). Meta-analyses of the methylenetetrahydrofolate reductase C677T and A1298C polymorphisms and risk of head and neck and lung cancer. *Cancer Lett.* 273, 55–61. doi: 10.1016/j.canlet.2008.07.026

Boccia, S., Hung, R., Ricciardi, G., Gianfagna, F., Ebert, M. P. A., Fang, J.-Y., et al. (2008). Meta- and pooled analyses of the methylenetetrahydrofolate reductase C677T and A1298C polymorphisms and gastric cancer risk: a huge-GSEC review. *Am. J. Epidemiol.* 167, 505–516. doi: 10.1093/aje/kwm344

Bollepalli, S., Dolan, L. M., Deka, R., and Martin, L. J. (2010). Association of FTO gene variants with adiposity in African-American adolescents. *Obesity* 18, 1959–1963. doi: 10.1038/oby.2010.82

Bordoni, L., Marchegiani, F., Piangerelli, M., Napolioni, V., and Gabbianelli, R. (2017). Obesity-related genetic polymorphisms and adiposity indices in a young Italian population. *IUBMB Life* 69, 98–105. doi: 10.1002/iub.1596

Bouchard, C., Sarzynski, M. A., Rice, T. K., Kraus, W. E., Church, T. S., Sung, Y. J., et al. (2011). Genomic predictors of the maximal O_2 uptake response to standardized exercise training programs. *J. Appl. Physiol.* 110, 1160–1170. doi: 10.1152/japplphysiol.00973.2010

Braithwaite, V. S., Jones, K. S., Schoenmakers, I., Silver, M., Prentice, A., and Hennig, B. J. (2015). Vitamin D binding protein genotype is associated with plasma 25OHD concentration in West African children. *Bone* 74, 166–170. doi: 10.1016/j.bone.2014.12.068

Brandt, T., Sack, L. M., Arjona, D., Tan, D., Mei, H., Cui, H., et al. (2019). Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. *Genet. Med.* 22, 336–344. doi: 10.1038/s41436-019-0655-2

Calvano Küchler, E., Arid, J., Palinkas, M., Ayumi Omori, M., de Lara, R. M., Napolitano Gonçalves, L. M., et al. (2020). Genetic polymorphisms in contribute to the etiology of bruxism in children. *J. Clin. Pediatr. Dent.* 44, 180–184. doi: 10.17796/1053-4625-44.3.8

Carswell, A. T., Howland, K., Martinez-Gonzalez, B., Baron, P., and Davison, G. (2020). The effect of caffeine on cognitive performance is influenced by CYP1A2 but not ADORA2A genotype, yet neither genotype affects exercise performance in healthy adults. *Eur. J. Appl. Physiol.* 120, 1495–1508. doi: 10.1007/s00421-020-04384-8

Ching, Y. K., Chin, Y. S., Appukutty, M., Ramanchadran, V., Yu, C. Y., Ang, G. Y., et al. (2019). Interaction of dietary linoleic acid and α-linolenic acids with rs174547 in gene on metabolic syndrome components among vegetarians. *Nutrients* 11:1686. doi: 10.3390/nu11071686

Choi, M., Kipps, T., and Kurzrock, R. (2016). ATM mutations in cancer: therapeutic implications. *Mol. Cancer Ther.* 15, 1781–1791. doi: 10.1158/1535-7163.MCT-15-0945

Choquet, H., Paylakhi, S., Kneeland, S. C., Thai, K. K., Hoffmann, T. J., Yin, J., et al. (2018). A multiethnic genome-wide association study of primary open-angle glaucoma identifies novel risk loci. *Nat. Commun.* 9, 2278. doi: 10.1038/s41467-018-04555-4

Chuaychoo, B., Tungtrongchitr, R., Kriengsinyos, W., Tuntipopipat, S., On-Nom, N., and Chupeerach, C. (2018). Correlation of vitamin D binding protein gene polymorphism and protein levels in chronic obstructive pulmonary disease compared with non-chronic obstructive pulmonary disease subjects. *Per. Med.* 15, 371–379. doi: 10.2217/pme-2018-0005

Cichon, S., Mühleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miró, X., Strohmaier, J., et al. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am. J. Hum. Genet.* 88, 372–381. doi: 10.1016/j.ajhg.2011.01.017

Clarke, R., Halsey, J., Bennett, D., and Lewington, S. (2011). Homocysteine and vascular disease: review of published results of the homocysteine-lowering trials. *J. Inherit. Metab. Dis.* 34, 83–91. doi: 10.1007/s10545-010-9235-y

Coonrod, E. M., Margraf, R. L., Russell, A., Voelkerding, K. V., and Reese, M. G. (2013). Clinical analysis of genome next-generation sequencing data using the Omicia platform. *Expert Rev. Mol. Diagn.* 13, 529–540. doi: 10.1586/14737159.2013.811907

Cormier, H., Rudkowska, I., Lemieux, S., Couture, P., and Vohl, M.-C. (2016). Expression and Sequence Variants of Inflammatory Genes; Effects on Plasma Inflammation Biomarkers Following a 6-Week Supplementation with Fish Oil. *Int. J. Mol. Sci.* 17:375. doi: 10.3390/ijms17030375

Cornelis, M. C., Kacprowski, T., Menni, C., Gustafsson, S., Pivin, E., Adamski, J., et al. (2016). Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Hum. Mol. Genet.* 25, 5472–5482. doi: 10.1093/hmg/ddw334

Corpas, M. (2012). A family experience of personal genomics. *J. Genet. Counsel.* 21, 386–391. doi: 10.1007/s10897-011-9473-7

Corpas, M., Valdivia-Granda, W., Torres, N., Greshake, B., Coletta, A., Knaus, A., et al. (2015). Crowdsourced direct-to-consumer genomic analysis of a family quartet. *BMC Genomics* 16:910. doi: 10.1186/s12864-015-1973-7

Cupeiro, R., González-Lamuño, D., Amigo, T., Peinado, A. B., Ruiz, J. R., Ortega, F. B., et al. (2012). Influence of the MCT1-T1470A polymorphism (rs1049434) on blood lactate accumulation during different circuit weight trainings in men and women. *J. Sci. Med. Sport* 15, 541–547. doi: 10.1016/j.jsams.2012.03.009

da Rocha, T. J., Blehm, C. J., Bamberg, D. P., Fonseca, T. L. R., Tisser, L. A., de Oliveira Junior, A. A., et al. (2014a). The effects of interactions between selenium and zinc serum concentration and SEP15 and SLC30A3 gene polymorphisms on memory scores in a population of mature and elderly adults. *Genes Nutr.* 9:377. doi: 10.1007/s12263-013-0377-z

da Rocha, T. J., Korb, C., Schuch, J. B., Bamberg, D. P., de Andrade, F. M., and Fiegenbaum, M. (2014b). SLC30A3 and SEP15 gene polymorphisms influence the serum concentrations of zinc and selenium in mature adults. *Nutr. Res.* 34, 742–748. doi: 10.1016/j.nutres.2014.08.009

de Luis, D. A., Gonzalez Sagrado, M., Aller, R., Izaola, O., and Conde, R. (2011). Effects of C358A missense polymorphism of the endocannabinoid degrading enzyme fatty acid amide hydrolase on weight loss after a hypocaloric diet. *Metabolism* 60, 730–734. doi: 10.1016/j.metabol.2010.07.007

Dean, L. (2012). "Warfarin therapy and genotype," in *Medical Genetics Summaries,* eds V. M. Pratt, S. A. Scott, M. Pirmohamed, B. Esquivel, M. S. Kane, B. L. Kattman et al. [Bethesda, MD: National Center for Biotechnology Information (US)].

Dean, L. (2018). "Warfarin therapy and VKORC1 and CYP genotype," in *Medical Genetics Summaries [Internet]*, eds V. M. Pratt, S. A. Scott, M. Pirmohamed, B. Esquivel, M. S. Kane, B. L. Kattman et al. [Bethesda, MD: National Center for Biotechnology Information (US)].

Dedoussis, G. V. Z., Yannakoulia, M., Timpson, N. J., Manios, Y., Kanoni, S., Scott, R. A., et al. (2011). Does a short breastfeeding period protect from FTO-induced adiposity in children? *Int. J. Pediatr. Obes.* 6, e326–e335. doi: 10.3109/17477166.2010.490269

Del Coso, J., Hiam, D., Houweling, P., Pérez, L. M., Eynon, N., and Lucía, A. (2019a). More than a "speed gene": ACTN3 R577X genotype, trainability, muscle damage, and the risk for injuries. *Eur. J. Appl. Physiol.* 119, 49–60. doi: 10.1007/s00421-018-4010-0

Del Coso, J., Moreno, V., Gutiérrez-Hellín, J., Baltazar-Martins, G., Ruíz-Moreno, C., Aguilar-Navarro, M., et al. (2019b). R577X genotype and exercise phenotypes in recreational marathon runners. *Genes* 10:413. doi: 10.3390/genes10060413

Del Coso, J., Salinero, J. J., Lara, B., Gallo-Salazar, C., Areces, F., Puente, C., et al. (2017). ACTN3 X-allele carriers had greater levels of muscle damage during a half-ironman. *Eur. J. Appl. Physiol.* 117, 151–158. doi: 10.1007/s00421-016-3507-7

Delgado-Lista, J., Perez-Jimenez, F., Tanaka, T., Perez-Martinez, P., Jimenez-Gomez, Y., Marin, C., et al. (2007). An apolipoprotein A-II polymorphism (-265T/C, rs5082) regulates postprandial response to a saturated fat overload in healthy men. *J. Nutr.* 137, 2024–2028. doi: 10.1093/jn/137.9.2024

Dickinson, J. B. (1994). Is omeprazole helpful in inflammatory bowel disease? *J. Clin. Gastroenterol.* 18, 317–319. doi: 10.1097/00004836-199406000-00012

Domínguez-Reyes, T., Astudillo-López, C. C., Salgado-Goytia, L., Muñoz-Valle, J. F., Salgado-Bernabé, A. B., Guzmán-Guzmán, I. P., et al. (2015). Interaction of dietary fat intake with APOA2, APOA5 and LEPR polymorphisms and its relationship with obesity and dyslipidemia in young subjects. *Lipids Health Dis.* 14:106. doi: 10.1186/s12944-015-0112-4

Doris, J. M., Millar, S. A., Idris, I., and O'Sullivan, S. E. (2019). Genetic polymorphisms of the endocannabinoid system in obesity and diabetes. *Diabetes Obes. Metab.* 21, 382–387. doi: 10.1111/dom.13504

Duell, E. J., Lujan-Barroso, L., Llivina, C., Muñoz, X., Jenab, M., Boutron-Ruault, M.-C., et al. (2013). Vitamin C transporter gene (SLC23A1 and SLC23A2) polymorphisms, plasma vitamin C levels, and gastric cancer risk in the EPIC cohort. *Genes Nutr.* 8, 549–560. doi: 10.1007/s12263-013-0346-6

Duicu, C., Mărginean, C. O., Voidăzan, S., Tripon, F., and Bănescu, C. (2016). FTO rs 9939609 SNP is associated with adiponectin and leptin levels and the risk of obesity in a cohort of romanian children population. *Medicine* 95:e3709. doi: 10.1097/MD.0000000000003709

Dwivedi, O. P., Tabassum, R., Chauhan, G., Ghosh, S., Marwaha, R. K., Tandon, N., et al. (2012). Common variants of FTO are associated with childhood obesity in a cross-sectional study of 3,126 urban Indian children. *PLoS ONE* 7:e47772. doi: 10.1371/journal.pone.0047772

Elkader, A., and Sproule, B. (2005). Buprenorphine: clinical pharmacokinetics in the treatment of opioid dependence. *Clin. Pharmacokinet.* 44, 661–680. doi: 10.2165/00003088-200544070-00001

Enlund-Cerullo, M., Koljonen, L., Holmlund-Suila, E., Hauta-Alus, H., Rosendahl, J., Valkama, S., et al. (2019). Genetic variation of the vitamin D binding protein affects vitamin D status and response to supplementation in infants. *J. Clin. Endocrinol. Metab.* 104, 5483–5498. doi: 10.1210/jc.2019-00630

Eynon, N., Alves, A. J., Sagiv, M., Yamin, C., Sagiv, M., and Meckel, Y. (2010). Interaction between SNPs in the NRF2 gene and elite endurance performance. *Physiol. Genomics* 41, 78–81. doi: 10.1152/physiolgenomics.00199.2009

Eynon, N., Ruiz, J. R., Bishop, D. J., Santiago, C., Gómez-Gallego, F., Lucia, A., et al. (2013). The rs12594956 polymorphism in the NRF-2 gene is associated with top-level Spanish athlete's performance status. *J. Sci. Med. Sport* 16, 135–139. doi: 10.1016/j.jsams.2012.05.004

Fahed, A. C., Wang, M., Homburger, J. R., Patel, A. P., Bick, A. G., Neben, C. L., et al. (2020). Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* 11:3635. doi: 10.1038/s41467-020-17374-3

Fedotovskaya, O. N., Mustafina, L. J., Popov, D. V., Vinogradova, O. L., and Ahmetov, I. I. (2014). A common polymorphism of the MCT1 gene and athletic performance. *Int. J. Sports Physiol. Perform.* 9, 173–180. doi: 10.1123/ijspp.2013-0026

Ferreira Todendi, P., de Moura Valim, A. R., Klinger, E., Reuter, C. P., Molina, S., Martínez, J. A., et al. (2019). The role of the genetic variants IRX3 rs3751723 and FTO rs9939609 in the obesity phenotypes of children and adolescents. *Obes. Res. Clin. Pract.* 13, 137–142. doi: 10.1016/j.orcp.2019.01.005

Ferreira, M. A. R., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* 40, 1056–1058. doi: 10.1038/ng.209

Figueroa, J. D., Ye, Y., Siddiq, A., Garcia-Closas, M., Chatterjee, N., Prokunina-Olsson, L., et al. (2014). Genome-wide association study identifies multiple loci associated with bladder cancer risk. *Hum. Mol. Genet.* 23, 1387–1398. doi: 10.1093/hmg/ddt519

Flygare, S., Hernandez, E. J., Phan, L., Moore, B., Li, M., Fejes, A., et al. (2018). The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool. *BMC Bioinformatics* 19:57. doi: 10.1186/s12859-018-2056-y

Gaedigk, A., Dinh, J. C., Jeong, H., Prasad, B., and Leeder, J. S. (2018). Ten years' experience with the CYP2D6 activity score: a perspective on future investigations to improve clinical predictions for precision therapeutics. *J. Pers. Med.* 8:15. doi: 10.3390/jpm8020015

Galeandro, V., Notarnicola, A., Bianco, A., Tafuri, S., Russo, L., Pesce, V., et al. (2017). ACTN3/ACE genotypes and mitochondrial genome in professional soccer players performance. *J. Biol. Regul. Homeost. Agents* 31, 207–213.

García-Solís, P., Reyes-Bastidas, M., Flores, K., García, O. P., Rosado, J. L., Méndez-Villa, L., et al. (2016). Fat mass obesity-associated (FTO) (rs9939609) and melanocortin 4 receptor (MC4R) (rs17782313) SNP are positively associated with obesity and blood pressure in Mexican school-aged children. *Br. J. Nutr.* 116, 1834–1840. doi: 10.1017/S0007114516003779

Gelernter, J., Kranzler, H. R., Sherva, R., Almasy, L., Koesterer, R., Smith, A. H., et al. (2014). Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* 19, 41–49. doi: 10.1038/mp.2013.145

Georgi, B., Mielke, J., Chaffin, M., Khera, A. V., Gelis, L., Mundl, H., et al. (2019). Leveraging human genetics to estimate clinical risk reductions achievable by inhibiting factor XI. *Stroke* 50, 3004–3012. doi: 10.1161/STROKEAHA.119.026545

Ghosh, S., Vivar, J. C., Sarzynski, M. A., Sung, Y. J., Timmons, J. A., Bouchard, C., et al. (2013). Integrative pathway analysis of a genome-wide association study of (V)O(2max) response to exercise training. *J. Appl. Physiol.* 115, 1343–1359. doi: 10.1152/japplphysiol.01487.2012

Gibbs, D. C., Bostick, R. M., McCullough, M. L., Um, C. Y., Flanders, W. D., Jenab, M., et al. (2020a). Association of prediagnostic vitamin D status with mortality among colorectal cancer patients differs by common, inherited vitamin D-binding protein isoforms. *Int. J. Cancer* 147, 2725–2734. doi: 10.1002/ijc.33043

Gibbs, D. C., Song, M., McCullough, M. L., Um, C. Y., Bostick, R. M., Wu, K., et al. (2020b). Association of circulating vitamin D with colorectal cancer depends on vitamin D-binding protein isoforms: a pooled, nested, case-control study. *JNCI Cancer Spectr.* 4:kz083. doi: 10.1093/jncics/pkz083

Glusman, G., Cariaso, M., Jimenez, R., Swan, D., Greshake, B., Bhak, J., et al. (2012). Low budget analysis of Direct-To-Consumer genomic testing familial data. *F1000Res.* 1:3. doi: 10.12688/f1000research.1-3.v1

Gómez-Gallego, F., Ruiz, J. R., Buxens, A., Altmäe, S., Artieda, M., Santiago, C., et al. (2010). Are elite endurance athletes genetically predisposed to lower disease risk? *Physiol. Genomics* 41, 82–90. doi: 10.1152/physiolgenomics.00183.2009

Graßmann, S., Pivovarova-Ramich, O., Henze, A., Raila, J., Ampem Amoako, Y., King Nyamekye, R., et al. (2020). SNP rs6564851 in the BCO1 gene is associated with varying provitamin a plasma concentrations but not with retinol concentrations among adolescents from rural Ghana. *Nutrients* 12:1786. doi: 10.3390/nu12061786

Grgic, J., Pickering, C., Bishop, D. J., Schoenfeld, B. J., Mikulic, P., and Pedisic, Z. (2020). CYP1A2 genotype and acute effects of caffeine on resistance exercise, jumping, and sprinting performance. *J. Int. Soc. Sports Nutr.* 17:21. doi: 10.1186/s12970-020-00349-6

Grimaldi, K. A., van Ommen, B., Ordovas, J. M., Parnell, L. D., Mathers, J. C., Bendik, I., et al. (2017). Proposed guidelines to evaluate scientific validity and evidence for genotype-based dietary advice. *Genes Nutr.* 12:35. doi: 10.1186/s12263-017-0584-0

Guest, N., Corey, P., Vescovi, J., and El-Sohemy, A. (2018). Caffeine, CYP1A2 genotype, and endurance performance in athletes. *Med. Sci. Sports Exerc.* 50, 1570–1578. doi: 10.1249/MSS.0000000000001596

Guevara-Cruz, M., Lai, C.-Q., Richardson, K., Parnell, L. D., Lee, Y.-C., Tovar, A. R., et al. (2013). Effect of a GFOD2 variant on responses in total and LDL cholesterol in Mexican subjects with hypercholesterolemia after soy protein and soluble fiber supplementation. *Gene* 532, 211–215. doi: 10.1016/j.gene.2013.09.055

Hällfors, J., Palviainen, T., Surakka, I., Gupta, R., Buchwald, J., Raevuori, A., et al. (2019). Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway. *Addict. Biol.* 24, 549–561. doi: 10.1111/adb.12618

Hazra, A., Kraft, P., Lazarus, R., Chen, C., Chanock, S. J., Jacques, P., et al. (2009). Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. *Hum. Mol. Genet.* 18, 4677–4687. doi: 10.1093/hmg/ddp428

He, Z., Hu, Y., Feng, L., Lu, Y., Liu, G., Xi, Y., et al. (2007). NRF2 genotype improves endurance capacity in response to training. *Int. J. Sports Med.* 28, 717–721. doi: 10.1055/s-2007-964913

Hindy, G., Mollet, I. G., Rukh, G., Ericson, U., and Orho-Melander, M. (2016). Several type 2 diabetes-associated variants in genes annotated to WNT signaling interact with dietary fiber in relation to incidence of type 2 diabetes. *Genes Nutr.* 11:6. doi: 10.1186/s12263-016-0524-4

Hindy, G., Sonestedt, E., Ericson, U., Jing, X.-J., Zhou, Y., Hansson, O., et al. (2012). Role of TCF7L2 risk variant and dietary fibre intake on incident type 2 diabetes. *Diabetologia* 55, 2646–2654. doi: 10.1007/s00125-012-2634-x

Hou, L., Bergen, S. E., Akula, N., Song, J., Hultman, C. M., Landén, M., et al. (2016). Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum. Mol. Genet.* 25, 3383–3394. doi: 10.1093/hmg/ddw181

Houweling, P. J., Papadimitriou, I. D., Seto, J. T., Pérez, L. M., Coso, J. D., North, K. N., et al. (2018). Is evolutionary loss our gain? The role of ACTN3 p.Arg577Ter (R577X) genotype in athletic performance, ageing, and disease. *Hum. Mutat.* 39, 1774–1787. doi: 10.1002/humu.23663

Howard, D. M., Adams, M. J., Shirali, M., Clarke, T.-K., Marioni, R. E., Davies, G., et al. (2018). Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* 9:1470. doi: 10.1038/s41467-018-03819-3

Howe, L. J., Lawson, D. J., Davies, N. M., St Pourcain, B., Lewis, S. J., Davey Smith, G., et al. (2019). Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. *Nat. Commun.* 10:5039. doi: 10.1038/s41467-019-12424-x

Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., and Yandell, M. (2013). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* 37, 622–634. doi: 10.1002/gepi.21743

Huang, M.-C., Chang, W.-T., Chang, H.-Y., Chung, H.-F., Chen, F.-P., Huang, Y.-F., et al. (2017). FADS gene polymorphisms, fatty acid desaturase activities, and HDL-C in type 2 diabetes. *Int. J. Environ. Res. Public Health* 14:572. doi: 10.3390/ijerph14060572

Hubacek, J. A., Jirsa, M., Bobak, M., Pelclova, D., and Zakharov, S. (2018). Aldehyde dehydrogenase 2 polymorphism affects the outcome of methanol poisoning in exposed humans. *Clin. Genet.* 94, 445–449. doi: 10.1111/cge.13411

Hunt, K. A., Zhernakova, A., Turner, G., Heap, G. A. R., Franke, L., Bruinenberg, M., et al. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* 40, 395–402. doi: 10.1038/ng.102

ICGC/TCGA and Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi: 10.1038/s41586-020-1969-6

Ikeda, M., Takahashi, A., Kamatani, Y., Okahisa, Y., Kunugi, H., Mori, N., et al. (2018). A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol. Psychiatry* 23, 639–647. doi: 10.1038/mp.2016.259

Itaka, T., Agemizu, K., Aruga, S., and Machida, S. (2016). G allele of the IGF2 ApaI polymorphism is associated with judo status. *J. Strength Cond. Res.* 30, 2043–2048. doi: 10.1519/JSC.0000000000001300

Jenab, M., McKay, J., Bueno-de-Mesquita, H. B., van Duijnhoven, F. J. B., Ferrari, P., Slimani, N., et al. (2009). Vitamin D receptor and calcium sensing receptor polymorphisms and the risk of colorectal cancer in European populations. *Cancer Epidemiol. Biomarkers Prev.* 18, 2485–2491. doi: 10.1158/1055-9965.EPI-09-0319

Jensen, D. P., Andreasen, C. H., Andersen, M. K., Hansen, L., Eiberg, H., Borch-Johnsen, K., et al. (2007). The functional Pro129Thr variant of the FAAH gene is not associated with various fat accumulation phenotypes in a population-based cohort of 5,801 whites. *J. Mol. Med.* 85, 445–449. doi: 10.1007/s00109-006-0139-0

Johansen, J.-M., Goleva-Fjellet, S., Sunde, A., Gjerløw, L. E., Skeimo, L. A., Freberg, B. I., et al. (2020). No change - no gain; the effect of age, sex, selected genes and training on physiological and performance adaptations in cross-country skiing. *Front. Physiol.* 11:581339. doi: 10.3389/fphys.2020.581339

Johnson, E. C., St Pierre, C. L., Meyers, J. L., Aliev, F., McCutcheon, V. V., Lai, D., et al. (2019). The genetic relationship between alcohol consumption and aspects of problem drinking in an ascertained sample. *Alcohol. Clin. Exp. Res.* 43, 1113–1125. doi: 10.1111/acer.14064

Jorgenson, E., Thai, K. K., Hoffmann, T. J., Sakoda, L. C., Kvale, M. N., Banda, Y., et al. (2017). Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* 22, 1359–1367. doi: 10.1038/mp.2017.101

Junyent, M., Parnell, L. D., Lai, C.-Q., Lee, Y.-C., Smith, C. E., Arnett, D. K., et al. (2009). Novel variants at KCTD10, MVK, and MMAB genes interact with dietary carbohydrates to modulate HDL-cholesterol concentrations in the Genetics of Lipid Lowering Drugs and Diet Network Study. *Am. J. Clin. Nutr.* 90, 686–694. doi: 10.3945/ajcn.2009.27738

Justice, A. C., Smith, R. V., Tate, J. P., McGinnis, K., Xu, K., Becker, W. C., et al. (2018). AUDIT-C and ICD codes as phenotypes for harmful alcohol use: association with ADH1B polymorphisms in two US populations. *Addiction* 113, 2214–2224. doi: 10.1111/add.14374

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). 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. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Karuwanarint, P., Phonrat, B., Tungtrongchitr, A., Suriyaprom, K., Chuengsamarn, S., Schweigert, F. J., et al. (2018). Vitamin D-binding protein and its polymorphisms as a predictor for metabolic syndrome. *Biomark. Med.* 12, 465–473. doi: 10.2217/bmm-2018-0029

Katsarou, M.-S., Karakonstantis, K., Demertzis, N., Vourakis, E., Skarpathioti, A., Nosyrev, A. E., et al. (2017). Effect of single-nucleotide polymorphisms in ADH1B, ADH4, ADH1C, OPRM1, DRD2, BDNF, and ALDH2 genes on alcohol dependence in a Caucasian population. *Pharmacol. Res. Perspect.* 5:e00326. doi: 10.1002/prp2.326

Katsarou, M.-S., Latsi, R., Papasavva, M., Demertzis, N., Kalogridis, T., Tsatsakis, A. M., et al. (2016). Population-based analysis of the frequency of HFE gene polymorphisms: Correlation with the susceptibility to develop hereditary hemochromatosis. *Mol. Med. Rep.* 14, 630–636. doi: 10.3892/mmr.2016.5317

Katsarou, M.-S., Papasavva, M., Latsi, R., and Drakoulis, N. (2019). Hemochromatosis: hereditary hemochromatosis and HFE gene. *Vitam. Horm.* 110, 201–222. doi: 10.1016/bs.vh.2019.01.010

Kelley, M. R., Logsdon, D., and Fishel, M. L. (2014). Targeting DNA repair pathways for cancer treatment: what's new? *Future Oncol.* 10, 1215–1237. doi: 10.2217/fon.14.60

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Kiemeney, L. A., Sulem, P., Besenbacher, S., Vermeulen, S. H., Sigurdsson, A., Thorleifsson, G., et al. (2010). A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer. *Nat. Genet.* 42, 415–419. doi: 10.1038/ng.558

Kiemeney, L. A., Thorlacius, S., Sulem, P., Geller, F., Aben, K. K. H., Stacey, S. N., et al. (2008). Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.* 40, 1307–1312. doi: 10.1038/ng.229

Kikuchi, N., Fuku, N., Matsumoto, R., Matsumoto, S., Murakami, H., Miyachi, M., et al. (2017a). The association between MCT1 T1470A polymorphism and power-oriented athletic performance. *Int. J. Sports Med.* 38, 76–80. doi: 10.1055/s-0042-117113

Kikuchi, N., Nakazato, K., Min, S.-K., Ueda, D., and Igawa, S. (2014). The ACTN3 R577X polymorphism is associated with muscle power in male Japanese athletes. *J. Strength Cond. Res.* 28, 1783–1789. doi: 10.1519/JSC.0000000000000338

Kikuchi, N., Yoshida, S., Min, S.-K., Lee, K., Sakamaki-Sunaga, M., Okamoto, T., et al. (2015). The ACTN3 R577X genotype is associated with muscle function in a Japanese population. *Appl. Physiol. Nutr. Metab.* 40, 316–322. doi: 10.1139/apnm-2014-0346

Kikuchi, N., Zempo, H., Fuku, N., Murakami, H., Sakamaki-Sunaga, M., Okamoto, T., et al. (2017b). Association between ACTN3 R577X polymorphism and trunk flexibility in 2 different cohorts. *Int. J. Sports Med.* 38, 402–406. doi: 10.1055/s-0042-118649

Knoll, N., Volckmar, A.-L., Pütter, C., Scherag, A., Kleber, M., Hebebrand, J., et al. (2012). The fatty acid amide hydrolase (FAAH) gene variant rs324420 AA/AC is not associated with weight loss in a 1-year lifestyle intervention for obese children and adolescents. *Horm. Metab. Res.* 44, 75–77. doi: 10.1055/s-0031-1291306

Kobylecki, C. J., Afzal, S., Davey Smith, G., and Nordestgaard, B. G. (2015). Genetically high plasma vitamin C, intake of fruit and vegetables, and risk of ischemic heart disease and all-cause mortality: a Mendelian randomization study. *Am. J. Clin. Nutr.* 101, 1135–1143. doi: 10.3945/ajcn.114.104497

Kobylecki, C. J., Afzal, S., and Nordestgaard, B. G. (2018). Genetically high plasma vitamin C and urate: a Mendelian randomization study in 106 147 individuals from the general population. *Rheumatology* 57, 1769–1776. doi: 10.1093/rheumatology/key171

Kranzler, H. R., Zhou, H., Kember, R. L., Vickers Smith, R., Justice, A. C., Damrauer, S., et al. (2019). Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* 10:1499. doi: 10.1038/s41467-019-09480-8

Laake, K., Jansen, L., Hahnemann, J. M., Brondum-Nielsen, K., Lönnqvist, T., Kääriäinen, H., et al. (2000). Characterization of ATM mutations in 41 Nordic families with ataxia telangiectasia. *Hum. Mutat.* 16, 232–246. doi: 10.1002/1098-1004(200009)16:3<232::AID-HUMU6>3.0.CO;2-L

Lai, D., Wetherill, L., Bertelsen, S., Carey, C. E., Kamarajan, C., Kapoor, M., et al. (2019). Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes Brain Behav.* 18:e12579. doi: 10.1111/gbb.12579

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi: 10.1093/nar/gkt1113

Lauria, F., Siani, A., Bammann, K., Foraita, R., Huybrechts, I., Iacoviello, L., et al. (2012). Prospective analysis of the association of a common variant of FTO (rs9939609) with adiposity in children: results of the IDEFICS study. *PLoS ONE* 7:e48876. doi: 10.1371/journal.pone.0048876

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Liew, S.-C., and Gupta, E. D. (2015). Methylenetetrahydrofolate reductase (MTHFR) C677T polymorphism: epidemiology, metabolism and the associated diseases. *Eur. J. Med. Genet.* 58, 1–10. doi: 10.1016/j.ejmg.2014.10.004

Liu, C.-C., Liu, C.-C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat. Rev. Neurol.* 9, 106–118. doi: 10.1038/nrneurol.2012.263

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359

Livingstone, K. M., Celis-Morales, C., Papandonatos, G. D., Erar, B., Florez, J. C., Jablonski, K. A., et al. (2016). FTO genotype and weight loss: systematic review and meta-analysis of 9563 individual participant data from eight randomised controlled trials. *BMJ* 354:i4707. doi: 10.1136/bmj.i4707

Lopez-Leon, S., Tuvblad, C., and Forero, D. A. (2016). Sports genetics: the PPARA gene and athletes' high ability in endurance sports. A systematic review and meta-analysis. *Biol. Sport* 33, 3–6. doi: 10.5604/20831862.1180170

López-Ortiz, M. M., Garay-Sevilla, M. E., Tejero, M. E., and Perez-Luque, E. L. (2016). Analysis of the interaction between transcription factor 7-like 2 genetic variants with nopal and wholegrain fibre intake: effects on anthropometric and metabolic characteristics in type 2 diabetes patients. *Br. J. Nutr.* 116, 969–978. doi: 10.1017/S0007114516002798

Machini, K., Ceyhan-Birsoy, O., Azzariti, D. R., Sharma, H., Rossetti, P., Mahanta, L., et al. (2019). Analyzing and reanalyzing the genome: findings from the MedSeq Project. *Am. J. Hum. Genet.* 105, 177–188. doi: 10.1016/j.ajhg.2019.05.017

Madden, K., Feldman, H. A., Chun, R. F., Smith, E. M., Sullivan, R. M., Agan, A. A., et al. (2015). Critically ill children have low vitamin D-binding protein, influencing bioavailability of vitamin D. *Ann. Am. Thorac. Soc.* 12, 1654–1661. doi: 10.1513/AnnalsATS.201503-160OC

Mahmutovic, L., Bego, T., Sterner, M., Gremsperger, G., Ahlqvist, E., Velija Asimi, Z., et al. (2019). Association of IRS1 genetic variants with glucose control and insulin resistance in type 2 diabetic patients from Bosnia and Herzegovina. *Drug Metab Pers Ther* 34. doi: 10.1515/dmpt-2018-0031

Major, J. M., Yu, K., Chung, C. C., Weinstein, S. J., Yeager, M., Wheeler, W., et al. (2012). Genome-wide association study identifies three common variants associated with serologic response to vitamin E supplementation in men. *J. Nutr.* 142, 866–871. doi: 10.3945/jn.111.156349

Major, J. M., Yu, K., Weinstein, S. J., Berndt, S. I., Hyland, P. L., Yeager, M., et al. (2014). Genetic variants reflecting higher vitamin e status in men are associated with reduced risk of prostate cancer. *J. Nutr.* 144, 729–733. doi: 10.3945/jn.113.189928

Major, J. M., Yu, K., Wheeler, W., Zhang, H., Cornelis, M. C., Wright, M. E., et al. (2011). Genome-wide association study identifies common variants associated with circulating vitamin E levels. *Hum. Mol. Genet.* 20, 3876–3883. doi: 10.1093/hmg/ddr296

Mangge, H., Renner, W., Almer, G., Weghuber, D., Möller, R., and Horejsi, R. (2011). Rs9939609 variant of the fat mass and obesity-associated gene and trunk obesity in adolescents. *J. Obes.* 2011:186368. doi: 10.1155/2011/186368

Martínez, C., Galván, S., Garcia-Martin, E., Ramos, M. I., Gutiérrez-Martín, Y., and Agúndez, J. A. G. (2010). Variability in ethanol biodisposition in whites is modulated by polymorphisms in the ADH1B and ADH1C genes. *Hepatology* 51, 491–500. doi: 10.1002/hep.23341

Martins, I. J., Hone, E., Foster, J. K., Sünram-Lea, S. I., Gnjec, A., Fuller, S. J., et al. (2006). Apolipoprotein E, cholesterol metabolism, diabetes, and the convergence of risk factors for Alzheimer's disease and cardiovascular disease. *Mol. Psychiatry* 11, 721–736. doi: 10.1038/sj.mp.4001854

März, W., Nauck, M. S., Fisher, E., Hoffmann, M. M., and Wieland, H. (2000). "The molecular mechanisms of inherited hypercholesterolemia," in *From Molecule to Men*, eds M. Zehender, H. Just, and G. Breithardt (Heidelberg: Steinkopff), 151–169. doi: 10.1007/978-3-642-57724-6_13

Masaoka, H., Ito, H., Gallus, S., Watanabe, M., Yokomizo, A., Eto, M., et al. (2017). Combination of ALDH2 and ADH1B polymorphisms is associated with smoking initiation: a large-scale cross-sectional study in a Japanese population. *Drug Alcohol Depend.* 173, 85–91. doi: 10.1016/j.drugalcdep.2016.12.015

Massidda, M., Voisin, S., Culigioni, C., Piras, F., Cugia, P., Yan, X., et al. (2019). ACTN3 R577X polymorphism is associated with the incidence and severity of injuries in professional football players. *Clin. J. Sport Med.* 29, 57–61. doi: 10.1097/JSM.0000000000000487

Matsuda, K., Takahashi, A., Middlebrooks, C. D., Obara, W., Nasu, Y., Inoue, K., et al. (2015). Genome-wide association study identified SNP on 15q24 associated with bladder cancer risk in Japanese population. *Hum. Mol. Genet.* 24, 1177–1184. doi: 10.1093/hmg/ddu512

Mattar, R., de Campos Mazo, D. F., and Carrilho, F. J. (2012). Lactose intolerance: diagnosis, genetic, and clinical factors. *Clin. Exp. Gastroenterol.* 5, 113–121. doi: 10.2147/CEG.S32368

Mbarek, H., Milaneschi, Y., Fedko, I. O., Hottenga, J.-J., de Moor, M. H. M., Jansen, R., et al. (2015). The genetics of alcohol dependence: Twin and SNP-based heritability, and genome-wide association study based on AUDIT scores. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 168, 739–748. doi: 10.1002/ajmg.b.32379

McLaren, C. E., Garner, C. P., Constantine, C. C., McLachlan, S., Vulpe, C. D., Snively, B. M., et al. (2011). Genome-wide association study identifies genetic loci associated with iron deficiency. *PLoS ONE* 6:e17390. doi: 10.1371/journal.pone.0017390

Mehramiz, M., Khayyatzadeh, S. S., Esmaily, H., Ghasemi, F., Sadeghi-Ardekani, K., Tayefi, M., et al. (2019). Associations of vitamin D binding protein variants with the vitamin D-induced increase in serum 25-hydroxyvitamin D. *Clin. Nutr.* 29, 59–64. doi: 10.1016/j.clnesp.2018.12.005

Meisner, A., Kundu, P., and Chatterjee, N. (2019). Case-only analysis of gene-environment interactions using polygenic risk scores. *Am. J. Epidemiol.* 188, 2013–2020. doi: 10.1093/aje/kwz175

Meng, X.-R., Song, J.-Y., Ma, J., Liu, F.-H., Shang, X.-R., Guo, X.-J., et al. (2014). Association study of childhood obesity with eight genetic variants recently identified by genome-wide association studies. *Pediatr. Res.* 76, 310–315. doi: 10.1038/pr.2014.88

Min, S.-K., Lim, S.-T., and Kim, C.-S. (2016). Association of ACTN3 polymorphisms with BMD, and physical fitness of elderly women. *J. Phys. Ther. Sci.* 28, 2731–2736. doi: 10.1589/jpts.28.2731

Moran, N. E., Thomas-Ahner, J. M., Fleming, J. L., McElroy, J. P., Mehl, R., Grainger, E. M., et al. (2019). Single nucleotide polymorphisms in β-carotene oxygenase 1 are associated with plasma lycopene responses to a tomato-soy juice intervention in men with prostate cancer. *J. Nutr.* 149, 381–397. doi: 10.1093/jn/nxy304

Mühleisen, T. W., Leber, M., Schulze, T. G., Strohmaier, J., Degenhardt, F., Treutlein, J., et al. (2014). Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* 5:3339. doi: 10.1038/ncomms4339

Muñoz, A., López-Samanes, Á., Aguilar-Navarro, M., Varillas-Delgado, D., Rivilla-García, J., Moreno-Pérez, V., et al. (2020). Effects of and genotypes on the ergogenic response to caffeine in professional handball players. *Genes* 11:933. doi: 10.3390/genes11080933

Murtagh, C. F., Brownlee, T. E., Rienzi, E., Roquero, S., Moreno, S., Huertas, G., et al. (2020). The genetic profile of elite youth soccer players and its association with power and speed depends on maturity status. *PLoS ONE* 15:e0234458. doi: 10.1371/journal.pone.0234458

Nievergelt, C. M., Maihofer, A. X., Mustapic, M., Yurgil, K. A., Schork, N. J., Miller, M. W., et al. (2015). Genomic predictors of combat stress vulnerability and resilience in U.S. Marines: a genome-wide association study across multiple ancestries implicates PRTFDC1 as a potential PTSD gene. *Psychoneuroendocrinology* 51, 459–471. doi: 10.1016/j.psyneuen.2014.10.017

Nissen, J., Rasmussen, L. B., Ravn-Haren, G., Andersen, E. W., Hansen, B., Andersen, R., et al. (2014). Common variants in CYP2R1 and GC genes predict vitamin D concentrations in healthy Danish children and adults. *PLoS ONE* 9:e89907. doi: 10.1371/journal.pone.0089907

Nissen, J., Vogel, U., Ravn-Haren, G., Andersen, E. W., Madsen, K. H., Nex,ø, B. A., et al. (2015). Common variants in CYP2R1 and GC genes are both determinants of serum 25-hydroxyvitamin D concentrations after UVB irradiation and after consumption of vitamin D3-fortified bread and milk during winter in Denmark. *Am. J. Clin. Nutr.* 101, 218–227. doi: 10.3945/ajcn.114.092148

Nofziger, C., Turner, A. J., Sangkuhl, K., Whirl-Carrillo, M., Agúndez, J. A. G., Black, J. L., et al. (2020). PharmVar GeneFocus: CYP2D6. *Clin. Pharmacol. Ther.* 107, 154–170. doi: 10.1002/cpt.1643

Nongmaithem, S. S., Joglekar, C. V., Krishnaveni, G. V., Sahariah, S. A., Ahmad, M., Ramachandran, S., et al. (2017). GWAS identifies population-specific new regulatory variants in FUT6 associated with plasma B12 concentrations in Indians. *Hum. Mol. Genet.* 26, 2551–2564. doi: 10.1093/hmg/ddx071

Noorshahi, N., Sotoudeh, G., Djalali, M., Eshraghian, M. R., Keramatipour, M., Basiri, M. G., et al. (2016). APOA II genotypes frequency and their interaction with saturated fatty acids consumption on lipid profile of patients with type 2 diabetes. *Clin. Nutr.* 35, 907–911. doi: 10.1016/j.clnu.2015.06.008

Nykamp, K., The Invitae Clinical Genomics, G.roup, Anderson, M., Powers, M., Garcia, J., Herrera, B., et al. (2017). Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet. Med.* 19, 1105–1117. doi: 10.1038/gim.2017.37

O'Connell, K., Posthumus, M., Schwellnus, M. P., and Collins, M. (2013). Collagen genes and exercise-associated muscle cramping. *Clin. J. Sport Med.* 23, 64–69. doi: 10.1097/JSM.0b013e3182686aa7

Oki, E., Norde, M. N., Carioca, A. A. F., Souza, J. M. P., Castro, I. A., Marchioni, D. M. L., et al. (2017). Polymorphisms of the TNF-α gene interact with plasma fatty acids on inflammatory biomarker profile: a population-based, cross-sectional study in São Paulo, Brazil. *Br. J. Nutr.* 117, 1663–1673. doi: 10.1017/S0007114517001416

Olfson, E., and Bierut, L. J. (2012). Convergence of genome-wide association and candidate gene studies for alcoholism. *Alcohol. Clin. Exp. Res.* 36, 2086–2094. doi: 10.1111/j.1530-0277.2012.01843.x

Palmer, J. R. (2019). Polygenic risk scores for breast cancer risk prediction: lessons learned and future opportunities. *J. Natl. Cancer Inst.* 112, 555–556. doi: 10.1093/jnci/djz176

Pataky, M. W., Womack, C. J., Saunders, M. J., Goffe, J. L., D'Lugos, A. C., El-Sohemy, A., et al. (2016). Caffeine and 3-km cycling performance: effects of mouth rinsing, genotype, and time of day. *Scand. J. Med. Sci. Sports* 26, 613–619. doi: 10.1111/sms.12501

Pavan, S., Rommel, K., Mateo Marquina, M. E., Höhn, S., Lanneau, V., and Rath, A. (2017). Clinical practice guidelines for rare diseases: the orphanet database. *PLoS ONE* 12:e0170365. doi: 10.1371/journal.pone.0170365

Pekkinen, M., Saarnio, E., Viljakainen, H. T., Kokkonen, E., Jakobsen, J., Cashman, K., et al. (2014). Vitamin D binding protein genotype is associated with serum 25-hydroxyvitamin D and PTH concentrations, as well as bone health in children and adolescents in Finland. *PLoS ONE* 9:e87292. doi: 10.1371/journal.pone.0087292

Peplonska, B., Adamczyk, J. G., Siewierski, M., Safranow, K., Maruszak, A., Sozanski, H., et al. (2017). Genetic variants associated with physical and mental characteristics of the elite athletes in the Polish population. *Scand. J. Med. Sci. Sports* 27, 788–800. doi: 10.1111/sms.12687

Petersen, R. A., Larsen, L. H., Damsgaard, C. T., Sørensen, L. B., Hjorth, M. F., Andersen, R., et al. (2017). Common genetic variants are associated with lower serum 25-hydroxyvitamin D concentrations across the year among children at northern latitudes. *Br. J. Nutr.* 117, 829–838. doi: 10.1017/S0007114517000538

Petr, M., Maciejewska-Skrendo, A., Zajac, A., Chycki, J., and Stastny, P. (2019). Association of elite sports status with gene variants of peroxisome proliferator activated receptors and their transcriptional coactivator. *Int. J. Mol. Sci.* 21:162. doi: 10.3390/ijms21010162

Petr, M., Stastny, P., Zajac, A., Tufano, J. J., and Maciejewska-Skrendo, A. (2018). The role of peroxisome proliferator-activated receptors and their transcriptional coactivators gene variations in human trainability: a systematic review. *Int. J. Mol. Sci.* 19:472. doi: 10.3390/ijms19051472

PGP-UK Consortium (2018). Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. *BMC Med. Genomics* 11:108. doi: 10.1186/s12920-018-0423-1

Pirastu, N., Joshi, P. K., de Vries, P. S., Cornelis, M. C., McKeigue, P. M., Keum, N., et al. (2017). GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat. Commun.* 8:1584. doi: 10.1038/s41467-017-01490-8

Płóciennik, Ł. A., Zaucha, J., Zaucha, J. M., Łukaszuk, K., and Józwicki, M., Płóciennik, M., et al. (2020). Detection of epistasis between ACTN3 and SNAP-25 with an insight towards gymnastic aptitude identification. *PLoS ONE* 15: e0237808. doi: 10.1371/journal.pone.0237808

Polimanti, R., and Gelernter, J. (2018). ADH1B: From alcoholism, natural selection, and cancer to the human phenome. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 177, 113–125. doi: 10.1002/ajmg.b.32523

Psychiatric, G. W. A. S., and Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* 43, 977–983. doi: 10.1038/ng.943

Puente, C., Abián-Vicén, J., Del Coso, J., Lara, B., and Salinero, J. J. (2018). The CYP1A2−163C>A polymorphism does not alter the effects of caffeine on basketball performance. *PLoS ONE* 13:e0195943. doi: 10.1371/journal.pone.0195943

Qi, Q., Downer, M. K., Kilpeläinen, T. O., Taal, H. R., Barton, S. J., Ntalla, I., et al. (2015a). Dietary intake, FTO genetic variants, and adiposity: a combined analysis of over 16,000 children and adolescents. *Diabetes* 64, 2467–2476. doi: 10.2337/db14-1629

Qi, Q., Zheng, Y., Huang, T., Rood, J., Bray, G. A., Sacks, F. M., et al. (2015b). Vitamin D metabolism-related genetic variants, dietary protein intake and improvement of insulin resistance in a 2 year weight-loss trial: POUNDS Lost. *Diabetologia* 58, 2791–2799. doi: 10.1007/s00125-015-3750-1

Quan, L.-L., Wang, H., Tian, Y., Mu, X., Zhang, Y., and Tao, K. (2015). Association of fat-mass and obesity-associated gene FTO rs9939609 polymorphism with the risk of obesity among children and adolescents: a meta-analysis. *Eur. Rev. Med. Pharmacol. Sci.* 19, 614–623.

Rafnar, T., Sulem, P., Thorleifsson, G., Vermeulen, S. H., Helgason, H., Saemundsdottir, J., et al. (2014). Genome-wide association study yields variants at 20p12.2 that associate with urinary bladder cancer. *Hum. Mol. Genet.* 23, 5545–5557. doi: 10.1093/hmg/ddu264

Rafnar, T., Vermeulen, S. H., Sulem, P., Thorleifsson, G., Aben, K. K., Witjes, J. A., et al. (2011). European genome-wide association study identifies SLC14A1 as a new urinary bladder cancer susceptibility gene. *Hum. Mol. Genet.* 20, 4268–4281. doi: 10.1093/hmg/ddr303

Rahimi, M. H., Mollahosseini, M., Mirzababaei, A., Yekaninejad, M. S., Maghbooli, Z., and Mirzaei, K. (2019). Interactions between vitamin D binding protein variants and major dietary patterns on the odds of metabolic syndrome and its components in apparently healthy adults. *Diabetol. Metab. Syndr.* 11:28. doi: 10.1186/s13098-019-0422-1

Ramírez, J., van Duijvenboden, S., Ntalla, I., Mifsud, B., Warren, H. R., Tzanis, E., et al. (2018). Thirty loci identified for heart rate response to exercise and recovery implicate autonomic nervous system. *Nat. Commun.* 9:1947. doi: 10.1038/s41467-018-04148-1

Ranzenhofer, L. M., Mayer, L. E. S., Davis, H. A., Mielke-Maday, H. K., McInerney, H., Korn, R., et al. (2019). The FTO gene and measured food intake in 5- to 10-year-old children without obesity. *Obesity* 27, 1023–1029. doi: 10.1002/oby.22464

Rasmussen, K. L., Tybjærg-Hansen, A., Nordestgaard, B. G., and Frikke-Schmidt, R. (2018). Absolute 10-year risk of dementia by age, sex and genotype: a population-based cohort study. *CMAJ* 190, E1033–E1041. doi: 10.1503/cmaj.180066

Ravindran, R. D., Sundaresan, P., Krishnan, T., Vashist, P., Maraini, G., Saravanan, V., et al. (2019). Genetic variants in a sodium-dependent vitamin C transporter gene and age-related cataract. *Br. J. Ophthalmol.* 103, 1223–1227. doi: 10.1136/bjophthalmol-2018-312257

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi: 10.1093/nar/gky1016

Rice, T. K., Sarzynski, M. A., Sung, Y. J., Argyropoulos, G., Stütz, A. M., Teran-Garcia, M., et al. (2012). Fine mapping of a QTL on chromosome 13 for submaximal exercise capacity training response: the HERITAGE Family Study. *Eur. J. Appl. Physiol.* 112, 2969–2978. doi: 10.1007/s00421-011-2274-8

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Rivera, M. A., and Fahey, T. D. (2019). Association between aquaporin-1 and endurance performance: a systematic review. *Sports Med Open* 5:40. doi: 10.1186/s40798-019-0213-0

Rivera, M. A., Fahey, T. D., López-Taylor, J. R., and Martínez, J. L. (2020). The association of aquaporin-1 gene with marathon running performance level: a confirmatory study conducted in male hispanic marathon runners. *Sports Med. Open* 6:16. doi: 10.1186/s40798-020-00243-0

Roach, J. C., Glusman, G., Hubley, R., Montsaroff, S. Z., Holloway, A. K., Mauldin, D. E., et al. (2011). Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* 89, 382–397. doi: 10.1016/j.ajhg.2011.07.023

Robien, K., Butler, L. M., Wang, R., Beckman, K. B., Walek, D., Koh, W.-P., et al. (2013). Genetic and environmental predictors of serum 25-hydroxyvitamin D concentrations among middle-aged and elderly Chinese in Singapore. *Br. J. Nutr.* 109, 493–502. doi: 10.1017/S0007114512001675

Ronald, A., Happ,é, F., Price, T. S., Baron-Cohen, S., and Plomin, R. (2006). Phenotypic and genetic overlap between autistic traits at the extremes of the general population. *J. Am. Acad. Child Adolesc. Psychiatry* 45, 1206–1214. doi: 10.1097/01.chi.0000230165.54117.41

Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J. D., et al. (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* 42, 978–984. doi: 10.1038/ng.687

Salinero, J. J., Lara, B., Ruiz-Vicente, D., Areces, F., Puente-Torres, C., Gallo-Salazar, C., et al. (2017). CYP1A2 genotype variations do not modify the benefits and drawbacks of caffeine during exercise: a pilot study. *Nutrients* 9:269. doi: 10.3390/nu9030269

Santiago, C., Ruiz, J. R., Buxens, A., Artieda, M., Arteta, D., González-Freire, M., et al. (2011). Trp64Arg polymorphism in ADRB3 gene is associated with elite endurance performance. *Br. J. Sports Med.* 45, 147–149. doi: 10.1136/bjsm.2009.061366

Saunders, C. J., Posthumus, M., O'Connell, K., September, A. V., and Collins, M. (2015). A variant within the AQP1 3'-untranslated region is associated with running performance, but not weight changes, during an Ironman Triathlon. *J. Sports Sci.* 33, 1342–1348. doi: 10.1080/02640414.2014.989535

Schadock, I., Schneider, A., Silva, E. D., Buchweitz, M. R. D., Correa, M. N., Pesquero, J. B., et al. (2015). Simple method to genotype the ACTN3 r577x polymorphism. *Genet. Test. Mol. Biomark.* 19, 253–257. doi: 10.1089/gtmb.2014.0299

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi: 10.1038/nature13595

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., et al. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 50, 928–936. doi: 10.1038/s41588-018-0142-8

Sen, A., and Stark, H. (2019). Role of cytochrome P450 polymorphisms and functions in development of ulcerative colitis. *World J. Gastroenterol.* 25, 2846–2862. doi: 10.3748/wjg.v25.i23.2846

Slattery, M. L., Wolff, R. K., Herrick, J. S., Caan, B. J., and Samowitz, W. (2010). Calcium, vitamin D, VDR genotypes, and epigenetic and genetic changes in rectal tumors. *Nutr. Cancer* 62, 436–442. doi: 10.1080/01635580903441204

Smith, C. E., Tucker, K. L., Arnett, D. K., Noel, S. E., Corella, D., Borecki, I. B., et al. (2013). Apolipoprotein A2 polymorphism interacts with intakes of dairy foods to influence body weight in 2 U.S. populations. *J. Nutr.* 143, 1865–1871. doi: 10.3945/jn.113.179051

Smith, D. M., Weitzel, K. W., Elsey, A. R., Langaee, T., Gong, Y., Wake, D. T., et al. (2019). CYP2D6-guided opioid therapy improves pain control in CYP2D6 intermediate and poor metabolizers: a pragmatic clinical trial. *Genet. Med.* 21, 1842–1850. doi: 10.1038/s41436-018-0431-8

Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W., et al. (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol. Psychiatry* 14, 755–763. doi: 10.1038/mp.2009.43

Søberg, S., Sandholt, C. H., Jespersen, N. Z., Toft, U., Madsen, A. L., von Holstein-Rathlou, S., et al. (2017). FGF21 is a sugar-induced hormone associated with sweet intake and preference in humans. *Cell Metab.* 25, 1045–1053.e6. doi: 10.1016/j.cmet.2017.04.009

Sonestedt, E., Roos, C., Gullberg, B., Ericson, U., Wirfält, E., and Orho-Melander, M. (2009). Fat and carbohydrate intake modify the association between genetic variation in the FTO genotype and obesity. *Am. J. Clin. Nutr.* 90, 1418–1425. doi: 10.3945/ajcn.2009.27958

Stein, M. B., Chen, C.-Y., Ursano, R. J., Cai, T., Gelernter, J., Heeringa, S. G., et al. (2016). Genome-wide association studies of posttraumatic stress disorder in 2 cohorts of US army soldiers. *JAMA Psychiatry* 73, 695–704. doi: 10.1001/jamapsychiatry.2016.0350

Sun, Y., Chang, S., Wang, F., Sun, H., Ni, Z., Yue, W., et al. (2019). Genome-wide association study of alcohol dependence in male Han Chinese and cross-ethnic polygenic risk score comparison. *Transl. Psychiatry* 9, 249. doi: 10.1038/s41398-019-0586-3

Szentkereszty-Kovács, Z., Fiatal, S., Szegedi, A., Kovács, D., Janka, E., Herszényi, K., et al. (2019). The prevalence of ADH1B and OPRM1 alleles predisposing for alcohol consumption are increased in the Hungarian psoriasis population. *Arch. Dermatol. Res.* 311, 435–442. doi: 10.1007/s00403-019-01915-y

Takahashi, M., Kahnoski, R., Gross, D., Nicol, D., and Teh, B. T. (2002). Familial adult renal neoplasia. *J. Med. Genet.* 39, 1–5. doi: 10.1136/jmg.39.1.1

Tanaka, T., Scheet, P., Giusti, B., Bandinelli, S., Piras, M. G., Usala, G., et al. (2009). Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. *Am. J. Hum. Genet.* 84, 477–482. doi: 10.1016/j.ajhg.2009.02.011

Tanofsky-Kraff, M., Han, J. C., Anandalingam, K., Shomaker, L. B., Columbo, K. M., Wolkoff, L. E., et al. (2009). The FTO gene rs9939609 obesity-risk allele and loss of control over eating. *Am. J. Clin. Nutr.* 90, 1483–1488. doi: 10.3945/ajcn.2009.28439

Tanwar, V. S., Chand, M. P., Kumar, J., Garg, G., Seth, S., Karthikeyan, G., et al. (2013). Common variant in FUT2 gene is associated with levels of vitamin B(12) in Indian population. *Gene* 515, 224–228. doi: 10.1016/j.gene.2012.11.021

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

Thompson, A., Cook, J., Choquet, H., Jorgenson, E., Yin, J., Kinnunen, T., et al. (2020). Functional validity, role, and implications of heavy alcohol consumption genetic loci. *Sci. Adv.* 6:eaay5034. doi: 10.1126/sciadv.aay5034

Timpson, N. J., Forouhi, N. G., Brion, M.-J., Harbord, R. M., Cook, D. G., Johnson, P., et al. (2010). Genetic variation at the SLC23A1 locus is associated with circulating concentrations of L-ascorbic acid (vitamin C): evidence from 5 independent studies with >15,000 participants. *Am. J. Clin. Nutr.* 92, 375–382. doi: 10.3945/ajcn.2010.29438

Todendi, P. F., Martínez, J. A., Reuter, C. P., Klinger, E. I., Fiegenbaum, M., and Rosane de Moura Valim, A. (2020). Influence of FTO (Fat mass and obesity) gene and parental obesity on Brazilian children and adolescents adiposity. *J. Pediatr. Endocrinol. Metab.* 33, 975–982. doi: 10.1515/jpem-2019-0594

Torkamani, A., and Topol, E. (2019). Polygenic risk scores expand to obesity. *Cell* 177, 518–520. doi: 10.1016/j.cell.2019.03.051

Touvier, M., Deschasaux, M., Montourcy, M., Sutton, A., Charnaux, N., Kesse-Guyot, E., et al. (2015). Determinants of vitamin D status in Caucasian adults: influence of sun exposure, dietary intake, sociodemographic, lifestyle, anthropometric, and genetic factors. *J. Invest. Dermatol.* 135, 378–388. doi: 10.1038/jid.2014.400

Uhlig, H. H., and Muise, A. M. (2017). Clinical genomics in inflammatory bowel disease. *Trends Genet.* 33, 629–641. doi: 10.1016/j.tig.2017.06.008

van Heel, D. A., Franke, L., Hunt, K. A., Gwilliam, R., Zhernakova, A., Inouye, M., et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* 39, 827–829. doi: 10.1038/ng2058

Vu, T. T., and Gooderham, M. (2017). Adverse drug reactions and cutaneous manifestations associated with anticoagulation. *J. Cutan. Med. Surg.* 21, 540–550. doi: 10.1177/1203475417716364

Wade, K. H., Forouhi, N. G., Cook, D. G., Johnson, P., McConnachie, A., Morris, R. W., et al. (2015). Variation in the SLC23A1 gene does not influence cardiometabolic outcomes to the extent expected given its association with L-ascorbic acid. *Am. J. Clin. Nutr.* 101, 202–209. doi: 10.3945/ajcn.114.092981

Walters, R. K., Polimanti, R., Johnson, E. C., McClintick, J. N., Adams, M. J., Adkins, A. E., et al. (2018). Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* 21, 1656–1669. doi: 10.1038/s41593-018-0275-1

Wang, B., Yang, L.-P., Zhang, X.-Z., Huang, S.-Q., Bartlam, M., and Zhou, S.-F. (2009). New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme. *Drug Metab. Rev.* 41, 573–643. doi: 10.1080/03602530903118729

Wang, M., Li, Z., Chu, H., Lv, Q., Ye, D., Ding, Q., et al. (2016). Genome-wide association study of bladder cancer in a chinese cohort reveals a new susceptibility locus at 5q12.3. *Cancer Res.* 76, 3277–3284. doi: 10.1158/0008-5472.CAN-15-2564

Wang, T., and Xu, L. (2019). Circulating vitamin E levels and risk of coronary artery disease and myocardial infarction: a mendelian randomization study. *Nutrients* 11:153. doi: 10.3390/nu11092153

Wang, Y., Tang, Y., Ji, Y., Xu, W., Ullah, N., Yu, H., et al. (2020). Association between rs174547 and levels of long-chain polyunsaturated fatty acids: a meta-analysis. *Br. J. Nutr.* doi: 10.1017/S0007114520005103. [Epub ahead of print].

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283. doi: 10.1038/ng.2797

Williams, C. J., Williams, M. G., Eynon, N., Ashton, K. J., Little, J. P., Wisloff, U., et al. (2017). Genes to predict VO trainability: a systematic review. *BMC Genomics* 18:831. doi: 10.1186/s12864-017-4192-6

Wilman, H. R., Parisinos, C. A., Atabaki-Pasdar, N., Kelly, M., Thomas, E. L., Neubauer, S., et al. (2019). Genetic studies of abdominal MRI data identify genes regulating hepcidin as major determinants of liver iron concentration. *J. Hepatol.* 71, 594–602. doi: 10.1016/j.jhep.2019.05.032

Wolf, J. M., Simon, D., Béria, J. U., Tietzmann, D. C., Stein, A. T., and Lunge, V. R. (2017). Analysis of the association of nonsynonymous polymorphisms in ADH genes with hazardous drinking in HIV-1-positive individuals. *Alcohol. Clin. Exp. Res.* 41, 1866–1874. doi: 10.1111/acer.13486

Wright, C. F., FitzPatrick, D. R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* 19, 253–268. doi: 10.1038/nrg.2017.116

Wu, X., Liu, Q., and Jiang, R. (2009a). Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* 25, 98–104. doi: 10.1093/bioinformatics/btn593

Wu, X., Ye, Y., Kiemeney, L. A., Sulem, P., Rafnar, T., Matullo, G., et al. (2009b). Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.* 41, 991–995. doi: 10.1038/ng.421

Yabuta, S., Urata, M., Wai Kun, R. Y., Masaki, M., and Shidoji, Y. (2016). Common SNP rs6564851 in the BCO1 gene affects the circulating levels of β-carotene and the daily intake of carotenoids in healthy Japanese women. *PLoS ONE* 11:e0168857. doi: 10.1371/journal.pone.0168857

Yang, M. M., Wang, J., Fan, J. J., Ng, T. K., Sun, D. J., Guo, X., et al. (2016). Variations in the obesity gene "LEPR" contribute to risk of type 2 diabetes mellitus: evidence from a meta-analysis. *J Diabetes Res* 2016:5412084. doi: 10.1155/2016/5412084

Yang, N., MacArthur, D. G., Gulbin, J. P., Hahn, A. G., Beggs, A. H., Easteal, S., et al. (2003). ACTN3 genotype is associated with human elite athletic performance. *Am. J. Hum. Genet.* 73, 627–631. doi: 10.1086/377590

Yao, P., Sun, L., Lu, L., Ding, H., Chen, X., Tang, L., et al. (2017). Effects of genetic and nongenetic factors on total and bioavailable 25(OH)D responses to vitamin D supplementation. *J. Clin. Endocrinol. Metab.* 102, 100–110. doi: 10.1210/jc.2016-2930

Yokoyama, A., Omori, T., and Yokoyama, T. (2020a). Risk factors for esophageal iodine-unstained lesions and changing trends among Japanese alcohol-dependent men (2003-2018). *Cancer Sci.* 112, 734–743. doi: 10.1111/cas.14753

Yokoyama, A., Taniki, N., Hara, S., Haysashi, E., Nakamoto, N., Mizukami, T., et al. (2018). Slow-metabolizing ADH1B and inactive heterozygous ALDH2 increase vulnerability to fatty liver in Japanese men with alcohol dependence. *J. Gastroenterol.* 53, 660–669. doi: 10.1007/s00535-017-1402-6

Yokoyama, A., Taniki, N., Nakamoto, N., Tomita, K., Hara, S., Mizukami, T., et al. (2020b). Associations among liver disease, serum lipid profile, body mass index, ketonuria, meal skipping, and the alcohol dehydrogenase-1B and aldehyde dehydrogenase-2 genotypes in Japanese men with alcohol dependence. *Hepatol. Res.* 50, 565–577. doi: 10.1111/hepr.13475

Yokoyama, A., Yokoyama, T., Matsui, T., Mizukami, T., Kimura, M., Matsushita, S., et al. (2020c). Impacts of interactions between ADH1B and ALDH2 genotypes on alcohol flushing, alcohol reeking on the day after drinking, and age distribution in Japanese alcohol-dependent men. *Pharmacogenet. Genomics* 30, 54–60. doi: 10.1097/FPC.0000000000000395

Yokoyama, A., Yokoyama, T., Omori, T., Maesato, H., Takimura, T., Iwahara, C., et al. (2019). Endoscopic screening using esophageal iodine staining and genotypes of ADH1B and ALDH2 in Japanese alcohol-dependent women. *PLoS ONE* 14:e0210546. doi: 10.1371/journal.pone.0210546

Yvert, T., Miyamoto-Mikami, E., Murakami, H., Miyachi, M., Kawahara, T., and Fuku, N. (2016). Lack of replication of associations between multiple genetic polymorphisms and endurance athlete status in Japanese population. *Physiol. Rep.* 4:e13003. doi: 10.14814/phy2.13003

Yvert, T., Santiago, C., Santana-Sosa, E., Verde, Z., Gómez-Gallego, F., López-Mojares, L. M., et al. (2015). Physical-capacity-related genetic polymorphisms in children with cystic fibrosis. *Pediatr. Exerc. Sci.* 27, 102–112. doi: 10.1123/pes.2014-0050

Zhang, M., Zhao, X., Cheng, H., Wang, L., Xi, B., Shen, Y., et al. (2014). Age- and sex-dependent association between FTO rs9939609 and obesity-related traits in Chinese children and adolescents. *PLoS ONE* 9:e97545. doi: 10.1371/journal.pone.0097545

Zhang, Q., Cao, Y., Chen, J., Shen, J., Ke, D., Wang, X., et al. (2019). ACTN3 is associated with children's physical fitness in Han Chinese. *Mol. Genet. Genomics* 294, 47–56. doi: 10.1007/s00438-018-1485-7

Zhang, Y., Tang, H.-Q., Peng, W.-J., Zhang, B.-B., and Liu, M. (2015). Meta-analysis for the association of apolipoprotein E ε2/ε3/ε4 polymorphism with coronary heart disease. *Chin. Med. J.* 128, 1391–1398. doi: 10.4103/0366-6999.156803

Zhao, J. V., and Schooling, C. M. (2017). Homocysteine-reducing B vitamins and ischemic heart disease: a separate-sample Mendelian randomization analysis. *Eur. J. Clin. Nutr.* 71, 267–273. doi: 10.1038/ejcn.2016.246

Zhao, X., Xi, B., Shen, Y., Wu, L., Hou, D., Cheng, H., et al. (2014a). An obesity genetic risk score is associated with metabolic syndrome in Chinese children. *Gene* 535, 299–302. doi: 10.1016/j.gene.2013.11.006

Zhao, X., Yang, Y., Sun, B.-F., Zhao, Y.-L., and Yang, Y.-G. (2014b). FTO and obesity: mechanisms of association. *Curr. Diab. Rep.* 14:486. doi: 10.1007/s11892-014-0486-0

Zheng, J.-S., Arnett, D. K., Parnell, L. D., Smith, C. E., Li, D., Borecki, I. B., et al. (2013). Modulation by dietary fat and carbohydrate of IRS1 association with type 2 diabetes traits in two populations of different ancestries. *Diabetes Care* 36, 2621–2627. doi: 10.2337/dc12-2607

Zhou, J.-C., Zhu, Y., Gong, C., Liang, X., Zhou, X., Xu, Y., et al. (2019). The haplotype of the vitamin D binding protein is a risk factor for a low plasma 25-hydroxyvitamin D concentration in a Han Chinese population. *Nutr. Metab.* 16:5. doi: 10.1186/s12986-019-0332-0

Zhou, T.-B., Jiang, Z.-P., Huang, M.-F., and Zhang, R. (2015). Association of vitamin D receptor gene polymorphism with the urine calcium level in nephrolithiasis patients. *J. Recept. Signal Transduct. Res.* 35, 127–132. doi: 10.3109/10799893.2014.936462

Conflict of Interest: At the time of writing, MC, KM, AM, and EL are associated with Cambridge Precision Medicine Limited. VM is a full-time employee of Fabric Genomics.

Copyright © 2021 Corpas, Megy, Mistry, Metastasio and Lehmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

POLICY AND PRACTICE REVIEWS
published: 31 March 2021
doi: 10.3389/fgene.2021.535340

Parental Access to Children's Raw Genomic Data in Canada: Legal Rights and Professional Responsibility

Michael J. S. Beauvais[1]*, Adrian M. Thorogood[2], Michael J. Szego[3,4], Karine Sénécal[5], Ma'n H. Zawati[1] and Bartha Maria Knoppers[1]

[1] Centre of Genomics and Policy, Faculty of Medicine, McGill University, Montreal, QC, Canada, [2] ELIXIR-LU, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg, [3] Centre for Clinical Ethics, Unity Health, Toronto, ON, Canada, [4] Departments of Family and Community Medicine and Molecular Genetics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada, [5] Independent Researcher, Montreal, QC, Canada

OPEN ACCESS

Edited by:
Georgia Theocharopoulou,
Ionian University, Greece

Reviewed by:
Debra Mathews,
Johns Hopkins University,
United States
Merle Weßel,
University of Oldenburg, Germany

***Correspondence:**
Michael J. S. Beauvais
michael.beauvais@mcgill.ca

Specialty section:
This article was submitted to
ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

Received: 19 February 2020
Accepted: 05 March 2021
Published: 31 March 2021

Citation:
Beauvais MJS, Thorogood AM, Szego MJ, Sénécal K, Zawati MH and Knoppers BM (2021) Parental Access to Children's Raw Genomic Data in Canada: Legal Rights and Professional Responsibility. Front. Genet. 12:535340. doi: 10.3389/fgene.2021.535340

Children with rare and common diseases now undergo whole genome sequencing (WGS) in clinical and research contexts. Parents sometimes request access to their child's raw genomic data, to pursue their own analyses or for onward sharing with health professionals and researchers. These requests raise legal, ethical, and practical issues for professionals and parents alike. The advent of widespread WGS in pediatrics occurs in a context where privacy and data protection law remains focused on giving individuals control-oriented rights with respect to their personal information. Acting in their child's stead and in their best interests, parents are generally the ones who will be exercising these informational rights on behalf of the child. In this paper, we map the contours of parental authority to access their child's raw genomic data. We consider three use cases: hospital-based researchers, healthcare professionals acting in a clinical-diagnostic capacity, and "pure" academic researchers at a public institution. Our research seeks to answer two principal questions: Do parents have a right of access to their child's raw WGS data? If so, what are the limits of this right? Primarily focused on the laws of Ontario, Canada's most populous province, with a secondary focus on Canada's three other most populous provinces (Quebec, British Columbia, and Alberta) and the European Union, our principal findings include (1) parents have a general right of access to information about their children, but that the access right is more capacious in the clinical context than in the research context; (2) the right of access extends to personal data in raw form; (3) a consideration of the best interests of the child may materially limit the legal rights of parents to access data about their child; (4) the ability to exercise rights of access are transferred from parents to children when they gain decision-making capacity in both the clinical and research contexts, but with more nuance in the former. With these findings in mind, we argue that professional guidelines, which are concerned with obligations to interpret and return results, may assist in furthering a child's best interests in the context of legal access rights. We conclude by crafting recommendations for healthcare professionals in the clinical and research contexts when faced with a parental request for a child's raw genomic data.

Keywords: whole genome sequencing, personal genomics, pediatrics, privacy, individual access, ethics, consent

INTRODUCTION

Children with rare and common diseases now undergo whole genome sequencing (WGS) in clinical and research contexts. Parents sometimes request access to their child's raw genomic data, to pursue their own analyses or for onward sharing with health professionals and researchers. These requests raise legal, ethical, and practical issues for both professionals and parents. In general, WGS provides a complete catalog of each nucleotide within an individual's genome. When analyzed with the appropriate tools and expertise, WGS potentially reveals inherited predispositions to a multiplicity of traits and disorders. WGS may also reveal information about a child's future health, which raises the issue of safeguarding the child's ethical right to an open future (Feinberg, 1980), viz. their ability to decide for themselves as adults whether or not to be tested for certain conditions. In turn, this raises the following questions: Should children be tested for adult-onset conditions? Should secondary or incidental findings from WGS be reported to children?

Parental access appears to be an increasingly pressing question for healthcare institutions, clinicians and researchers. The prevalence of parental access requests has not been well-studied, though there is some preliminary empirical evidence of the prevalence of individual access requests generally (Narayanasamy et al., 2020). Patients and caregivers at pediatric institutions occasionally ask for their raw genomic data following WGS tests. Indeed, geneticists at a large pediatric hospital in Ontario report that these requests occur (personal communication) and one of the authors of this article (MS) has been contacted by clinicians and researchers asking for guidance on how to respond to requests for raw genomic data by parents.

Parents may seek access to their child's genomic data for a number of reasons: to seek a second medical opinion about the child's condition, to inform the parents' health or reproductive choices, to share data with a health research project or repository, or to analyze the data themselves to better understand health conditions affecting their child or entire family (though, importantly, their motivation may be unknown in the context of legal access requests). We expect parental access to become more pressing in coming years as a result of three trends: (1) patients are taking a greater role in directing their care and managing their data[1], (2) sequencing of children is expanding, particularly to new clinical and newborn screening contexts[2], and (3) a growing ecosystem of third party interpretation services and data sharing platforms are emerging directed toward patients (Capaci et al., 2020; Guerrini et al., 2020).

For health professionals and researchers, parental access raises concerns that children's data will be misinterpreted or misused by parents or third party services, thus putting at risk the child's current and future health interests, development, autonomy, and privacy. Parents may pursue unnecessary analyses discouraged or prohibited by professional organizations, such as analyzing genomes for predispositions for untreatable, adult-onset conditions or predictive adult-onset disease (Borry et al., 2009; Knoppers et al., 2014; Botkin et al., 2015). This may lead to psychosocial harms for the child (e.g., anxiety, low self-esteem) or affect familial relationships (Kesserwan et al., 2016). Parents might publish a child's genome on an open-access recreational genomic database, share it with various researchers and service providers, or unintentionally allow it to be leaked (Bala, 2020). This poses potential risks of genetic discrimination by employers or insurers, and unfettered searches from law enforcement seeking to identify criminal suspects. Parental access to their child's genetic data adds a new molecular dimension to the larger policy debate over the ethics and regulation of "sharenting," where parents post photos, videos or comments about their children on social media. "Sharenting" can expose children to risks including discrimination, identity theft, reputational harm, and intimidation (Steinberg, 2016).

Previous literature has addressed issues for adults seeking access to raw WGS data. Individuals in many countries have a general right to access their health information[3] (Ries, 2010; Ogbogu et al., 2014; European Union, 2016; Guerrini et al., 2019). When WGS is adopted in clinical contexts, this right presumably extends to raw WGS data (Thorogood et al., 2018). It remains to be determined whether access to raw data extends to the research context. Many countries exempt researchers from obligations to provide participants access to their personal information (Thorogood et al., 2018). Genomics, however, often blurs clinic and research contexts, raising uncertainty as to the applicability of these exceptions (Schickhardt et al., 2020). Ethically, some commentators express concern that raw genomic data is incomprehensible to most people, offering limited benefit while presenting health and privacy risks to sequenced individuals or their family members stemming from misinterpretation or mismanagement of data (Bredenoord et al., 2011). These concerns are counteracted by principled arguments, including that such data belongs to the individual, who should be free to decide what to do with the data (Schickhardt et al., 2020), arguments of beneficence that individuals can improve

[1] Genetic Alliance Promise for Engaging Everyone Responsibly|GeneticAlliance.org. Available at: http://www.geneticalliance.org/programs/biotrust/peer (Accessed February 1, 2021).

[2] Department of Health and Social Care (England) Health minister: NHS must lead the world in genomic healthcare. GOV.UK. Available at: https://www.gov.uk/government/news/health-minister-nhs-must-lead-the-world-in-genomic-healthcare (Accessed February 15, 2020).

[3] *McInerney v.* MacDonald, 1992 CanLII 57 (SCC).

their health through sharing data with clinicians or via self-analysis, and finally, utilitarian arguments that providing access may attract research participants as well as provide them with opportunities to accelerate research by sharing their data with other research projects (Kish and Topol, 2015). There are also practical questions about who will provide individuals requesting access with interpretive support and who will pay for this support.

These legal and ethical debates have largely overlooked the rights of children themselves. The legal access rights of parents and the duties of health professionals toward children must be considered in light of their human rights. The United Nations *Convention on the Rights of the Child* (CRC) mandates that the best interests of children are to be a primary consideration in all matters concerning children (United Nations General Assembly, 2007). Parents are authorized by law to act on behalf of their children, while children have a right to be heard and a right to participate in decisions concerning their health (to the extent possible). At the same time, children have a right to appropriate guidance from their parents, amounting to a zone of deference to parental decision-making (Kamchedzera, 2012). Parental authority, however, has a fiduciary character and must be exercised in the child's best interests, not in the parents' personal interests or those of other family members (Tobin and Varadan, 2019). Health professionals are obliged not only to promote the health of children but also to protect children from parental decisions that are contrary to children's actual and future health and well-being (Schwarz et al., 2015). Medical "neglect" under child protection legislation can include both over-treatment or the failure to prevent or treat diseases in children. Parental access therefore raises new legal and ethical questions. When does parental access to their child's information support the child's best interests? When does it threaten them? And who is ultimately responsible for making this determination?

This article examines the legal rights of parents to access their child's raw WGS data generated in healthcare and health research contexts. We begin with legal questions about the scope of parental authority to request access, the rights of children themselves, and the scope of professional responsibility toward minors to justify withholding access. The analysis looks primarily at the freedom of information and health privacy laws of Ontario, Canada's most populous province and the site of much WGS, but also highlights important similarities and differences with the laws of other Canadian provinces and the European Union. While the specific results of our analysis may be largely jurisdiction-specific, we believe the structure of our analysis can be generalized. More specifically, we aim to answer the following questions:

1. Do parents have a legal right to access their child's health information upon request in clinical and research contexts?
2. Do access rights, where applicable, extend to raw genomic sequence data (e.g., BCL, FASTQ, SAM, BAM, or VCF files)?
3. Under what circumstances, if any, can a healthcare institution or researcher refuse a parental access request (e.g., to protect the child's best interests)?
4. Where a minor is sufficiently mature to understand and appreciate the consequences of access requests, does the legal right of access ultimately reside with the minor?

We then turn to contextualizing these findings within the broader, ongoing discussion in ethical and professional guidelines in pediatric genomics surrounding the reporting of secondary and incidental findings (Jarvik et al., 2014; Knoppers et al., 2014; Zawati et al., 2014; Boycott et al., 2015; Sénécal et al., 2015a; Vears et al., 2018). Professional obligations to report secondary or incidental findings to patients or participants are admittedly distinct from the legal obligation to provide an individual access on request. With secondary or incidental findings, professionals have obligations to interpret and "push" information of clinical significance to individuals. In the access context, patients have informational rights to "pull" information (e.g., raw WGS data) upon request from data custodians. While distinct, debates over reporting secondary and incidental findings suggest that health professionals and researchers responding to parental access requests are confronted with important ethical issues surrounding the child's well-being, privacy, and developing autonomy.

Ultimately, our legal analysis aims to guide health researchers, clinicians, and health-care organizations confronted with formal requests from parents to access their child's raw genomic sequence data. In contributing to a better understanding of the law, our research findings can inform access policy and communication between health professionals, parents, and children to ensure the child's health, privacy, and developing autonomy are given full consideration.

WHOLE GENOME SEQUENCING IN CLINICAL AND RESEARCH CONTEXTS

Over the past 35 years, scientists have discovered and studied genetic variants involved in monogenic diseases, resulting in the development of genetic tests for the diagnosis or prediction of monogenic diseases. Advances in molecular biology and other biotechnological advances have contributed to a rapid increase in the supply of genetic tests for hereditary diseases. Since around 2010, next-generation sequencing (NGS) technologies, of which whole genome sequencing (WGS) is a part, have been an important addition to existing genetic testing strategies (Hall et al., 2014). Introduced first in the research realm, they have also had a tremendous impact in the clinical context (Brown and Meloche, 2016; Vaxillaire and Froguel, 2016). For example, the use of WGS presents the possibility of identifying the causes of variable clinical responses among patients with the same condition (Eckford et al., 2019).

Pediatrics has seen some of the first clinical applications of genomics. Genomic sequencing allows for faster diagnosis of inherited and de novo disease and increases the likelihood of diagnosing a child with a rare disease, or of excluding, based on the knowledge at the time of the analysis, the possibility of a rare genetic disorder (Goh and Choi, 2012). Obtaining a genetic diagnosis for a child can help clinicians and families identify and anticipate future health problems (Wilson et al., 2014), while also informing the health and reproductive choices of family members (Wright et al., 2018).

In pediatric oncology in particular, WGS can inform treatment choices via the characterization of cancers and through the identification of markers relevant for drug metabolism, i.e., pharmacogenomics (Hawcutt et al., 2013). In this context, using WGS to its fullest potential involves comparing the tumor genome to the germline in order to identify cancer-specific genetic variants, implicating both somatic and germline genomic data (Bombard et al., 2013). Thus, WGS technologies help to identify novel genetic alterations contributing to oncogenesis, cancer progression and metastasis, and assist in studying tumor complexity, heterogeneity, and evolution (Shyr and Liu, 2013). The use of genomic sequencing allows for the identification of more effective personalized targeted therapies that lead to increased cure rates and decreased treatment-related morbidity and mortality for the patient affected by relapses or hard-to-treat cancers. Despite these insights, the available clinical care options nevertheless remain insufficient in the case of some pediatric cancers, especially those at advanced stages (Khan et al., 2018).

To overcome these challenges as best as possible, children with cancer may be enrolled in research study such as Terry Fox PROFYLE 2 (PRecision Oncology for Young people 2). PROFYLE 2 targets young patients with difficult-to-treat cancers by sequencing their tumors and, upon recommendation by a molecular tumor board, enrolling those patients in relevant clinical trials. Similar efforts are also underway in other countries (Chakradhar, 2018). In Canada, 17 pediatric oncology centers, in conjunction with the Children's Oncology Group (COG), conduct clinical research studies with the aim to cure and prevent childhood and adolescent cancer through scientific discovery and collaborative research. Some of these studies sequence the child's tumor with the hopes of identifying or testing targeted therapies.

In summary, an increasing number of children with rare diseases and cancer in Canada and around the world are undergoing WGS in clinical or translational research contexts. It is therefore timely to consider the legal framework governing parental requests for access to their child's raw genomic data, and the ramifications of such access for the child's health, privacy, and overall well-being.

METHODOLOGY

The principal method of research employed was doctrinal (Hutchinson, 2018). The two principal laws governing information held by either healthcare institutions[4] or public bodies[5] in Ontario were consulted. The statutes were reviewed comprehensively, with a focus on those provisions applicable to parental access to information about their child. Where relevant provisions were found, a search was conducted for related case law from the Information and Privacy Commissioner of Ontario (IPC) or Ontario Superior Courts using WestlawNext Canada and CanLII, two legal databases. IPC and court cases that were responsive to the search were read and analyzed to grasp both the meaning and application of the provisions. We also reviewed the laws that govern personal information held by public- and private-sector organizations of other populous Canadian provinces (Quebec, Alberta, and British Columbia) as well as the European Union. With regards to the former group, a case law search using the legal databases WestlawNext Canada, CanLII, and SOQUIJ was also conducted. Our goal was not to conduct a comprehensive comparative analysis, but to at least hint at the range of potential legal divergence one may expect if our research questions were posed in other jurisdictions.

Raw WGS data includes any one of the common underlying files that are generated in the sequencing process (For a more nuanced introduction to the concept of rawness with regard to genomic data, see Schickhardt et al., 2020). Specifically, this refers to either BCL (base call) files, FASTQ files, SAM (sequence alignment map) files, BAM (binary alignment map) files or VCF (variant call format) files (Evans, 2017). This definition is meant to capture sequence data that has not been subject to any interpretation beyond the data's bare representation, without prejudice to the idea that a representation *per se* implicates an interpretive process (Gitelman and Jackson, 2013). Data having undergone processes such as annotation and interpretation are excluded from this definition (Abril and Castellano, 2019). Return of results and incidental findings are furthermore excluded from this definition of raw WGS data as both such types of information are only generated through the interpretation of sequence data. Nevertheless, as the objects of professional ethical obligations, we also examined the professional guidelines of the American College of Medical Genetics, the Canadian College of Medical Genetics, and the European Society of Human Genetics to understand the relationship of access rights to professional obligations.

RESULTS

Our research revealed the following responses to our four legal research questions. First, the right to access health information in Ontario applies generally in healthcare institutions, to both clinical and research data unless the data is held "solely" for research purposes (see **Table 1**). For all intents and purposes, there is no legal right to access health information associated with research projects at academic institutions. Other provinces and countries may provide broader access rights in research contexts. Regardless, research projects may consider providing access as a matter of policy and ethics while recognizing the limits of such data. Second, access rights in Canada extend to raw WGS data. A patient's legal right of access incorporates raw data, provided their clinician would also hypothetically have access to this data. Third, parents have authority to exercise their child's right of access to health information, as long as they exercise that right *on behalf of* their child. There may be grounds for a health information custodian to refuse parental access requests that are manifestly made to serve the interests of the parent or another party and are not in the best interests of the child. The best interests, however, may not always be an effective ground for constraining parental

[4] *Personal Health Information Protection Act*, 2004, SO 2004, c 3, Schedule A.
[5] *Freedom of Information and Protection of Privacy Act*, RSO 1990, c F.31.

access requests. Parents are generally given great deference in deciding the best interests of their child (at least under privacy law), as the actions of parents are largely perceived as aligning with their child's best interests. Parental access requests do not necessarily provide sufficient information for health information custodians to assess if access will serve or undermine the child's interests. Finally, children–not their parents–have authority to request access in healthcare institutions if they are over 16 (unless it is demonstrated that the adolescent lacks the capacity to consent), or, even if under 16, if they are sufficiently mature to understand and appreciate the consequences of data access (the mature-minor exception). Where parents request access to information from adolescents, health information custodians may be required to determine if the child is capable of consenting to access before allowing parents to do so. See **Figure 1** for a flowchart with the summary of findings.

Parental Rights of Access to Information About Their Children

The question regarding whether parents have a legal right of access to their child's raw WGS data must be framed through the prism of the general law concerning individuals' abilities to access information about themselves held by others. As a surrogate for the child's interests, parents enjoy a general ability to exercise legal rights *on behalf of* their children, including informational rights[6,7]. This parental authority is, however, not unfettered, a point which will be developed later through examining how considerations for a child's welfare feature in legal analysis regarding informational rights. What is more, whether personal information is generated for a clinical or research purpose has relevance for the availability of a right to access information (*FIPPA*, s65(8.1); General, O Reg 329/04 (*PHIPA*), s24). To best elucidate the different contours of parental access rights, our legal analysis primarily concerns itself with three different contexts: hospital-based researchers, healthcare professionals acting in a clinical-diagnostic capacity, and "pure" academic researchers at a public institution.

For the three envisaged use cases, there are two relevant laws regarding a parent's potential ability to receive their child's raw WGS data in Ontario: the *Personal Health Information Protection Act* (herein "health privacy law") and the *Freedom of Information and Protection of Privacy Act* (herein "FOI law"). While there is also the federal *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5 (*PIPEDA*), it applies to only private-sector organizations engaged in commercial activities and who are not also health information custodians under the health privacy law (Office of the Privacy Commissioner of Canada, 2015). Given the wide scope of health information custodians under the health privacy law as well as the public-dominated research and clinical landscape, the situations in which *PIPEDA* applies are limited.

The health privacy law applies only to personal health information (PHI) held by particular custodians of personal health information, which may be either public or private entities. The FOI law, on the other hand, is concerned with information held by public-sector organizations generally. Accordingly, hospital-based researchers and clinicians will generally be subject to the health privacy law whereas academic researchers at universities will be subject to the FOI law.

Research Exemptions: Tempering Individual Access

Under certain circumstances, both the health privacy law and FOI law exempt information custodians from their obligations to provide an individual with access to their personal information. The health privacy law applies to healthcare institutions and so covers situations where healthcare professionals are acting in a clinical-diagnostic capacity, as well as to hospital-based research. The FOI law applies to public-sector organizations, and so covers academic institutions (see **Table 1**). Under the health privacy law, access rights apply to health information by hospitals and supporting clinical laboratories. While genetics laboratories are considered health information custodians, only those laboratories where tests are "performed to obtain information for diagnosis, prophylaxis or treatment" are within the ambit of the health privacy law[8] [*PHIPA*, s3(1)(4)(iv)]. Both laws include exemptions to the access right for research (see **Table 2**), defined as "a systematic investigation designed to develop or establish principles, facts or generalizable knowledge or any combination of them, and includes the development, testing and evaluation of research"[9](*PHIPA*, s2). Note, however, that these research exemptions do not prevent researchers from voluntarily providing access to data, provided that such access is compatible with other legal and ethical norms.

Under the FOI law, records "respecting or associated with" research are exempted from the entirety of the law[10,11]. This broad exemption is meant to protect the academic endeavor from freedom of information requests, but also exempts individual access rights. Arguably, the research exemption's justification loses its persuasiveness where only an *individual*'s health or WGS information is requested because there is no risk of swiping research results or using research data for other improper purposes (Ries, 2010). Under the health privacy law, the research exemption is narrower—health information used "*solely* for the purposes of research" [our emphasis] is excluded from the access provisions [General, O Reg 329/04 (*PHIPA*), s24(1)]. This suggests that research information also used for clinical-care purposes is subject to the individual's access right [*PHIPA*, s1(b)]. Hospital-based genomics research, especially in pediatric contexts, often has a translational component, where WGS may also be used to assist clinical decision-making (Knoppers et al., 2016; Graaf et al., 2018). Hospital-based research projects that

[6] *Freedom of Information and Protection of Privacy Act* (*FIPPA*), RSO 1990, c F.31, s66(c).

[7] *Personal Health Information Protection Act (PHIPA)*, 2004, SO 2004, c 3, Schedule A, s23.

[8] *Laboratory and Specimen Collection Centre Licensing Act*, RSO 1990, c L.1, s5.

[9] *McMaster University (Re)*, 2008 CanLII 36902 (ON IPC).

[10] *Carleton University v. Information and Privacy Commissioner of Ontario and John Doe, requester*, 2018 ONSC 3696, 2018.

[11] *McMaster University (Re)*, 2008 CanLII 36902 (ON IPC).

TABLE 1 | Sources and scope of informational access rights in Ontario.

Name of law and/or regulation	Scope/applicability of law and/or regulation	Contexts in which a parental access right exists (clinical, research, or both)	Applicability of access right to raw genomic sequence data	Doctrines that may reduce the scope of parental access right	Recognition of "mature minor" doctrine for informational rights
Personal Health Information Protection Act, 2004, SO 2004, c 3, Sch A ("health privacy law") Complemented by *Ontario Regulation* 329/04.	Private- and public-sector organizations designed as "health information custodians" **Use case:** hospital-based researchers and healthcare professionals acting in a clinical-diagnostic capacity	Clinical: yes Research: yes, but with narrow exceptions	Yes	Best interests of the child (BIC)	Yes
Freedom of Information and Protection of Privacy Act, RSO 1990, c F.31 ("FOI law")	Public-sector organizations **Use case:** "pure" academic research at a public institution	Clinical: yes Research: yes, but with broad exceptions	Likely	Best interests of the child (BIC)	No

report incidental or secondary findings are likely to be subject to access rights, as the data would no longer be used solely for research purposes. On the other hand, the return of such findings in university-based research would not likely trigger access rights.

It is important to note that the existence and scope of research exemptions vary across provinces and internationally. In Quebec, neither the FOI law nor the law governing public medical records contain a research exemption for access rights[12,13]. Individuals in Quebec thus appear to enjoy a general right of access to research information about themselves. In British Columbia, one FOI law governs all public entities, including hospitals, but its research exemption only applies to post-secondary educational institutions[14]. Access rights therefore appear to apply to all hospital-based researchers. For universities, Alberta's FOI law follows the same position as British Columbia's FOI law[15]. Alberta's health privacy law is restricted to information related to diagnosis, treatment or care, and so does not have information generated during research as a general concern[16]. Although, as is the case with Ontario's access laws, the robustness the clinical-research distinction may be questioned. Under the European *General Data Protection Regulation* (*GDPR*), access rights apply generally, but Member States are permitted to limit access rights in the research context "in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes" [European Union, 2016, Arts 9(2)(j) and 89]. Such is the case of Germany's *GDPR* implementation; access rights apply to research data unless the access rights are likely to render impossible or seriously impair the achievement of the research (Guerrini et al., 2019; Schickhardt et al., 2020).

Identifiability and the Characterization of Raw Whole Genome Sequence Data

Assuming no barriers to access rights by way of research exemptions, a legal right to access a child's raw WGS data further requires that the data be considered personally identifiable. For both the health privacy and FOI laws, regardless of clinical or research context, the standard of identifiability is the same: identifying information is information that either directly identifies an individual or information for which it is reasonably foreseeable that it could be used either alone or through combining information to identify an individual [*FIPPA*, s2(1); *PHIPA*, s4(1)]. In clinical care and in many genomic research contexts, genomic and health-related data are either nominally identifiable or coded (a link is maintained between the individual's name and their genomic data), thus maintaining its personal identifiability and consequently allowing for the data to be subject to an access right (Thorogood et al., 2018).

Both the health privacy and FOI laws in Ontario have been found to furnish individuals a legal right of access to raw data, with the IPC having rejected a distinction between raw data and information[17,18]. Central to this position under the health privacy law is that distinguishing between raw data and information would bring raw data outside of the data protection regime entirely. For genomic sequencing, the distinction between raw data and information would mean that BCL, FASTQ, BAM, and other such files are not subject to the security safeguards and other associated obligations created by law and that only a final lab report or other file resulting from a process of analysis or interpretation would be.

The information governed by the health privacy and FOI laws is nevertheless not coextensive with information to which an individual has a right of access. For example, if a record contains the information of other individuals or information subject to other access exemptions {e.g., quality of care information [*PHIPA*, s51(1)(a)]}, an individual only has access to information that can be reasonably severed (separated) from information to

[12] *Act respecting Access to documents held by public bodies and the Protection of personal information*, CQLR c A-2.1.

[13] *Act respecting health services and social services*, CQLR c S-4.2.

[14] *Freedom of Information and Protection of Privacy Act*, RSBC 1996, c 165, s3(1)(e).

[15] *Freedom of Information and Protection of Privacy Act*, RSA 2000, c F-25, s4(1)(i).

[16] *Health Information Act*, RSA 2000, c H-5, ss1(1)(i) and 1(1)(k).

[17] *Ontario (Natural Resources) (Re)*, 2003 CanLII 53917 (ON IPC).

[18] *St. Michael's Hospital (Re)*, 2017 CanLII 70006 (ON IPC).

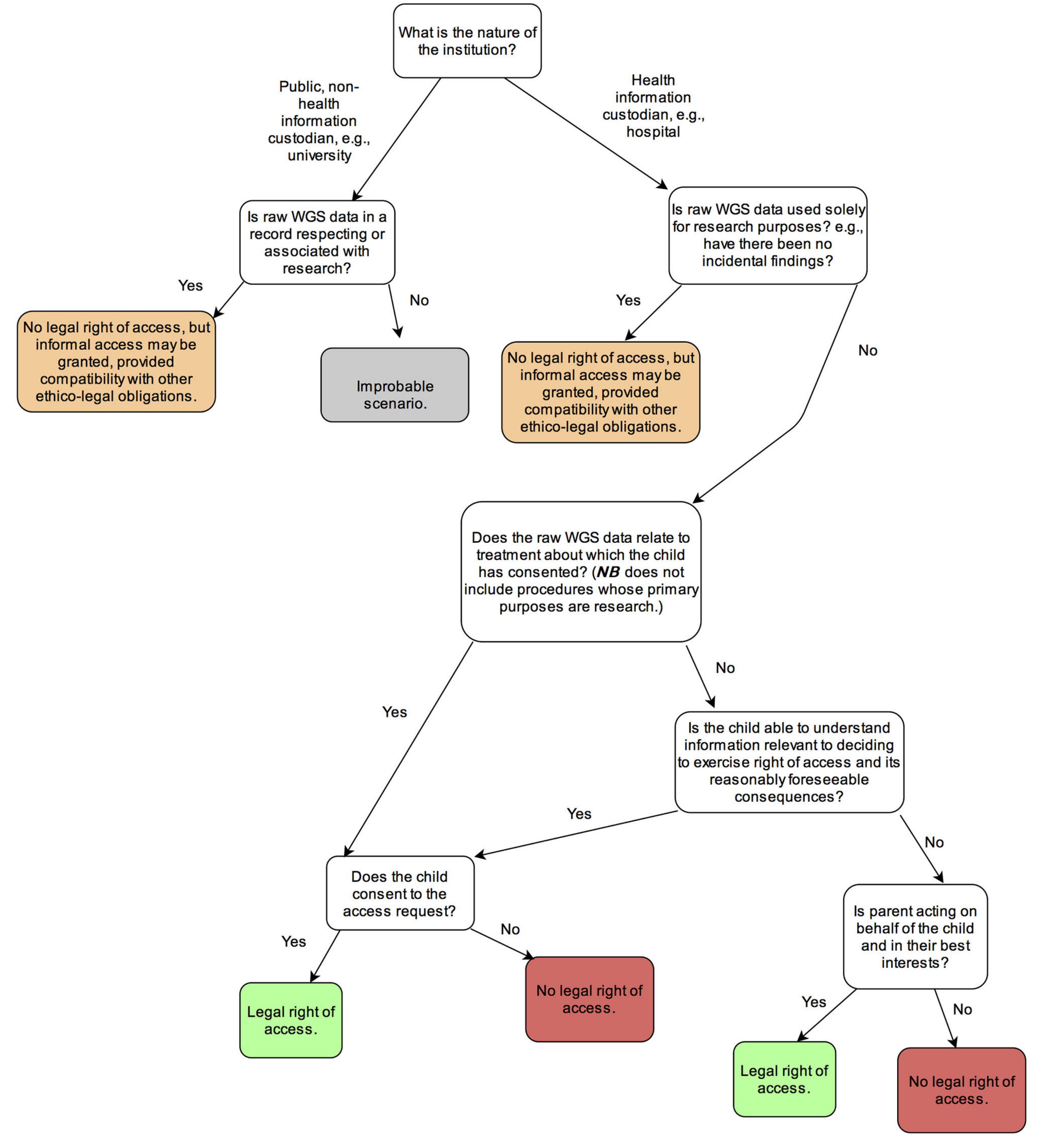

What is the nature of the institution?
Public, non-health information custodian, e.g., university
Health information custodian, e.g., hospital
Is raw WGS data in a record respecting or associated with research?
Yes
No
No legal right of access, but informal access may be granted, provided compatibility with other ethico-legal obligations.
Improbable scenario.
Is raw WGS data used solely for research purposes? e.g., have there been no incidental findings?
Yes
No
No legal right of access, but informal access may be granted, provided compatibility with other ethico-legal obligations.
Does the raw WGS data relate to treatment about which the child has consented? (NB does not include procedures whose primary purposes are research.)
Yes
No
Is the child able to understand information relevant to deciding to exercise right of access and its reasonably foreseeable consequences?
Yes
No
Does the child consent to the access request?
Yes
No
Legal right of access.
No legal right of access.
Is parent acting on behalf of the child and in their best interests?
Yes
No
Legal right of access.
No legal right of access.

FIGURE 1 | Decisional tree for determining whether parents have a legal right of access to their child's personal information in Ontario.

which an individual does not have a right of access [*FIPPA*, s10(2); *PHIPA*, s51(2)].

As regards the health privacy law, information subject to an access right has been found to include raw data from diagnostic equipment from which information in an individual's health record had been derived[19]. A guiding principle in determining which information an individual has access to under the health privacy law is informational reciprocity in the clinician-patient

[19] *Family Services of Peel (Re)*, 2019 CanLII 75908 (ON IPC).

TABLE 2 | Comparison of research exemptions in Ontario's health privacy and FOI laws.

	Health Privacy Law	FOI Law
Wording and Scope of Research Exemption	"Personal health information that a researcher uses solely for the purposes of research" is not subject to individual access rights. "Solely" for purposes of research suggests a stringent standard for information to fall within the exemption.	Information contained in records "respecting or associated with research" is not subject to the law. This includes clinical trial data conducted by a person employed by or associated with a hospital. "Respecting or associated with" research requires a substantial connection between the content of the record at issue and specific research being conducted.
Effect of Research Exemption	Exempts research data from the access provisions of the law. The other security safeguards with respect to personal health information continue to apply, however.	Exempts research data from the law in its entirety. No rights and obligations with respect to access, security safeguards, etc.
Situations Where Exemption Lacks Clarity	Individual undergoing whole genome sequencing as part of a research project and sequence data then informs clinical care of the individual, e.g., in cases of incidental or secondary findings.	When data at issue does not implicate academic freedom, e.g., individuals undergoing whole genome sequencing as part of a research project and seek access to this data.

relationship[20]. Hence, a patient has a legal right of access to any data, including raw data, to which their clinician would also reasonably have access. For example, an individual would not have a right of access to raw data used in machine processing and that a clinician cannot reasonably use[21]. Furthermore, an individual's access right extends to data that may be extracted via custom queries using currently available software and formats, but not if accessing the information would require the development of novel methods[22].

In Alberta, British Columbia, and Quebec, an individual's right of access generally extends to raw data about themself[23,24,25]. Only in Quebec, however, has it been found that individuals also have a legal right to access raw clinical data about themselves, even where an individual has already received a report based on the interpretation or analysis of this data[26]. The underlying logic behind this position is that the governing law does not distinguish between raw data and other types of information. Consequently, raw data forming part of an individual's health record can be the object of an access request. Information and privacy commissioner decisions in Alberta and British Columbia have not dealt with raw health-related data other than in the context of psychological tests. Given that no distinction exists between raw data and information as a matter of law, it appears strongly probable that, when faced with an individual access request to raw health-related data, an access right to raw data will be recognized. Indeed, in the European context, it has been argued that under the *GDPR*, individuals have a general right to raw data about themselves (Schickhardt et al., 2020).

As with the case in Ontario, an individual's right of access to raw data is not unfettered. For example, Quebec's FOI law does not give individuals a right of access to documents that require "computation or comparison of information"[27]. That is, an information custodian does not need to create a document or file of assembled information solely for the purposes of providing access[28]. Extracting data from an information system does not, however, constitute the creation of a new document[29]. In contrast, laws in British Columbia and Alberta require that the information custodian create a record for an individual exercising an access right[30,31,32]. However, and much as is the position under Ontario's health privacy law, individuals in Alberta, British Columbia, and Quebec do not have a right of access to data to which the information custodian does not have access through existing software and/or normal technical expertise[33,34,35,36,37]. Moreover, it is a commonality that the information must be reasonably severed if contained in a record not dedicated primarily to the individual's personal information[38,39,40,41]. Applied to raw WGS data, it then seems that an individual will have an access right to their raw WGS data, pending no other potential exclusions, as explored in the next sections.

[20] *St. Michael's Hospital (Re)*, 2017 CanLII 70006 (ON IPC).
[21] *St. Michael's Hospital (Re)*, 2017 CanLII 70006 (ON IPC).
[22] *St. Michael's Hospital (Re)*, 2017 CanLII 70006 (ON IPC).
[23] *Alberta Innovates – Technology Futures (Re)*, 2012 CanLII 70603 (AB OIPC), 2012.
[24] *G.F. c. Centre de réadaptation en déficience intellectuelle et en troubles envahissants du développement du Saguenay-Lac-Saint-Jean*, 2015 QCCAI 160 (CanLII).
[25] *Ministry of Forests, Re*, 2003 CanLII 49186 (BC IPC).
[26] *G.F. c. Centre de réadaptation en déficience intellectuelle et en troubles envahissants du développement du Saguenay-Lac-Saint-Jean*, 2015 QCCAI 160 (CanLII).
[27] *Act respecting Access to documents held by public bodies and the Protection of personal information*, CQLR c A-2.1, s15.
[28] *C.S. c. Société de l'assurance automobile du Québec*, 2017 QCCAI 251 (CanLII).
[29] *Stratégie 360 inc. c. Laval (Ville de)*, 2012 QCCAI 238 (CanLII).
[30] *Freedom of Information and Protection of Privacy Act*, RSA 2000, c F-25, s10.
[31] *Freedom of Information and Protection of Privacy Act*, RSBC 1996, c 165, s6.
[32] *Health Information Act*, RSA 2000, c H-5, s10.
[33] *Act respecting Access to documents held by public bodies and the Protection of personal information*, CQLR c A-2.1, s15.
[34] *Freedom of Information and Protection of Privacy Act*, RSA 2000, c F-25, s10.
[35] *Freedom of Information and Protection of Privacy Act*, RSBC 1996, c 165, s6.
[36] *Health Information Act*, RSA 2000, c H-5, s10.
[37] *Stratégie 360 inc. c. Laval (Ville de)*, 2012 QCCAI 238 (CanLII).
[38] *Act respecting Access to documents held by public bodies and the Protection of personal information*, CQLR c A-2.1, s14.
[39] *Freedom of Information and Protection of Privacy Act*, RSA 2000, c F-25, s6(2).
[40] *Freedom of Information and Protection of Privacy Act*, RSBC 1996, c 165, s4(2).
[41] *Health Information Act*, RSA 2000, c H-5, s7(2).

Best Interests of the Child: A Limit on Parental Access?

The best interests of the child (BIC) is a legal standard whose most important source is the *CRC*. The BIC is at the forefront of legal and ethical considerations in making decisions concerning a child (Sénécal et al., 2015b). The *CRC* itself deems BIC the "basic concern" of parents (United Nations General Assembly, 2007, 18). In Canada, "the values and principles of the Convention recognize the importance of being attentive to the rights and best interests of children when decisions are made that relate to and affect their future"[42].

The multifactorial, context-specific nature of the BIC has been criticized for its failure to produce clear, bright-line rules (Parker, 1994). It is, however, largely by virtue of the BIC's context-specific nature that gives it its potential to be applied using localized meanings and understandings in a way that serves its overarching purpose—the treatment of the child in a way that promotes their welfare while also being responsive to the child's age and capacities (Parker, 1994; Lansdown, 2005).

The BIC standard is relevant for the exercise of informational rights, and in particular the right of access. BIC has been applied in the context of access requests under both the health privacy and FOI laws. The approach of applying BIC under both laws is the same and can act as a limit on parents' ability to access information about their children[43,44]. For example, the Ontario IPC has found that a father making an information access request, despite having done so in good faith, was nevertheless not acting *on behalf of* the child, but rather for his own collateral purpose and so access to the information at issue was not granted[45]. The adjudicator further found, "based on the sensitive nature of the materials contained in the records, that the release of the son's personal information would not serve the best interests of the child."[46] The decision's reasoning that the exercise of rights *on behalf of* a child requires a connection to that child's best interests finds further support in Ontario's *Children's Law Reform Act*, which states that incidents of custody of the child, such as exercising an access to information right, are to be determined with reference to the BIC[47]. In this way, the intersection of BIC and informational rights ensures that the parent is in fact acting on behalf of the child in a way that coheres with that child's best interests.

Beyond Ontario, the BIC remains a primordial consideration in all decisions concerning a child. As in Ontario, however, each province had limited case law concerning the intersection of the BIC standard and informational rights. In Quebec, the Civil Code requires that all decisions concerning a child take into account that child's interests and rights, including the right of the child to be involved in the decision-making process in a way that is compatible with their maturity[48]. British Columbia similarly has found that acting on behalf of a child is synonymous with acting in the best interests of the child, even where informational rights are concerned[49]. The BIC is furthermore relevant in Alberta, where the disclosure of confidential information may be justified if it is in the child's best interests[50]. Similarly, the BIC and the child's right to be involved in decisions affecting themself are fundamental rights in the European Union (European Union, 2012, Art 24). The intersection of child's rights and informational rights has nevertheless garnered criticism on the basis that the child's evolving capacities are not adequately taken into consideration (Buitelaar, 2018).

Mature Minors in Access Contexts

Children are both bearers of rights and in need of protection owing to their vulnerability. As they age and mature, children present distinctive rights and needs, in the informational context and elsewhere. Central for our purposes is the *CRC*'s "participatory/emancipatory concept," whereby rights are transferred from the parent to the child in recognition of the child's developing maturity (Lansdown, 2005). Concern for the child's developing autonomy finds its principal expression in the involvement of the child in decision-making processes, such as the informed consent process (United Nations General Assembly, 1989, Art 12). The informed consent process must mediate between concerns for a child's developing autonomy, self-awareness, values, ability to understand, and the overarching concern for a child's best interests (Coughlin, 2018). Elucidating this mediation process, the Supreme Court of Canada has stated that the BIC "must be interpreted in a way that reflects and addresses an adolescent's evolving capacities for autonomous decision-making."[51] For children who possess a high level of maturity, the concerns for the child's welfare (concretized in the BIC standard) on one hand, and their autonomy on the other, "will collapse altogether and the child's wishes will become the controlling factor."[52]

Assuming that an access right exists in relation to the information, an information custodian must determine who is capable of exercising the right. In what follows, we examine the rights of access of children and of parents under both the FOI and health privacy laws (see **Figure 1** for summary). We will show that the two laws share a common point of departure—parents may exercise access rights where the child is under 16 years of age. But there is a lack of clarity regarding cases where the sampling and sequencing procedure serves both clinical and research purposes or the procedure is undergone solely for research purposes.

For the purposes of our analysis, we assume that the parent requesting access is a custodial parent and that the child to

[42] *Baker v. Canada (Minister of Citizenship and Immigration)*, 1999 CanLII 699 (SCC), para 71.

[43] *Family Services of Peel (Re)*, 2019 CanLII 75908 (ON IPC).

[44] *Ontario (Community Safety and Correctional Services) (Re)*, 2016 CanLII 25549 (ON IPC).

[45] *Ontario (Community and Social Services) (Re)*, 1994 CanLII 6595 (ON IPC).

[46] *Ontario (Community and Social Services) (Re)*, 1994 CanLII 6595 (ON IPC), p 3.

[47] *Children's Law Reform Act*, RSO 1990, c C.12, s19(a).

[48] *Civil Code of Québec*, CQLR c CCQ-1991, 1991, Arts 32–34.

[49] *British Columbia (Children and Family Development) (Re)*, 2018 BCIPC 47 (CanLII).

[50] *High Prairie School Division No. 48 (Re)*, 2012 CanLII 70631 (AB OIPC).

[51] *A.C. v. Manitoba (Director of Child and Family Services)*, 2009 SCC 30 (CanLII), para 88.

[52] *A.C. v. Manitoba (Director of Child and Family Services)*, 2009 SCC 30 (CanLII), para 87.

whom the WGS data relates follows a unidirectional, progressive trajectory with regard to their capacity for autonomous decision-making, i.e., that capacity is not present at one time but then is lost at a later time. We use the term "mature minor" in the narrow sense to refer to the mature minor legal doctrine, as well as in the broad sense to refer to minors who either have capacity to consent to treatment or who have informational capacity.

Both the FOI and health privacy laws share a common starting point: where a child is under 16 years of age, the parent or other LAR may exercise the right of access on behalf of the child (*FIPPA*, s66; *PHIPA*, s23). Contrary to the health privacy law, the FOI law does not incorporate the mature minor doctrine into its access provisions. The FOI law's bright-line approach, whereby parents exercise informational rights on behalf of a child under 16 years of age without regard to the circumstances, may be understood as fusing the interests of parent and child [*FIPPA*, s66(c)]. The only potential for a separation of the interests of the parent and child is through reference to the BIC standard (*infra*).

The health privacy law contains two key exceptions to the general rule that a parent or LAR may exercise the right of access on behalf of a child under 16 years of age. The first is that parents and other LARs do not have a right of access where the information relates to treatment or care to which the child has consented on their own [*PHIPA,* s23(2)(i)]. The health privacy law works in concert with the law governing capacity to consent to clinical treatment, ensuring that informational rights traces authority with regard to clinical decision-making (*HCCA*). The second exception to the general rule that parents or other LARs have a right to access information about a child under 16 years of age concerns minors who are informationally capable. We explore each in turn.

In Ontario, all individuals, including children, are presumed to be capable of consenting, unless the individual is unable to understand information relevant to the treatment[53]. Capacity to consent to treatment revolves around the notion that treatment, "means anything that is done for a therapeutic, preventive, palliative, diagnostic, cosmetic or other health-related purpose," (*HCCA*, s2). Capacity is determined on a treatment-by-treatment basis with regard to the capacity of the patient to understand the information relevant to that decision and to appreciate the associated reasonably foreseeable consequences (*HCCA*, s4). Consequently, a minor may be competent to consent to a low-risk procedure such as the removal of a mole, while for higher risk procedures, such as novel chemotherapies, that same minor would not be competent to consent. Where a mature minor has consented to a sampling and sequencing procedure for a clinical-diagnostic purpose, a parent does not have a right of access to any of the sequence data, raw or otherwise, unless the minor consents to releasing the information to the parent.

For a procedure whose "primary purpose" is research, however, the general rules regarding parental access under the health privacy law apply. Consequently, if the child is under 16 years of age and the primary purpose of the sequencing and sampling was research, then the parent will have *prima facie* a right of access to that information. The breadth of scenarios covered by the primary purpose criterion is broad. Where a child undergoes sampling and sequencing as part of participation in a research study in the hope that the data generated will be relevant for treatment, the primary purpose appears to remain research as the research study is the reason for which the data is generated. Any potential clinical application is merely secondary. This is significant as the primary purpose criterion implicates both pure research and research-clinical scenarios. Recall that, as regards the former, the return of incidental findings should hypothetically trigger an access right and that the research exemption would not apply to the latter because such information would no longer be used exclusively for research purposes.

The effect of the foregoing is that where adolescent children undergo a sampling and sequencing procedure, a parent will *prima facie* have a right to data generated in either the pure research or research-clinical contexts, but not to data generated in relation to the clinical-diagnostic context. Informational rights do not trace decisions by a minor regarding research participation because there is no legal mature minor doctrine for research participation. As an ethical process, assent does not directly affect legal rights with regard to information related to research.

The second exception to the general rule that parents or other LARs have a right to access information about a child under 16 years of age concerns minors who are informationally capable. Where a minor child is able to "understand the information that is relevant to deciding whether to consent" and to "appreciate the reasonably foreseeable consequences" of a decision regarding their information, they are recognized as having capacity for the purposes of the health privacy law (herein "informational capacity") [*PHIPA*, s21(1)]. Where the decision of a parent or other substitute decision-maker differs from a capable child as regards that decision, the child's decision prevails over the conflicting decision of the parent[54] [*PHIPA*, s23(3)].

Informational capacity gives the minor a voice, even where they have not consented to the procedure to which the information relates. However, it introduces complexities for information custodians. Examining whether the child consented to the procedure at issue is only a first step. Even if they did not, they may still possess informational capacity such that the access right must be exercised by the child themself. The relevant point in time for undertaking the analysis is at the time of the access request. Consequently, an adolescent may be likely to possess informational capacity for information that relates to a procedure they underwent in their tender years.

One may still question the significance of informational capacity in practice. Consider that when a minor's decision regarding treatment or care is at issue, the clinician has directly interacted with the minor-patient and so is in an appropriate position to judge that minor's capacity to make a choice for treatment. Yet in the informational context, an access request will likely be handled by an administrator without personal knowledge of the child to whom the information relates. There is no explicit obligation for an information custodian to determine whether a child is informationally capable upon receipt of an

[53] *Health Care Consent Act* (*HCCA*), 1996, SO 1996, c 2, Sch A.

[54] *Family Services of Peel (Re)*, 2019 CanLII 75908 (ON IPC).

access request from a parent (Perun et al., 2005). It appears likely that, unless an information custodian knows that a minor disagrees with a parent's access request, access is likely to be granted. The general duty of information custodians to act in "in good faith and reasonably in the circumstances" may, however, give rise to an obligation for custodians to take into consideration the minor's decision-making capacity at the time of the access request [*PHIPA*, s71(1)].

The circumstances under which a parent has access to their child's information varies widely by province in Canada. For consent to clinical care, Quebec follows an age-based criterion that presumes any individual above 14 years of age may consent to treatment required for their health unless there is reason to believe the individual does not have sufficient decision-making capacity[55]. Similar to Ontario, informational rights map onto this age of consent: minors over the age of 14 have access rights under the statute that governs individuals' medical files[56]. Where a parent requests access to information in a medical record that relates to a child who is 14 or older, the custodian must first consult the child and the child's decision regarding whether or not to provide access to the parent is binding[57].

Other provinces follow the "mature minor" doctrine, initially developed in England and Wales, whereby a minor who is capable of understanding the proposed course of clinical action and is capable of expressing their own wishes regarding this course of action may consent to care, provided that it is in their best interests and notwithstanding their general lack of legal capacity due to their age (e.g., British Columbia and Alberta)[58,59,60,61] (Dalpé et al., 2019). In British Columbia, a parent may only exercise a child's access right where the child is incapable of exercising it themselves[62]. In practice, informational competence tends to be recognized at the age of 12 and so parents require their child's authorization to access their health files[63] (see, e.g., Health Information Management, 2020). Likewise, Alberta's health privacy law is also consistent with the mature minor doctrine regarding consent to clinical care. Under Alberta's health privacy law, a parent making an access request for information about their child under the age of 18 bears the onus of proving that their child "does not understand the nature of the right… and the consequences of exercising the right" at the time that the request is made[64,65]. Notably, "the level of understanding that is required for an individual to understand the nature and consequences of exercising rights or powers under [Alberta's health privacy law] is not a particularly onerous standard,"[66]. The kind of information at issue appears to be irrelevant for the purpose of determining informational capacity. Thus, there is no clear support that there is a higher capacity required for exercising access rights over raw WGS data than other kinds of health information. Alberta's FOI law takes a novel approach in that a parent may exercise the child's rights provided that such exercise does not cause "an unreasonable invasion of the personal privacy of the minor," which presumably is intended to be a case-by-case determination with the minor's evolving capacities taken into account[67]. The position under the *GDPR* is largely a question for Member State law, as the regulation only makes specific provision for the age at which children may consent in relation to information society services, and not to data processing activities generally (European Union, 2016, Art 8). In view of the wide diversity of approaches and the multiple considerations at play (e.g., age of consent to clinical treatment, to research, to data processing, duties to assess a minors' capacity before allowing parents to exercise their rights, duties to consult minors before releasing data, etc.), health professionals should carefully consider the potential interface of these factors under local law.

DISCUSSION

Individuals in Ontario have a legal right to access their genomic data used for clinical and translational research. Ethical and legal literature in genomics has mainly focused on the obligations of professionals implicated in the bioinformatics pipeline with respect to test interpretation and the return of incidental or secondary findings, with raw data receiving less attention (Borry et al., 2018). Previous articles in the Canadian context have focused on access requests in the context of academic health research (Ries, 2010), or on individual control over genetic information (Ogbogu et al., 2014). We expand upon these articles by identifying legal access rights to clinical data and by clarifying the scope of research exemptions in Ontario. In healthcare institutions, only data solely used for research is exempted from the individual right of access. In academic institutions, all data associated with research is exempted. We also find case law indicating that access rights should encompass raw WGS data. Exempt research projects can still decide to offer access as a matter of ethics and participant engagement.

Our study is the first to trace the contours of parental access rights where children undergo WGS. We find that parents' authority to request access must be exercised on behalf of the child and in that child's best interests. Health information custodians would have grounds to refuse an access request manifestly not in the child's best interests. Furthermore, we find that health information custodians likely have a duty to ensure parents are not granted access to a mature minor's information, unless the minor consents or the parent demonstrates that the minor lacks capacity to make decisions about the disclosure of their health information. The position concerning access rights

[55] *Civil Code of Québec*, CQLR c CCQ-1991, 1991, Arts 14 and 17.

[56] *Act respecting health services and social services*, CQLR c S-4.2, ss17, 20, 21.

[57] *Gagné c. Hôpital Ste-Justine* [1999] CAI 261.

[58] *Infants Act*, RSBC 1996, c 223, 1996, s17.

[59] *A.C. v. Manitoba (Director of Child and Family Services)*, 2009 SCC 30 (CanLII).

[60] *Gillick v West Norfolk and Wisbech AHA* [1985] UKHL 7.

[61] J.S.C. v. Wren, 1986 ABCA 249 (CanLII).

[62] *Freedom of Information and Protection of Privacy Regulation*, BC Reg 155/2012, 2012, s3(1).

[63] MyHealthPortal Available at: https://www.interiorhealth.ca/YourHealth/MyHealthPortal/Pages/default.aspx [Accessed February 14, 2020].

[64] *Calgary Health Region (Re)*, 2006 CanLII 80851 (AB OIPC).

[65] *Health Information Act*, RSA 2000, c H-5, s104(1)(c).

[66] *Calgary Health Region (Re)*, 2006 CanLII 80851 (AB OIPC), para 74.

[67] *Freedom of Information and Protection of Privacy Act*, RSA 2000, c F-25, s84(1)(e).

is most clear in the case of sequencing for exclusively clinical-diagnostic purposes across jurisdictions, albeit with importance nuances among them. Parental access to research data typically is not possible, due to the research exemptions. If it is possible, however, the provinces also differ greatly in their approaches.

Despite the clear evidence of BIC's relevance for the exercise of informational rights, the cases in which BIC has been applied in the context of parental access requests are limited. Indeed, this was true across all Canadian provinces under study. The small number of cases suggests either that information custodians do not generally deny parental access requests or that such denials are not appealed to the provincial IPC (Cases only appear in front of the provincial IPC in circumstances where an individual is challenging a decision made by an information custodian). It is moreover difficult to envisage the circumstances in which an information custodian would be able to easily distinguish between the BIC and any ulterior motives on the part of a parent. For example, one of the few cases where BIC was an express consideration was when a parent had requested information from a police report on behalf of their children, but the children were fearful of them and did not desire contact with the parent[68]. It thus seems that the BIC standard exists in principle in relation to information access requests, but the circumstances in which information custodians may meaningfully invoke it are limited. Importantly, we note that the vast majority of parents are likely making decisions that are in keeping with their child's best interests. As such, information custodians should not be quick to second-guess parental motives in most circumstances. To this end, see section Conclusion and Points to Consider for recommendations in.

Leveraging Professional Expertise With Access Rights

While law provides an important framework in this area, ensuring parental access supports the welfare, privacy, and developing autonomy of children will primarily depend on the ethical behavior of both professionals and parents. One important challenge for information custodians and professionals is the difficulty of distinguishing beneficial parental access requests from improper ones. Likewise, it may be difficult to craft legislation or professional guidelines that effectively make this distinction. Too much intervention risks depriving children of their right to receive parental guidance in keeping with their age and capacities.

The existence of legal access rights, rather than trumping professional obligations, invites us to reconsider how the child's best interests can be furthered. Professional expertise should be leveraged to further the child's best interests, which should include the involvement of the most important individuals in a child's life—their parents. Professionals should thus engage with parents and help them decide if access is the right decision for their child, and how to responsibly handle the data once accessed. The potential for parental access may also encourage professionals to more carefully consider whether or not to sequence children in the first place. Ultimately, much of the responsibility to act on behalf of, and thus safeguard, the child's interests will rest with parents. Careful management of a child's personal information is an increasingly important parental responsibility–this responsibility also extends to genomic data.

A fruitful starting point in ensuring that the exercise of access rights is in keeping with the child's best interests are professional guidelines developed to address the return of incidental or secondary genomic findings in children. These guidelines highlight the ethical challenges with respect to handling the genomic data of children, particularly where it reveals health risks that may only materialize after the child has reached maturity. On the one hand, returning predictive information to children and their parents may inform childhood or adulthood actions that could improve the child's future health (Johnson et al., 2017). The return of information may also better inform the health choices of family members, which can improve the overall well-being of the child (Hardart and Chung, 2014). On the other hand, returning predictive information may threaten the child's future autonomy and ethical right not to know (Feinberg, 1980). Return may also lead to psychological harms (e.g., anxiety, low self-esteem), harms to family relationships, and potential discrimination (McGuire et al., 2020). Flowing from these competing concerns, professional guidelines have made different recommendations about the reporting of adult-onset genomic findings in pediatrics.

Clinicians using WGS tests may look to their professional associations for guidance on how to deal with requests to provide parents access to their child's raw WGS data. The American College of Medical Genetics (ACMG), the Canadian College of Medical Genetics (CCMG), and the European Society of Human Genetics (ESHG) have not published any policies on responding to access requests to raw data generally ("pulling" data). Nevertheless, each organization does have a position on the return of incidental or secondary findings, i.e., "pushing" data (Green et al., 2013; van El et al., 2013; Boycott et al., 2015).

Secondary findings describe pathogenic variants that are identified in the genome of a patient unrelated to the primary purpose of the testing (Knoppers et al., 2015). Secondary findings and raw data are undoubtedly distinct from one another: the former are curated (and, thus, the product of an interpretive process) and the latter are merely the subject-matter of that interpretive process. However, both represent different forms of genomic information that can be returned to individuals if requested. With this common characteristic considered, and in the absence of any guidance on the return of raw data, it is worth briefly exploring positions on the return of secondary findings.

The ACMG has the most permissive policy on returning secondary findings, recommending that a predetermined list of variants associated with medically actionable disorders be returned to patients, provided consent is obtained, in addition to primary test results (Green et al., 2013). Importantly, while the majority of these conditions are adult-onset, the ACMG also recommends returning these variants when found in children as the results may have immediate implications for family members and for the child when they are older. The ACMG also highlights the importance of parental decision-making when it comes to

[68] *Ontario (Community Safety and Correctional Services) (Re)*, 2016 CanLII 25549 (ON IPC).

genetic testing. Despite its nuances, this approach not been without detractors (Garrett et al., 2019).

The ESHG and CCMG take a more cautious and classical approach by suggesting the creation of a bioinformatics pipeline that minimizes the identification of secondary findings (van El et al., 2013; Boycott et al., 2015). The CCMG nevertheless recognizes that labs may want to search for secondary findings and provide guidance on what results to return. They suggest that labs searching for secondary findings ought to return results for highly penetrant conditions that are medically actionable in childhood. Variants associated with adult-onset medically actionable conditions should only be returned upon request, when the data has the potential to prevent serious harm to the health of a parent or family member. The ESHG highlights concerns over respecting the emerging autonomy of children, while the CCMG suggests that there may be psychosocial harms associated with returning secondary findings as a rationale for their cautious approach.

While providing secondary findings and returning raw data both involve returning genomic information that may have nothing to do with the primary indication for testing, the scale of data being returned is vastly different. For example, the ACMG suggests screening for pathogenic variants in only 59 genes (Kalia et al., 2017). In contrast, raw genomic data contains information on all genes and intervening sequence in the genome. Returning raw data could be considered analogous to returning all variants, depending on what is done with the data. Raw data could be analyzed to identify variants associated with adult-onset non-medically actionable diseases, variants of unknown significance, and the carrier status of the child. To our knowledge, no professional guideline or policy has even contemplated returning this type of information to parents.

Despite the silence of professional norms regarding the return of raw sequence data, many laboratories performing clinical WGS permit raw data release. A recent study examined the content of publicly available consent forms to determine whether they complied with recommendations made by the ACMG and the Presidential Commission for the Study of Bioethical Issues (Fowler et al., 2018). Germane to this discussion was the recommendation made by the Presidential Commission that patients be informed of what data and information may be returned (Presidential Commission for the Study of Bioethical Issues, 2012). Of the 18 consent forms analyzed, 44% provided for return of the raw data to the clinician, with commercial laboratories being more likely to permit raw data release compared to academic labs (Fowler et al., 2018). This study suggests that a large minority of patients are made aware that raw data release is possible and that clinicians are the gatekeepers for this information. Regardless, patients generally have legal rights to access health information held by laboratories in Ontario either directly or indirectly through their clinician[69].

In this vein, laboratory data retention practices are noteworthy. Despite health information retention laws, and professional recommendations for retention of some data files by clinical genetics laboratories, both policy and practice remain unclear and variable. For example, the CCMG recommends that clinical genetics laboratories retain the VCF file for a minimum of 2 years and possibly even longer for the testing of minors or for inherited disorders with familial implications (Hume et al., 2019). Surprisingly, the CCMG's recommendation of retaining a VCF file for at least 2 years is markedly shorter than the periods established by other legal and ethical norms for retention of health information, e.g., 10 years in the case of health information and 5–10 years for diagnostic imaging records[70,71]. One possible reason for this discrepancy could be that clinical genetics laboratories do not typically have direct contact with patients and the VCF file represents an intermediate step between the act of sequencing and the information relayed to a patient by their clinician. The existence or accessibility of the file over time clearly has implications for parental access.

Our analysis of legal rights of parental access is connected to another debate regarding parents' ability to have their child tested through direct to consumer (DTC) genetic testing services. Usually, children are only sequenced in health care where there is an important clinical indication, and in research where there is a need to improve our understanding of serious childhood conditions. With DTC, parents can seek genetic testing of healthy children or children with non-serious health conditions. Some of the health information they may receive, such as information about adult-onset conditions, raise the ethical issues highlighted above between access to health information for children and families, and closing of the child's future choices not to know their health status. Furthermore, parents can generally access their child's genetic data in the DTC context because *PIPEDA* sees the parent effectively exercising the child's legal informational rights on behalf of the child and does not expressly consider the rights of children with regard to access. Parents may then share the child's data with third party interpretation services, clinicians, researchers, and even open-access recreational genomics sites. While this may offer interesting health, research, and recreational opportunities for both parents and children, there is also the potential for important privacy risks and discrimination.

With the increase in sequencing in the research and clinical contexts, coupled with the advent of DTC genetic testing services, parents have greater freedom to test their children for various health risks and to direct the sharing of their children's data. A recent study counted as many as 35 raw genomic data interpretation services available to consumers online (Capaci et al., 2020). Parents are already attracting more responsibilities for safeguarding their children's privacy with their social media interactions. Such responsibilities are likely to extend to understanding the health and privacy implications of genetic testing and data sharing for children.

Our study focused on describing the application of current law to parental genomic access requests. Future legal studies could explore if laws should be adapted to be more responsive to the challenges of genomic and children's privacy. Future legal research questions include the following: Are individual access

[69] General, O Reg 329/04 (*Personal Health Information Protection Act*, 2004), 24(1)(2).

[70] General, O Reg 114/94 (*Medicine Act*, 1991).

[71] Hospital Management, RRO 1990, Reg 965 (*Public Hospitals Act*).

rights an appropriate and effective way to empower patients? For example, it has been highlighted that proactive approaches of providing individuals with access to their data would be better for all parties involved, as formal access requests are clunky and time-consuming (Kwoka, 2017). Should health privacy laws incorporate more explicit protection and consideration for the child's best interests? (Buitelaar, 2018; Savirimuthu, 2019) Is direct regulation of parents regarding their children's genetic data desirable? Feasible? What about greater regulation of third-party interpretation services, especially when it comes to children's genetic data? (Guerrini et al., 2020)

CONCLUSION AND POINTS TO CONSIDER

Health professionals, researchers, and their organizations must carefully consider the legal and ethical implications of parental access when handling requests or designing personal genomic access policies and processes. They need to be able to determine when parents have a legal right to access their child's health information, the ethical implications of parental access for the child, and their corresponding professional duties to protect the best interests and developing capacity of young patients and research participants.

While our study has focused on the legal rights of access of parents to their child, the avenues of inquiry may be generalized for other jurisdictions. Individuals should identify the controlling legal framework regarding individual access rights, which will most often be contained in either privacy and data protection or freedom of information laws. It is also essential to determine the existence and ambit of any exemptions in the research context. Furthermore, individuals should examine if raw data constitutes personal information under relevant privacy and data protection and/or freedom of information laws. Where parents are requesting the personal information of their children, two additional and interrelated issues are present: the BIC and provisions for "mature minors"/the need to involve a child in appropriate manner based on age and competence. Either of these aspects of children's rights may temper the parental access right.

General recommendations for personal genomic access in healthcare and health research contexts have already been developed in the German context by the Ethical and Legal Aspects of Whole Genome Sequencing project (EURAT) (Winkler et al., 2019). We endorse EURAT's core recommendations of pre- and post-access education. Such an approach sees professional expertise working together with legal rights to further the health, privacy, and general welfare of probands and their families. EURAT recommends that an initial conversation be held with requestors to explain the access process and assess their capacity and motivation. At this stage, general information about the nature and implications of raw genomic data should be provided to help requestors determine if access will serve their purposes. This information may include disclaimers about quality and fitness for medical use, information about the limited meaning of the information, the need for expert interpretation, and the health and privacy risks to the individual and family members that can arise from sharing genomic data. The individual can then be offered an opportunity for sober reflection and reconsideration after this initial conversation. If the individual proceeds with the request, they can be offered general written information about the health and privacy implications of the raw data should be provided, as well as an opportunity for personal consultation, while making it clear this is not individualized genetic counseling. Each of these steps should be carefully documented.

Overall, EURAT also recommends that healthcare organizations and research projects should establish a clear and accessible policy to facilitate handling of requests, describing the scope of the right to access, the process for requesting access, and opportunities to receive information and consultation. Moreover, appropriate quality-control mechanisms for sample and data tracking and identity authentication processes must be in place to ensure the right data is returned to the right person. One final general consideration is that access requests should be directed through the ordering physician, and not the laboratory directly.

While a useful source of guidance, EURAT's recommendations are neither specific to the pediatric context nor to the unique contours of legal access rights in Canada. As such, we propose these additional considerations:

- If possible, professionals in the child's circle of care should speak to parents who are requesting access to raw sequence data to better understand the context of the request. It may turn out that the parents' request may be better satisfied by other avenues, e.g., returning interpreted results. An explanation of the interpretation processes the sequence data have already undergone can assist parents in understanding the nature of their request. For example, if a search has already been conducted for highly penetrant conditions that are clinically actionable in childhood, parents may decide that having the raw sequence data is not needed.
- Pre- and post-access informational materials and consultations should inform parent requestors about the implications of raw data for the child's well-being, privacy, and developing autonomy. They should also inform parental requestors of their ethical responsibilities for handling, using and sharing their child's genome responsibly.
- Information custodians should withhold access if it is manifestly clear to the professional that the parent is not acting on behalf of the child, viz. for an ulterior purpose such as uploading the child's sequence data to an online portal for a reason disconnected from the child's best interests. Nevertheless, we recognize that professionals may rarely have clear evidence about the motivations to justify refusing parental access. Moreover, parents can always lodge an appeal to an information custodian's decision with which they disagree.
- Steps should be taken to determine that only the individual who is legally authorized to exercise the child's access right is permitted to access data (parent, mature-minor, LAR, or no one). This will generally be determined by the age

of consent, but also exceptionally by the child's level of maturity in the clinical context. We provide a flow chart to aid with this determination (see **Figure 1**). In particular, health information custodians should consider an older child's developing maturity: they should seek to determine if the child has the capacity to exercise informational rights alone before granting access to parents, and to ensure the child has been consulted about the request in an age-appropriate manner.

- In pediatric research contexts where there is no legal right to access, a governance decision should be made before recruitment as to whether or not the project will provide access to parents, considering the consequences for research integrity, available resources, and expectations of participants. The specific research context may be important. Parents of sick children with rare diseases, chronic conditions, or cancer may deserve greater deference in managing their child's genetic data in order to drive their care and related research, than parents of healthy children. If providing access may bias research outcomes, then access may require the participant to withdraw from the study. If a research project voluntarily opts to provide access, the considerations above for doing so responsibly are applicable.

AUTHOR CONTRIBUTIONS

MB conducted the doctrinal review. MB and AT interpreted the data. MB, AT, MS, and KS drafted the manuscript while critical revision was provided by MZ and BK. All the authors approved the manuscript for publication.

FUNDING

BK, AT, and MB received financial support from the Government of Canada through Genome Canada, Genome Quebec, and the Ontario Genomics Institute (OGI-148). KS and MZ received financial support from the PRecision Oncology for Young peopLE 2 (PROFYLE 2) program and through funds from the Terry Fox Research Institute and all other funders supporting PROFYLE 2.

ACKNOWLEDGMENTS

We thank Genome Canada, Genome Quebec, the Ontario Genomics Institute, the Terry Fox Research Institute, and PROFYLE 2 for their support.

REFERENCES

Abril, J. F., and Castellano, S. (2019). "Genome annotation," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press), 195–209.

Bala, N. (2020). Opinion | Why Are You Publicly Sharing Your Child's DNA Information? *N. Y. Times*. Available at: https://www.nytimes.com/2020/01/02/opinion/dna-test-privacy-children.html (accessed January 7, 2020).

Bombard, Y., Robson, M., and Offit, K. (2013). Revealing the incidentalome when targeting the tumor genome. *JAMA* 310, 795–796. doi: 10.1001/jama.2013.276573

Borry, P., Bentzen, H. B., Budin-Ljøsne, I., Cornel, M. C., Howard, H. C., Feeney, O., et al. (2018). The challenges of the expanded availability of genomic information: an agenda-setting paper. *J. Community Genet.* 9, 103–116. doi: 10.1007/s12687-017-0331-7

Borry, P., Evers-Kiebooms, G., Cornel, M. C., Clarke, A., Dierickx, K., and Public and Professional Policy Committee (PPPC) of the European Society of Human Genetics (ESHG) (2009). Genetic testing in asymptomatic minors: background considerations towards ESHG Recommendations. *Eur. J. Hum. Genet.* 17, 711–719. doi: 10.1038/ejhg.2009.25

Botkin, J. R., Belmont, J. W., Berg, J. S., Berkman, B. E., Bombard, Y., Holm, I. A., et al. (2015). Points to consider: ethical, legal, and psychosocial implications of genetic testing in children and adolescents. *Am. J. Hum. Genet.* 97, 6–21. doi: 10.1016/j.ajhg.2015.05.022

Boycott, K., Hartley, T., Adam, S., Bernier, F., Chong, K., Fernandez, B. A., et al. (2015). The clinical application of genome-wide sequencing for monogenic diseases in Canada: position statement of the Canadian College of Medical Geneticists. *J. Med. Genet.* 52, 431–437. doi: 10.1136/jmedgenet-2015-103144

Bredenoord, A. L., Onland-Moret, N. C., and Van Delden, J. J. M. (2011). Feedback of individual genetic results to research participants: in favor of a qualified disclosure policy. *Hum. Mutat.* 32, 861–867. doi: 10.1002/humu.21518

Brown, T. L., and Meloche, T. M. (2016). Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* 108, 109–114. doi: 10.1016/j.ygeno.2016.06.003

Buitelaar, J. C. (2018). Child's best interest and informational self-determination: what the GDPR can learn from children's rights. *Int. Data Priv. Law* 8, 293–308. doi: 10.1093/idpl/ipy006

Capaci, M., Crombag, N., Devriendt, T., Demuynck, R., Kalokairinou, L., Pasquier, L., et al. (2020). "Fifteen years of consumer genomics: growing fragmentation and greater diversity of genomic services," in *European Society of Human Genetics Conference* (Berlin). Available online at: https://www.abstractsonline.com/pp8/#!/9102/presentation/2281 (accessed May 10, 2020).

Chakradhar, S. (2018). Matching up. *Nat. Med.* 24, 882–884. doi: 10.1038/s41591-018-0113-8

Coughlin, K. W. (2018). Medical decision-making in paediatrics: infancy to adolescence. *Paediatr. Child Health* 23, 138–146. doi: 10.1093/pch/pxx127

Dalpé, G., Thorogood, A., and Knoppers, B. M. (2019). A tale of two capacities: including children and decisionally vulnerable adults in biomedical research. *Front. Genet.* 10:289. doi: 10.3389/fgene.2019.00289

Eckford, P. D. W., McCormack, J., Munsie, L., He, G., Stanojevic, S., Pereira, S. L., et al. (2019). The CF Canada-sick kids program in individual CF therapy: a resource for the advancement of personalized medicine in CF. *J. Cystic Fibr.* 18, 35–43. doi: 10.1016/j.jcf.2018.03.013

European Union (2012). Charter of Fundamental Rights of the European Union.

European Union (2016). Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natral Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection *Regulation).*

Evans, B. J. (2017). "Genomic Data Commons," in *Governing Medical Knowledge Commons*, eds K. J. Strandburg, B. M. Frischmann, and M. J. Madison (Cambridge: Cambridge University Press), 74–101.

Feinberg, J. (1980). "The child's right to an open future," in *Whose Child? Children's Rights, Parental authority, and State Power*, eds W. Aiken and H. LaFollette (Totowa, NJ: Rowman and Littlefield), 124–53.

Fowler, S. A., Saunders, C. J., and Hoffman, M. A. (2018). Variation among consent forms for clinical whole exome sequencing. *J. Genet. Couns.* 27, 104–114. doi: 10.1007/s10897-017-0127-2

Garrett, J. R., Lantos, J. D., Biesecker, L. G., Childerhose, J. E., Chung, W. K., Holm, I. A., et al. (2019). Rethinking the "open future" argument against predictive genetic testing of children. *Genet. Med.* 21, 2190–2198. doi: 10.1038/s41436-019-0483-4

Gitelman, L., and Jackson, V. (2013). "Introduction," in *"Raw data" Is an Oxymoron Infrastructures Series*, ed L. Gitelman (Cambridge, MA; London: The MIT Press), 1–14.

Goh, G., and Choi, M. (2012). Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inform.* 10, 214–219. doi: 10.5808/GI.2012.10.4.214

Graaf, R., van der, Dekking, S. A., Vries, M. C., de, Zwaan, C. M., and Delden, J. J. M., van (2018). Pediatric oncology as a Learning Health System: ethical implications for best available treatment protocols. *Learn. Health Syst.* 2:e10052. doi: 10.1002/lrh2.10052

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Guerrini, C. J., Botkin, J. R., and McGuire, A. L. (2019). Clarify the HIPAA right of access to individuals' research data. *Nat. Biotechnol.* 37, 850–852. doi: 10.1038/s41587-019-0190-3

Guerrini, C. J., Wagner, J. K., Nelson, S. C., Javitt, G. H., and McGuire, A. L. (2020). Who's on third? regulation of third-party genetic interpretation services. *Genet. Med.* 22, 4–11. doi: 10.1038/s41436-019-0627-6

Hall, A., Finnegan, T., Alberg, C., and PHG Foundation (2014). *Realising Genomics in Clinical Practice*. Cambridge: PHG Foundation.

Hardart, G. E., and Chung, W. K. (2014). Genetic testing of children for diseases that have onset in adulthood: the limits of family interests. *Pediatrics* 134, S104–S110. doi: 10.1542/peds.2014-1394F

Hawcutt, D. B., Thompson, B., Smyth, R. L., and Pirmohamed, M. (2013). Paediatric pharmacogenomics: an overview. *Arch. Dis. Child.* 98, 232–237. doi: 10.1136/archdischild-2012-302852

Health Information Management (2020). *Authorization for the Release of Health Records*. Available online at: http://www.himconnect.ca/Documents/Authorization-for-the-Release-of-Health-Records.pdf (accessed February 9, 2021).

Hume, S., Nelson, T. N., Speevak, M., McCready, E., Agatep, R., Feilotter, H., et al. (2019). CCMG practice guideline: laboratory guidelines for next-generation sequencing. *J. Med. Genet.* 56, 792–800. doi: 10.1136/jmedgenet-2019-106152

Hutchinson, T. (2018). "Doctrinal research: researching the jury," in *Research Methods in Law*, eds D. Watkins and M. Burton (Abingdon, Oxon; New York, NY: Routledge), 8–39.

Jarvik, G. P., Amendola, L. M., Berg, J. S., Brothers, K., Clayton, E. W., Chung, W., et al. (2014). Return of genomic results to research participants: the floor, the ceiling, and the choices in between. *Am. J. Hum. Genet.* 94, 818–826. doi: 10.1016/j.ajhg.2014.04.009

Johnson, L.-M., Hamilton, K. V., Valdez, J. M., Knapp, E., Baker, J. N., and Nichols, K. E. (2017). Ethical considerations surrounding germline next-generation sequencing of children with cancer. *Expert Rev. Mol. Diagn.* 17, 523–534. doi: 10.1080/14737159.2017.1316665

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). 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. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Kamchedzera, G. (2012). *Article 5: The Child's Right to Appropriate Direction and guidaNce*. Leiden; Boston, MA: Martinus Nijhoff Publishers.

Kesserwan, C., Friedman Ross, L., Bradbury, A. R., and Nichols, K. E. (2016). The advantages and challenges of testing children for heritable predisposition to cancer. *Am. Soc. Clin. Oncol. Educ. Book Am. Soc. Clin. Oncol. Annu. Meet.* 35, 251–269. doi: 10.14694/EDBK_160621

Khan, T., Stewart, M., Blackman, S., Rousseau, R., Donoghue, M., Cohen, K., et al. (2018). Accelerating pediatric cancer drug development: challenges and opportunities for pediatric master protocols. *Ther. Innov. Regul. Sci.* 53, 270–278. doi: 10.1177/2168479018774533

Kish, L. J., and Topol, E. J. (2015). Unpatients—why patients should own their medical data. *Nat. Biotechnol.* 33, 921–924. doi: 10.1038/nbt.3340

Knoppers, B. M., Avard, D., Sénécal, K., and Zawati, M. H. (2014). Return of whole-genome sequencing results in paediatric research: a statement of the P3G international paediatrics platform. *Eur. J. Hum. Genet.* 22, 3–5. doi: 10.1038/ejhg.2013.176

Knoppers, B. M., Nguyen, M. T., Sénécal, K., Tass,é, A. M., and Zawati, M. H. (2016). Next-generation sequencing and the return of results. *Cold Spring Harb. Perspect. Med.* 6:a026724. doi: 10.1101/cshperspect.a026724

Knoppers, B. M., Zawati, M. H., and Sénécal, K. (2015). Return of genetic testing results in the era of whole-genome sequencing. *Nat. Rev. Genet.* 16, 553–559. doi: 10.1038/nrg3960

Kwoka, M. B. (2017). First-Person FOIA. *Yale Law J.* 127, 2204–2269.

Lansdown, G. (2005). *The Evolving Capacities of the Child*. Florence: UNICEF Office of Research–Innocenti.

McGuire, A. L., Pereira, S., Gutierrez, A. M., and Majumder, M. A. (2020). "Ethics in Genetic and Genomic Research," in *Ethical Issues in Pediatric Hematology/Oncology*, eds K. A. Mazur and S. L. Berg (Cham: Springer International Publishing), 91–110.

Narayanasamy, S., Markina, V., Thorogood, A., Blazkova, A., Shabani, M., Knoppers, B. M., et al. (2020). Genomic sequencing capacity, data retention, and personal access to raw data in Europe. *Front. Genet.* 11:303. doi: 10.3389/fgene.2020.00303

Office of the Privacy Commissioner of Canada (2015). *The Application of PIPEDA to Municipalities, Universities, Schools, and Hospitals*. Available online at: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/r_o_p/02_05_d_25/ (accessed February 15, 2020).

Ogbogu, U., Burningham, S., and Caulfield, T. (2014). The right to control and access genetic research information: does mcinerney offer a way out of the consent/withdrawal conundrum. *UBC Law Rev.* 275–292. doi: 10.7939/R38K75B2C

Parker, S. (1994). The best interests of the child–principles and problems. *Int. J. Law Fam.* 8, 26–41. doi: 10.1093/lawfam/8.1.26

Perun, H., Orr, M., Dimitriadis, F., and Krever, H. (2005). "Substitute Decision-Making," in *Guide to the Ontario Personal Health Information Protection Act* (Toronto, ON: Irwin Law Inc.). Available online at: http://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=3317134 (accessed October 27, 2019).

Presidential Commission for the Study of Bioethical Issues (2012). *Privacy and Progress in Whole Genome Sequencing*. Available online at: https://heinonline.org/HOL/P?h=hein.prescomm/prescommaaahp0001andi=2 (accessed February 14, 2020).

Ries, N. M. (2010). Research participants' rights to access information about themselves held by public research institutions. *Health Law Rev.* 18, 5–14.

Savirimuthu, J. (2019). Datafication as parenthesis: reconceptualising the best interests of the child principle in data protection law. *Int. Rev. Law Comput. Technol.* 34, 310–341. doi: 10.1080/13600869.2019.1590926

Schickhardt, C., Fleischer, H., and Winkler, E. C. (2020). Do patients and research subjects have a right to receive their genomic raw data? an ethical and legal analysis. *BMC Med. Ethics* 21:7. doi: 10.1186/s12910-020-0446-y

Schwarz, K., Sisk, B., Schreiber, J., and Malik, F. (2015). A common thread: pediatric advocacy training. *Pediatrics* 135, 7–9. doi: 10.1542/peds.2014-2675

Sénécal, K., Rahimzadeh, V., Knoppers, B. M., Fernandez, C. V., Avard, D., and Sinnett, D. (2015a). Statement of principles on the return of research results and incidental findings in paediatric research: a multi-site consultative process. *Genome* 58, 541–548. doi: 10.1139/gen-2015-0092

Sénécal, K., Vears, D. F., Bertier, G., Knoppers, B. M., and Borry, P. (2015b). Genome-based newborn screening: a conceptual analysis of the best interests of the child standard. *Pers. Med.* 12, 439–441. doi: 10.2217/pme.15.28

Shyr, D., and Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biol. Proced. Online* 15:4. doi: 10.1186/1480-9222-15-4

Steinberg, S. B. (2016). Sharenting: children's privacy in the age of social media. *Emory Law J.* 66, 839–884.

Thorogood, A., Bobe, J., Prainsack, B., Middleton, A., Scott, E., Nelson, S., et al. (2018). APPLaUD: access for patients and participants to individual level uninterpreted genomic data. *Hum. Genomics* 12:7. doi: 10.1186/s40246-018-0139-5

Tobin, J., and Varadan, S. (2019). "Art. 5 the right to parental direction and guidance consistent with a child's evolving capacities," in *The UN Convention on the Rights of the Child: a commentary Oxford commentaries on international law*, ed J. Tobin (Oxford: Oxford University Press) 159–185.

United Nations General Assembly (1989). *Convention on the Rights of the Child*. GA Res. 44/25, UN GAOR, 44th Sess., UN Doc. A/RES/44/25.

United Nations General Assembly (2007). Convention on the Rights of Persons with Disabilities. https://www.google.com/search?sxsrf=

ALeKk006PkDX3ZtieBwZlabAdsx7PBj3Dg:1615564992511 New York, NY: United Nations.

van El, C. G., Cornel, M. C., Borry, P., Hastings, R. J., Fellmann, F., Hodgson, S. V., et al. (2013). Whole-genome sequencing in health care. *Eur. J. Hum. Genet.* 21, 580–584. doi: 10.1038/ejhg.2013.46

Vaxillaire, M., and Froguel, P. (2016). Monogenic diabetes: implementation of translational genomic research towards precision medicine. *J. Diabetes* 8, 782–795. doi: 10.1111/1753-0407.12446

Vears, D. F., Sénécal, K., Clarke, A. J., Jackson, L., Laberge, A. M., Lovrecic, L., et al. (2018). Points to consider for laboratories reporting results from diagnostic genomic sequencing. *Eur. J. Hum. Genet.* 26, 36–43. doi: 10.1038/s41431-017-0043-9

Wilson, G. R., Sunley, J., Smith, K. R., Pope, K., Bromhead, C. J., Fitzpatrick, E., et al. (2014). Mutations in SH3PXD2B cause Borrone dermato-cardio-skeletal syndrome. *Eur. J. Hum. Genet.* 22, 741–747. doi: 10.1038/ejhg.2013.229

Winkler, E., Idler, I., Beck, K., Brors, B., Cornelius, K., Dikow, N., et al. (2019). *Stellungnahme zur Herausgabe genomischer rohdaten an Patient_innen und Studienteilnehmende*. Heidelberg: EURAT–Ethische und rechtliche Aspekte der Totalsequenzierung des menschlichen Genoms.

Wright, C. F., FitzPatrick, D. R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* 19, 253–268. doi: 10.1038/nrg.2017.116

Zawati, M. H., Parry, D., and Knoppers, B. M. (2014). The best interests of the child and the return of results in genetic research: international comparative perspectives. *BMC Med. Ethics* 15:72. doi: 10.1186/1472-6939-15-72

Conflict of Interest: The 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.

Copyright © 2021 Beauvais, Thorogood, Szego, Sénécal, Zawati and Knoppers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership