PLOS COMPUTATIONAL BIOLOGY

EDUCATION

A FAIR guide for data providers to maximise sharing of human genomic data

Manuel Corpas¤*, Nadezda V. Kovalevskaya, Amanda McMurray, Fiona G. G. Nielsen

Repositive Ltd, Betjeman House, Cambridge, United Kingdom

¤ Current address: Cambridge Precision Medicine Ltd., Future Business Centre, Cambridge, United Kingdom
* manuel@cambridgeprecisionmedicine.com

Abstract

It is generally acknowledged that, for reproducibility and progress of human genomic research, data sharing is critical. For every sharing transaction, a successful data exchange is produced between a data consumer and a data provider. Providers of human genomic data (e.g., publicly or privately funded repositories and data archives) fulfil their social contract with data donors when their shareable data conforms to FAIR (findable, accessible, interoperable, reusable) principles. Based on our experiences via Repositive (https://repositive.io), a leading discovery platform cataloguing all shared human genomic datasets, we propose guidelines for data providers wishing to maximise their shared data's FAIRness.

Citation: Corpas M, Kovalevskaya NV, McMurray A, Nielsen FGG (2018) A FAIR guide for data providers to maximise sharing of human genomic data. PLoS Comput Biol 14(3): e1005873. https://doi.org/10.1371/journal.pcbi.1005873

Editor: Francis Ouellette, Genome Quebec, CANADA

Published: March 15, 2018

Funding: Authors were supported by Repositive at the time of writing this article. The company had a role in the decision to publish and the preparation of the manuscript.

Competing interests: I have read the journal's policy and have the following conflict: at the time of writing, MC, NVK, and FGGN are employees of Repositive Ltd.

Introduction

Making research data available for reuse is an essential component for repeatable research [1]. Sharing data generated through publicly funded projects maximises return on investment and increases the likelihood of obtaining funding in future rounds [2]. Genomic data of human origin, when adequately shared, constitutes a direct measure of the current advancement in risk prediction, diagnosis, and treatment of genomic disorders [3]. Not only does human genome data have value to the individual, it is also of value to biological relatives of the individual, as well as to the wider research community, particularly when clinically actionable [4].

Sharing of human genomic data by researchers is governed by both legal and implicit obligations. Legal obligations include responsibilities and liabilities to protect the confidentiality and privacy of research participants, who, as data donors, intend and expect their data to be reused [5]. This is what is commonly referred to as the implicit 'social contract', which must be taken into account when developing the governance mechanisms around research participants. Although expectations from the social contract are implemented differently depending on local jurisdictions, for a concerned human genomics data generator (e.g., a researcher), it is important to be aware of local and international governmental regulations in place affecting the individual's genetic data. In the United Kingdom, for example, there is currently a moratorium on the use of an individual's genetic data for life insurance purposes [6]. In the United States, the Genetic Information Nondiscrimination Act (GINA) prevents employers from requesting, requiring, or purchasing the genetic information of their employees. GINA also

prohibits health insurers acquiring genetic information for underwriting purposes and prior to enrolment [7]. In the European Union, there is no legislation specific to genetic information, although genetic data pertaining to health is considered 'sensitive data', and discrimination based on genetic features is prohibited [8]. This disparity among jurisdictions influences governance models for data providers who wish to maximise their data sharing in a FAIR (findable, accessible, interoperable, reusable) manner.

Sharing of human genomic data with external collaborators usually requires formal agreements and compliance with institutional review board (IRB) rules. An IRB may request the establishment of a data access committee (DAC) to regulate access to the data and define acceptable (re)use conditions. For researchers funded by international public funding bodies (e.g., National Institutes of Health [NIH] [9], Cancer Research UK [9,10], Wellcome Trust [11], Medical Research Council [MRC] [12], and others), it is common for investigators to share human genomic data broadly for secondary research purposes, in all cases, consistent with applicable laws, regulations, and policies.

Human genomic datasets are often accompanied by clinical phenotypes and other sensitive metadata, including pictures, medical history, sex, age, etc., building a picture of the patient to facilitate diagnostics and therapies. Despite a single genetic mutation datum itself not being a threat to the individual's privacy, if whole genome data for an individual is made publicly available, removing direct identifiers (name, date of birth, etc.) may not be enough to conceal the identity of the individual. According to Homer et al. [13], it is straightforward to assess the probability that a person or relative participated in a study, especially if phenotype and clinical metadata are also available. However, the risk of re-identification may be mitigated. For example, Genomics England (GeL) [14] provides protected access (allowing authorised data users to access the de-identified data within the system) and enables export of only completely anonymised results. The consent framework implemented by GeL is thus in place for data to be accessible only to authorised users. Such solutions may work for specific research scenarios, yet commonly, researchers may require complete access to the research data.

Even when privacy risks are managed, an underlying problem may still remain: the systemic lack of standardised protocols for secure interoperability of genomic data globally. It is in particular this problem, combined with the lack of interoperable protocols and an increasing awareness of the need for human genomic data sharing, that led to the establishment of the Global Alliance for Genomics and Health (GA4GH) [15]. FAIR principles have been embraced by GA4GH and the community in general, providing a framework for data-sharing infrastructures [16]. FAIR principles are ideally suited to data repositories developing specialised strategies to facilitate the sharing of clinical data [17]. Examples of such data repositories include the European Genome-phenome Archive (EGA) [18] and the database of Genotypes and Phenotypes (dbGaP) [19]. Both store patient data of genetic and phenotypic origin, for which the patient has consented to reutilisation, approved for predetermined research uses via controlled data access. Specialised data journals such as *Scientific Data* [20], *GigaScience* [18], and *Human Genome Variation* [19] may also enforce best practices for publishing data whilst providing an incentive for researchers to share their data via a data paper.

Here, we propose five tips for providers of human genomic data wishing to use FAIR principles as a context of reference. We acknowledge that the act of sharing is a two-way process: the data producer may delegate the provision of the data to a trusted repository (data provider), where a data consumer finds and accesses the shared data. Our focus on human genomic data sharing from the data provider's perspective originates as a consequence of developing Repositive [21], a global catalogue of human genomic data and metadata from data archives and repositories. Our mission and ongoing work to collate and connect the global landscape of data sources and datasets for genomics through an intuitive platform like

Repositive has given us insight into common practices, enabling us to contribute to the discussion on how to maximise the FAIRness of shared human genomic datasets.

This is a *PLOS Computational Biology* Education paper.

Tip 1: Establish a FAIR-aware patient consent framework

Consent frameworks dictate the extent to which human genomic data can be accessed and reused. Ensuring appropriate consent to collect genotype, phenotype, and any other type of human data is achieved will usually be the responsibility of the principal investigator (PI) overseeing the study. Data archives and repositories will be required to check that the consent forms of deposited datasets specify the goals of the immediate project. It is essential to explicitly describe in clear terms if the data is intended to be shared beyond the current scope of the project (i.e., general research use). If wider data sharing is intended, the consent form should set out potential risks and benefits to participants, as well as any data anonymisation procedures to be undertaken. Consent frameworks require special considerations from the data producer's point of view, given their extreme variability. To allow standardisation of consent frameworks, GA4GH has developed consent codes that facilitate the integration of distinct consent types across different legal systems [22].

The level of anonymisation that will be applied to the data should be clearly explained in consent forms, since different levels are possible. Participant consent requirements should be considered prior to data collection, alongside approval from an IRB. Different research questions may necessitate variable degrees of identity exposure by study participants. For example, the Personal Genomes Project (PGP) provides complete access to study participants' identities and phenotypic traits [23] under a Creative Commons Zero (CC0) license waiver [24]. This radically open consent framework is, however, a highly unusual one for clinical genomic data. NIH-funded studies require third-party researchers to submit a Data Access Request describing how they intend to use the data. A Data Use Certification Agreement is then produced, which must adhere to the NIH Genomic Data Sharing Policy's ethical principles governing data access and privacy safeguards [25]. In the UK, GeL consent forms are classified according to whether patients are affected with cancer or rare diseases, with the consent framework allowing access to summary statistics in a controlled environment to authorised users [26].

Patient data sharing consent frameworks vary country to country, funder to funder, and study to study. We thus suggest that, for interoperability purposes, data sharing consent frameworks adopt existing standards for digital consent formats and include, at a minimum:

1. Goals of the current research project and why data generation/sharing is being carried out.
2. Potential risks to the individual participant from the (ab)use of the data.
3. Confirmation that these issues have been discussed in person, with the individual and/or guardian involved in signing the form.
4. Contract of data access for the current research project and the extent to which the data custodian commits to make the data findable, accessible, interoperable, and reusable for future research projects.

Some consent forms (e.g., PGP or Genomes Unzipped [27]) may make the patient/donor's identity known. Others require the identity of research participants anonymised. The Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) database [28], a provider of anonymised human copy number variation (CNV) data

and phenotypes (not datasets), offers a consent framework compliant with European Union guidelines for clinical sharing [29], allowing anonymous sharing of genomic and phenotypic data of patients. At all events, it is always advised that ethical and genetic counselling experts are consulted when choosing the appropriate consent form.

Tip 2: Define FAIR data types specifying their intended uses and limitations

Providing sufficient information about the type of data being shared is fundamental for maximising data reuse. Deciding on their shared data and metadata descriptions also affects how datasets are found, accessed, and interoperated. Data consumers need to be able to find in the metadata what format the data is in and what size the files are as well as its provenance. It is also important to clearly define the technologies the data originated from, the experimental conditions, and any limitations as to how the data can be reused to ensure compliance with participants' original consent forms. Phenotype or clinical history data is also essential for generating research outcomes. It may include controlled vocabularies or extensive free text. A number of controlled vocabularies such as the Human Phenotype Ontology [30] facilitate phenotypic annotation but offer no guarantee of having all needed fine-grained detail. Clinical history and further vital measurements may also vary according to study, instrument, or clinical need. Hence, extreme care must be ensured in establishing the procedure with which the data itself is to be transferred, e.g., 'pretty good privacy' encryption [31] or Aspera [32]. The choice for data transfer will be greatly influenced by the characteristics of the dataset, the consent framework, the amount of data to be accessed, and the repository where the data is stored.

For human genome–based data, it is important to make a distinction between raw and processed data types. Raw sequencing data must be processed before it can be interpreted. The processing of raw data is usually dependent on the software and parameters chosen to create interpretable data (e.g., variation calls). The choices of both software and parameters for processing raw data (and its intermediary files) are deeply influenced by the research questions being tested. Being able to capture the processing methodology in the metadata descriptions may be crucial for some experiments. Sometimes, however, researchers may choose not to redo the processing steps and simply reuse the interpretable data, but this might not be out of choice: the size of raw reads from whole human genome experiments can be prohibitively voluminous, depending on the coverage of the run. Therefore, the size of both the raw and processed data files to be shared will impact on the ease with which they can be reused. For example, variant call format (VCF) processed files [33] are much 'lighter' in storage footprint than binary alignment/map format (BAM) files [34] (at the cost of losing some information), so it may be better to provide Fastq files as raw data files from which BAM and VCF files can be derived.

Tip 3: Maximise machine-readable data and metadata findability and interoperability

Maximising the likelihood for data to be found is a vital component of the data sharing process. For this, capturing of health/clinical data with complete, coherent, and standard descriptions is critical. The richness, granularity, and compliance to standards with which metadata descriptors are captured are determinant in influencing the user's ability to reuse and draw any value from the dataset. A good template that incorporates specific pathophenotypic descriptions and patient annotations such as health constants, smoking status, clinical history, etc., is the PGP-Harvard data collection [23] and its accompanying raw and processed human genome data. This ideal level of data and metadata capture may not be attained by the

sometimes constrained experimental conditions, data access options, or consent frameworks. The investigation, study, and assay (ISA) modelling tool may help guide experimental metadata collection [35] using the BioSharing (now FAIRsharing) catalogue of known standards [36]. The use of the minimum information about a microarray experiment (MIAME) standard for microarray data [37], for example, increases the discoverability of microarray data in MIAME-compliant repositories such as National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) [38] and European Bioinformatics Institute's (EBI) ArrayExpress [39].

Despite rich, standard metadata capture increasing the interoperability of datasets in any given repository, there is the additional problem of an increasing number of data repositories existing around the world. Thus, a number of metadata catalogues have been created to increase the discoverability of datasets. These catalogues include Repositive [40], DataMed [41], and OmicsDI [42], all with different aims and scope. Such meta-indexing solutions make it even more critical to standardise metadata descriptions.

Apart from (a) the great hurdle of scattered data sources, (b) different regulatory frameworks, and (c) heterogeneity of data types, there is the problem of how to incentivise scientists to contribute to the best possible standard of metadata annotation. Arguably, producers/generators of data are disincentivised to share as much data as possible in a standard, coherent, and complete manner, as the perceived risks are high and their rewards may be few. Making the data shareable means that the producers of data can be scrutinised by the community. Allowing the data to be discoverable, accessible, and reusable may also increase the risk that others will reap the reward of the effort of making the data shareable. To counteract these disincentives, data producers may allow early access to their data if they retain the privilege of publishing it first. Data papers may also increase the findability of the dataset, since the article will be indexed in bibliographic databases such as PubMed, where more discoverability of the dataset can be attained. Publishing discoverable, citable data increases the amount of citations for scientific output, resulting in greater incentive for FAIR data publishing as a measure of acknowledgment for work and scientific merit [43].

It is worthwhile to plan for ways in which the data itself can be made discoverable, as well as considering both the human and machine accessibility of the dataset. For example, it was recently shown that human gene symbols were converted to dates in the supplementary data files of some published papers [37], which meant these gene symbols were not machine readable. This is an issue particularly when data is not part of the manuscript review process in a publishing context, i.e., data validation checks are not in place (as opposed to automated/manual checks in databases). Simple strategies can be used to avoid these errors, such as including data units in tables and keeping data types consistent across columns or rows to avoid mixing of strings with numbers. It is best to avoid the use of acronyms where possible and to make sure they are defined if their use is unavoidable.

It is essential to ensure human genomic data is shared in a citable way. Data citations' growing importance as a way to incentivise FAIR data sharing is being attested by the way in which researchers can gain recognition for making data available as well as providing provenance for it. A FAIR-aware data repository will enable data to be cited by providing a persistent and unique identifier for each data archive so that large-scale data interoperability is attained [44]. The main NCBI and EBI databases use accession identifiers, and other repositories may use DataCite's Digital Object Identifiers (DOIs). Both accession IDs and DOIs can be cited in scholarly works, as in the guidance *Scientific Data* provides to its authors on how to cite data [45]. Similarly, scholarly works associated with a dataset should be referenced in the uniquely identified data record in the data archive: e.g., PubMed IDs might be added to the sequencing studies using either the interactive or programmatic route [46].

Tip 4: Choose the most findable and accessible genomic data repository

The dataset type may impact on the type of repository that can be used to share the data. For example, the generalist repository figshare offers single file uploads of 5 terabytes per file but does not support controlled access to sensitive data. Using specialist data repositories for human genomic data may help ensure that this data is archived and preserved in a data type–specific way. For example, array-based human data would usually be submitted to repositories such as GEO [38] or ArrayExpress [46], while raw sequence data should usually go to repositories such as the Sequence Read Archive (SRA) [40] or the European Nucleotide Archive (ENA) [41]. Both SRA and ENA also store aligned data and data analysis (e.g., genome assemblies, taxonomic and gene class, etc.). For clinical genomic dataset deposition, the European Genome/phenome Archive (EGA) and the NCBI equivalent database of Genotypes and Phenotypes (dbGaP) are well-recognised controlled-access data archives. Both resources allow submission of sequencing, array-based data, and phenotypes as well. Care should be taken to archive controlled-access data in repositories that have workflows in place to ensure data access is only given to those requestors that fulfil the relevant consent requirements. Several funders have published lists of recommended repositories for specific types of research output (e.g., Wellcome Trust [47] and NIH [48]).

Tip 5: Set FAIR data access governance

Data access governance is in great measure influenced by the consent framework (see Tip 1), the applicable jurisdiction, and the experimental design. The implementation of the data access policy will also be influenced by the technical strategy set in place. We expect that both technical and ethical/legal implementations will continue to evolve as types of human genomic data and their characteristics continue to change. Thus, a flexible approach is much needed. We turn again to GA4GH as a good guide for researchers and clinicians in choosing the right policy and technical implementation. We currently envisage a spectrum of access [49]. At one extreme, we have the open access approach with complete disclosure of the individual's identity as exemplified by the PGP. At the other extreme, we have DAC-regulated access, in which access to data is subject to a contract between the user and the DAC, to be reviewed by the DAC and granted only if approved. In the case of dbGaP [50], the contract is signed with the US government, while for EGA, the contract is signed with the Wellcome Trust Sanger Institute. Both EGA and dbGaP act as the conduits that allow contract exchange implementation via their respective platforms. The access to the data is ultimately granted by the DAC. Access standards are not coordinated between international institutions, thus creating a huge overhead burden when data consumers require access from different studies across disparate data sources. There is at least (through GA4GH) an effort to facilitate the mutual recognition of independent separate DACs to save having to apply separately to multiple DACs when needing to access data from separate studies or independent sources.

The benefit of having regulated access to data via DACs is evident. With a DAC, every access request is evaluated against the consent given by patients and individual data donors, and the access to this data is provided only to intended recipients. The flip side of this is that undergoing the whole application process from dataset identification to data access can take time. Intermediary implementations of regulated access have been developed independently to avoid wasted time and effort applying to a DAC, reducing the likelihood of requesting access to the wrong type of dataset. GA4GH, for example, has developed the Beacon project, which allows standard programmatic querying of distributed sources for presence or absence of genetic features, given a dataset. The dbGaP Data Browser [51] has minimised the process of viewing general research use (GRU) data (currently 13% of all dbGaP subjects) to take less

than 2 weeks. The dgGaP Data Browser reduces the number of unnecessary data downloads, allowing researchers to assess patient data before downloading it while decreasing chances for this data to be abused. Gaining download access to dbGaP data, however, still requires the submission of a dbGaP Project Request approval for each dataset [51].

Establishing the governance of access to shared datasets requires the awareness that once the data is acquired by the user, there is no easy way to track the usage of data by the data receiver. In any case, the DAC can always be utilised as the point of contact should there be need for consent frameworks to be modified given the evolving nature of research questions. Similarly, Data Access Agreements specify the custodianship of the data as well as the exceptional requirements of reporting incidental findings or what to do when inadvertently identifying an individual.

Conclusion

Multiple funders and experts in data curation agree that sharing of personal health-related data must be planned from the start of the research project in order for it to be FAIR. Whenever it is possible to anonymise research data, this is the advised procedure for data producers to follow before data is shared. For data that has not been consented for open access, additional governance procedures for data access need to be established. For a compelling overview of all aspects to do with human genome data sharing, we direct our readers to [52].

This work contextualises current best practices for data providers assuming the role of dissemination agents for data producers. We specify that, in every sharing transaction of human genomic data, both a data consumer and a data provider are involved in establishing a secure data exchange. We embrace the FAIR data sharing principles described by Wilkinson et al. [16] and apply them to our particular 'data provider' context, which we have worked on as part of our wider efforts to catalogue the human genomic data landscape via the Repositive platform [21]. As precision medicine starts to impact patient lives, it is expected that sharing of datasets containing potentially sensitive information will become more widespread. Hence, having a set of guiding tips that help keep patient genomic data reusable whilst complying with consent frameworks is crucial if we are to leverage the power of FAIR principles to realise the promise of better diagnostics and more personalised therapies.

Acknowledgments

We are grateful to Varsha Khodiyar and Petra Ten Hoopen for valuable comments on the manuscript. We also thank Repositive employees Charlotte Whicher and Tom Byers for their feedback. We are grateful to the public repositories that make it possible for scientists to upload their genomic datasets at no cost.

References

1. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009; 41: 149–155. https://doi.org/10.1038/ng.295 PMID: 19174838
2. van Schaik TA, Kovalevskaya NV, Protopapas E, Wahid H, Nielsen FGG. The need to redefine genomic data sharing: A focus on data accessibility. Appl Transl Genom. 2014; 3: 100–104. https://doi.org/10.1016/j.atg.2014.09.013 PMID: 27294022
3. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014; 15: 409–421. https://doi.org/10.1038/nrg3723 PMID: 24805122
4. Website [Internet]. [cited 5 Jun 2017]. Available from: http://blogs.nature.com/scientificdata/2016/05/13/enabling-the-effective-sharing-of-clinical-data/
5. W-C. Open Access Science | Sanger Institute [Internet]. [cited 5 Jun 2017]. Available from: http://www.sanger.ac.uk/about/who-we-are/policies/open-access-science

6. Agreement extended on predictive genetic tests and insurance—GOV.UK [Internet]. [cited 5 Jun 2017]. Available from: https://www.gov.uk/government/publications/agreement-extended-on-predictive-genetic-tests-and-insurance
7. Website [Internet]. [cited 19 Aug 2017]. Available from: http://blogs.harvard.edu/billofhealth/2017/03/15/will-the-recent-workplace-wellness-bill-really-undermine-employee-health-privacy/
8. Soini S. Genetic testing legislation in Western Europe—a fluctuating regulatory target. J Community Genet. 2012; 3: 143–153.
9. NOT-OD-14-124: NIH Genomic Data Sharing Policy [Internet]. [cited 5 Jun 2017]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html
10. Data sharing guidelines. In: Cancer Research UK [Internet]. 21 Mar 2014 [cited 5 Jun 2017]. Available from: http://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines
11. Policy on data management and sharing | Wellcome [Internet]. [cited 5 Jun 2017]. Available from: https://wellcome.ac.uk/funding/managing-grant/policy-data-management-and-sharing
12. Website [Internet]. [cited 27 Feb 2018]. Available from: https://www.mrc.ac.uk/research/policies-and-guidance-for-researchers/data-sharing/
13. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008; 4: e1000167. https://doi.org/10.1371/journal.pgen.1000167 PMID: 18769715
14. Current research | Genomics England. In: Genomics England [Internet]. 15 Jan 2016 [cited 3 Jul 2017]. Available from: https://www.genomicsengland.co.uk/the-100000-genomes-project/data/current-research/
15. Mission & Founding Principles | Global Alliance for Genomics and Health [Internet]. [cited 15 Nov 2017]. Available from: https://www.ga4gh.org/aboutus/
16. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3: 160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
17. Hrynaszkiewicz I, Khodiyar V, Hufton AL, Sansone S-A. Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations. Research Integrity and Peer Review. 2016; 1. https://doi.org/10.1186/s41073-016-0015-6 PMID: 29451541
18. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015; 47: 692–695. https://doi.org/10.1038/ng.3312 PMID: 26111507
19. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014; 42: D975–9. https://doi.org/10.1093/nar/gkt1211 PMID: 24297256
20. Scientific Data [Internet]. 30 May 2017 [cited 5 Jun 2017]. Available from: http://www.nature.com/sdata/
21. One-click access to human genomic data | Repositive [Internet]. [cited 13 Jul 2017]. Available from: https://repositive.io
22. Dyke SOM, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent Codes: Upholding Standard Data Use Conditions. PLoS Genet 2016; 12: e1005772. https://doi.org/10.1371/journal.pgen.1005772 PMID: 26796797
23. PersonalGenomes.org. Personal Genome Project: Harvard Medical School [Internet]. [cited 27 Feb 2018]. Available from: https://pgp.med.harvard.edu/
24. CC0 [Internet]. [cited 8 Jun 2017]. Available from: https://creativecommons.org/choose/zero/
25. GENOMIC DATA SHARING (GDS) [Internet]. [cited 27 Feb 2018]. Available from: https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/
26. Website [Internet]. [cited 8 Jun 2017]. Available from: https://www.genomicsengland.co.uk/taking-part/patient-information-sheets-and-consent-forms/
27. Author G, MacArthur D, Wright C, Pickrell J. Genomes Unzipped [Internet]. [cited 10 Jul 2017]. Available from: http://genomesunzipped.org/
28. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. Cell Press; 2009; 84: 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010 PMID: 19344873
29. About—DECIPHER v9.15 [Internet]. [cited 6 Jun 2017]. Available from: https://decipher.sanger.ac.uk/about#downloads/documents

30. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017; 45: D865–D876. https://doi.org/10.1093/nar/gkw1039 PMID: 27899602

31. Pretty Good Privacy—Wikipedia [Internet]. [cited 8 Jun 2017]. Available from: https://en.wikipedia.org/wiki/Pretty_Good_Privacy

32. Aspera High-Speed File Transfer Software [Internet]. [cited 8 Jun 2017]. Available from: http://asperasoft.com/

33. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

35. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010; 26: 2354–2356. https://doi.org/10.1093/bioinformatics/btq415 PMID: 20679334

36. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database. 2016; 2016. https://doi.org/10.1093/database/baw075 PMID: 27189610

37. Brazma A. Minimum Information About a Microarray Experiment (MIAME)–Successes, Failures, Challenges. The Scientific World JOURNAL. 2009; 9: 420–423. https://doi.org/10.1100/tsw.2009.57 PMID: 19484163

38. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2017; 45: D12–D17. https://doi.org/10.1093/nar/gkw1071 PMID: 27899561

39. Brazma A, Parkinson H. ArrayExpress service for reviewers/editors of DNA microarray papers. Nat Biotechnol. 2006; 24: 1321–1322. https://doi.org/10.1038/nbt1106-1321 PMID: 17093465

40. Kovalevskaya NV, Whicher C, Richardson TD, Smith C, Grajciarova J, Cardama X, et al. DNAdigest and Repositive: Connecting the World of Genomic Data. PLoS Biol 2016; 14: e1002418. https://doi.org/10.1371/journal.pbio.1002418 PMID: 27011302

41. Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, et al. Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet. 2017; 49: 816–819. https://doi.org/10.1038/ng.3864 PMID: 28546571

42. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017; 35: 406–409. https://doi.org/10.1038/nbt.3790 PMID: 28486464

43. Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2 2007;(3): e308. https://doi.org/10.1371/journal.pone.0000308 PMID: 17375194

44. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol 2017; 15: e2001414. https://doi.org/10.1371/journal.pbio.2001414 PMID: 28662064

45. Guide for Authors | Scientific Data [Internet] [cited 20 Nov 2017] Available from: https://www.nature.com/sdata/publish/for-authors

46. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015; 43: D1113–6. https://doi.org/10.1093/nar/gku1057 PMID: 25361974

47. Data repositories and database resources | Wellcome Trust [Internet] [cited 20 Nov 2017] Available from: https://wellcome.ac.uk/funding/managing-grant/data-repositories-and-database-resources

48. NIH Data Sharing Repositories [Internet] [cited 20 Nov 2017] Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

49. Paradise Or Inferno The Future Of Data Notes From The Uk Anonymisation Symposium [Internet] [cited 20 Nov 2017] Available from: http://labs.theodi.org/blog/2014/09/12/paradise-or-inferno-the-future-of-data-notes-from-the-uk-anonymisation-symposium/

50. Young M. How to successfully apply for access to dbGaP. In: Genomics & software development blog posts | Repositive [Internet]. 15 Mar 2016 [cited 12 Jul 2017]. Available from: https://blog.repositive.io/how-to-successfully-apply-for-access-to-dbgap/

51. Wong KM, Langlais K, Tobias GS, Fletcher-Hoppe C, Krasnewich D, Leeds HS, et al. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. Nucleic Acids Res. 2017; 45: D819–D826. https://doi.org/10.1093/nar/gkw1139 PMID: 27899644

52. Data Sharing 101 | University of Leicester's Department of Genetics and Genome Biology [Internet] [cited 21 Nov 2017] Available from: https://datasharing-101.le.ac.uk/DataSharing_101/