Genomics England Airlock Policy

Document Key/Version Number: RES-AP-001

Function	Science		**Version**	1.1
Document Key	RES-AP-001			
Document Owner	Simon Thompson	*Data Wrangler*	**Review Date**	31/07/2019
Document Author and Job Title	Simon Thompson	*Data Wrangler*	**Status**	Draft ☐ Live ☒ Archive ☐
Document Reviewers and Job Title	*Anna Need* *James Holman*	*GeCIP Lead Research* *Environment Project Lead*	**Effective Date**	17/07/2018
Document Approvers and Job Title	*Tom Fowler*		*Deputy Chief Scientist*	
Electronic Signature	Tom Fowler Tom Fowler (Jul 31, 2018)		**Date Approved**	Jul 31, 2018
Transaction Number	Adobe Sign Transaction Number: CBJCHBCAABAA_7cEBeKmun7f6Q0LfQ3_Qm_o8gzdh0ld			

Contents

1. Scope and applicability

The following policy covers acceptable uses of the airlock for those conducting research on the 100,000 Genomes Project using de-identified data inside the Research Environment. Airlock application is not required for export of data to be used in clinical care (i.e. export of data by GMCs), such applications can be made to the Genomics England Bioinformatics team . Furthermore, Airlock application is not required for import or export (including whitelisting of websites) by Genomics England Service Providers for the purpose of service provision or service improvement. These requests can be made directly to Genomics England Platforms team (who may reject if they have security or support concerns).

Where individuals have access to both the identified data (for clinical care) and de-identified data (for research), for example GMC members of staff who are members of a GeCIP domain, it is imperative that a clear distinction is made between what activities are carried out on which version of the data. Genomics England will consider any cases of research being carried out on identified data to be a serious offense and will act accordingly.

Genomics England considers research to be any activity using the Dataset where the motivating purpose is eventual submission of that work for publication in the scientific literature, or presentation to the scientific community.

2. Definitions

100,000 Genomes Data - information associated with, or derived from, any individual's participation in the 100,000 Genomes Project (as defined by the accompanying data model), or associated with, or derived from, the samples they provide over the course of the project. This includes, but is not limited to:

- Clinical and personal data provided at registration, or as part of any longitudinal/surveillance monitoring specific to 100,000 Genomes Project;
- Data from any tests or procedures carried out on the participant's samples given at registration or during follow-up;
- Any and all 'omics data (including all genomic, transcriptomic, metabolomic, proteomic data) derived from the participant's samples.

Airlock – the process by which information and data is securely moved in or out of the Research Environment.

Airlock Review Team – a delegation of the Chief Scientist responsible for oversight of all airlock requests in accordance with the Airlock Policy and the group's Terms of Reference. It comprises:

- Senior Information Risk Office (SIRO)
- Technical Lead

- User Community Representative
- Bioinformatics Director
- Caldicott Guardian
- Chief Scientist representative

ARC approval – The current approval from the Genomics England Access Review Committee under which any analyses pertaining to an airlock request have been undertaken.

The Dataset - all data linked to all study participants, of all types and origins, accessible from the Research Environment i.e. the totality of 100,000 Genomes Data and External data.

Dataset Owner – The person Genomics England considers to be the nominated contact person for a particular piece of External data (e.g. a nominated individual from the study PIs). For some external data, this responsibility may be delegated to the Office of the Chief Scientist.

Docker container - a packaged software tool that contains all the requisite components required to run a certain process or pipeline isolated from the host environment. In this context, Docker containers will be used to import tools into the Research Environment that are not, and will not be, centrally provided, and so need to be run in a secure, contained manner.

External data - data present within the Research Environment that was not collected as part of the 100,000 Genomes Project. This includes, but is not limited to:

- additional data on 100,000 Genomes Project participants that has been collected as part of nationwide programmes to collect and collate population-wide datasets e.g. Hospital Episode Statistics;
- additional data on 100,000 Genomes Project participants that was collected as part of a separate research project, and has subsequently been made available in the Research Environment;
- data from public data repositories (e.g. 1000 Genomes Project) that has been made available within the Research Environment (either via direct data import, or via 'whitelisting' of web resources);
- data from restricted access datasets (e.g. dbGAP) that have been made available centrally within the Genomics England Datacentre;
- data from 'private' datasets, comparable to the 100,000 Genomes Project, that are within the Research Environment (either with restricted or unrestricted access).

Genomics England Publication Policy – the policy that all publications detailing analyses carried out on 100,000 Genomes Data must adhere to.

Identifiable data - Data that can be used on its own, or with other easily obtained information, to identify, contact, or locate a single person, or that might identify an individual within a given context.

Individual-level data - data, both raw and calculated, from the Dataset that is presented on a per-participant basis (regardless of whether the participant is anonymised or not), presented in written or graphical format.

Information - is distinct from data in this context in that information is not derived from participants in a research study (examples include a SNP list, chromosomal coordinates, or non-human cell line 'omics data).

Output checker – an individual who assesses statistical data within an export request for any risk of disclosure, following the principles described in the Airlock Policy Guidelines, and who will ultimately recommend to the ART whether the data should or should not be approved for export.

Pipeline - a linear sequence of specialized modules for performing complex analyses. In this context a pipeline can be run using data within the Research Environment, but does not explicitly contain any data within it.

Research - any activity using the Dataset where the motivating purpose is eventual submission of that work for publication in the scientific literature, or presentation to the scientific community, this must be in accordance with the Publication Moratorium (as detailed elsewhere).

Research Environment – The Genomics England analysis environment through which access to deidentified 100,000 Genomes Data is given.

Research Registry – The community-maintained catalogue of all research projects being undertaken on the Dataset.

Airlock Policy Guidelines – an evolving document overseen by the Airlock Review Team that describes the various likely data types that will submitted for airlock export, and the principles/rules-of-thumb that will guide the researcher during the preparation of their analyses, and the output checker during their review and decision-.

Summary data - is defined as data calculated from multiple observations across multiple participants or measurements within the Dataset, presented in written or graphical format. Graphical summaries presenting individual data points (e.g. scatter plots) would not be considered summary data.

Whitelisted website - a website to which access is possible from within the Research Environment (an environment that otherwise cannot access the internet).

3. Common policies for airlock

1. An airlock request will be assessed considering:
    a. its compatibility with the relevant ARC approval;
    b. its implications for the security of the Dataset;
    c. the risk for participant identification in light of its intended use and the nature of any agreements established between the applicant and Genomics England;
    d. the technical feasibility of the request;
    e. any costs Genomics England may incur by fulfilling the request;
    f. and, when importing, its value to the community of researchers within the Research Environment;
2. Genomics England will aim to review all requests within ten working days, any delays to review beyond this will be communicated to the applicant in writing, stating the reason for the delay and an expected decision date.
3. Airlock requests will not necessarily be processed in the order in which they are received. Succinct, clear, well-prepared requests consisting of a small number of items will be prioritised over more complex requests.
4. The decision of Genomics England will be communicated to the applicant within five working days of a decision being made.
5. Approved requests will aim to be actioned within five working days of a decision being made, though complex requests (e.g. import of software packages) may take longer.
6. Applicants will be expected to promptly reply to any questions or queries raised during the review process, or make any required changes to the material in a timely manner. Where this is not the case, Genomics England reserves the right to reject the request due to insufficient information being provided or lack of engagement from the applicant.
7. The applicant may appeal any final decision by Genomics England and should do so in writing to the Chief Scientist within 25 working days of the decision, clearly stating how they believe the decision of the Airlock Review Team contradicts the terms of this policy.
8. The Chief Scientist will consider all appeals and respond with a final decision within 25 working days of the receipt of the appeal.
9. Breach of the airlock (for example by 'screen-shots' of the Research Environment or copying data long hand), or contravention of the policies of the airlock (for example, using exported materials for uses other than those within the original request) will be considered a serious breach of the terms of data access. Genomics England reserves the right to ban the researcher's institution, and all their researchers, from access to the 100,000 Genomes Project dataset and pursue legal or criminal action against both the individual and the institution.

4. Policies for airlock export

10. It is expected that all analyses will be carried out within the Research Environment and that export will be of analysis results only. Export of data for use in further analysis (this does not include changing the format of the data) will only be permitted if that analysis is deemed necessary and cannot reasonably be carried out within the Research Environment.
11. You will comply with all reasonable requests associated with bringing data in and out of the research environment.
12. Graphical representations of summary data should be accompanied by the data represented i.e. any bar chart should be accompanied by the frequency table used to draw the graph. It is preferable that the data itself be exported and the creation of the graph be carried out outside of the Research Environment.
13. Binary files will not be considered for airlock export.
14. Any data for export must have been generated in line with the applicant's current ARC approval, and must have been generated as part of a specific project registered within the Research Registry.
15. Generally, only summary data will be eligible for export. Cells with counts lower than 5 will be scrutinized. Export of individual-level data will be considered for, but not limited to, instances in which:
    a. it forms part of a case study intended for publication within an academic journal;
    b. It forms part of a graphical summary of the data (e.g. a scatter plot) that is sufficiently complex that individual values for the data point are not discernible.
16. While Genomics England may allow the export of individual-level data on a case-by-case basis, it will be considered a breach of the airlock for researchers to use this process to reconstruct individual-level datasets external to the Research Environment.
17. Where individual-level data has been approved for export and Genomics England feels there is identifiable data within the export (either a single piece of data or a combination of pieces of data within the export) that cannot be avoided (e.g. the presence of necessary images/photos that identify the individual), export of the material will be dependent on express consent from the individuals affected..
18. Any export should contain the minimum set of information required to fulfil its purposes i.e. no extraneous variables or data points should be included.
19. Any export should not contain participant identifiers, or other labels that might be considered identifiable data.
20. Any export for the purposes of public sharing, e.g. presentation, publication, or other purpose covered by the Genomics England Publication Policy, must be submitted referencing a publication record within the Research Registry and be formatted in accordance with the publication policy.

21. Any data for export must have been prepared considering the relevant statistical disclosure control principles as outlined in the Genomics England Airlock Policy Guidelines. The review of statistical output will be conducted by output-checkers who will consider its suitability for export taking into account its disclosure risk given the intended use.

5. Exported material

a. cannot be shared with external commercial organisations (organisations that do not currently have any formal agreement with Genomics England) unless there is express permission from Genomics England;
b. cannot be distributed to media outlets unless there is express permission from Genomics England;
c. can be shared privately with individuals who have been approved for data access by Genomics England provided they are made aware of the sensitivity of the data and do not distribute further;
d. can only be shared publically if this was stated in the original airlock request;
e. can only be used for those purposes stated, and agreed to, in the original airlock request.

22. Any intellectual property contained within, or produced from, the exported material will, unless otherwise agreed, be wholly owned by Genomics England. For further details on the IP policy of the 100,000 Genomes Project see 'Intellectual Policies for 100,000 Genomes Project'
23. The Research Environment contains External Data (for example Hospital Episodes Statistics [HES]) which is subject to data sharing framework contracts and data sharing agreements between Genomics England and other parties that dictate how the data may be used and what can be exported. Where an export contains External Data, Genomics England will always apply the requirements placed on them as conditions of having access to the data. In some cases, particularly concerning the export of individual-level data, these will be more conservative than those applied to 100,000 Genomes Data alone.
24. Any export that features External Data will be brought to the attention of the Dataset Owner, who may, if raised in sufficient time, veto the export.

6. Policies for airlock import

25. Import of any data, tools, or other material will only be considered if its intended use is in-line with the applicant's ARC approval, the relevant ethical approvals and is part of a research project registered on the Research Registry.
26. Import of Docker containers will only be considered if it is not possible or practical for an equivalent pipeline or process to be established within the Research Environment.

27. Import of licensed software, whether within a Docker Container or otherwise, will only be considered if:
    a. there is no comparable software within the Research Environment;
    b. there is no equivalent open source software package available;
    c. it does not require excessive support, or maintenance by Genomics England, or does not require significant re-configuration of the Research Environment;
    d. a license is provided by the applicant that clearly supports its deployment into the Research Environment.
28. Import of additional data will be considered if
    a. it is clearly comparable to 100,000 Genomes Data;
    b. participant consent is consistent with its use in the Genomics England Datacentre.
29. Import of information will be considered if relevant to the research aims of the project and there is justification for its use.