EBiSC Guidance: Genetic / Genomic Data Deposition and Access

Version 0.1

2022-08-16

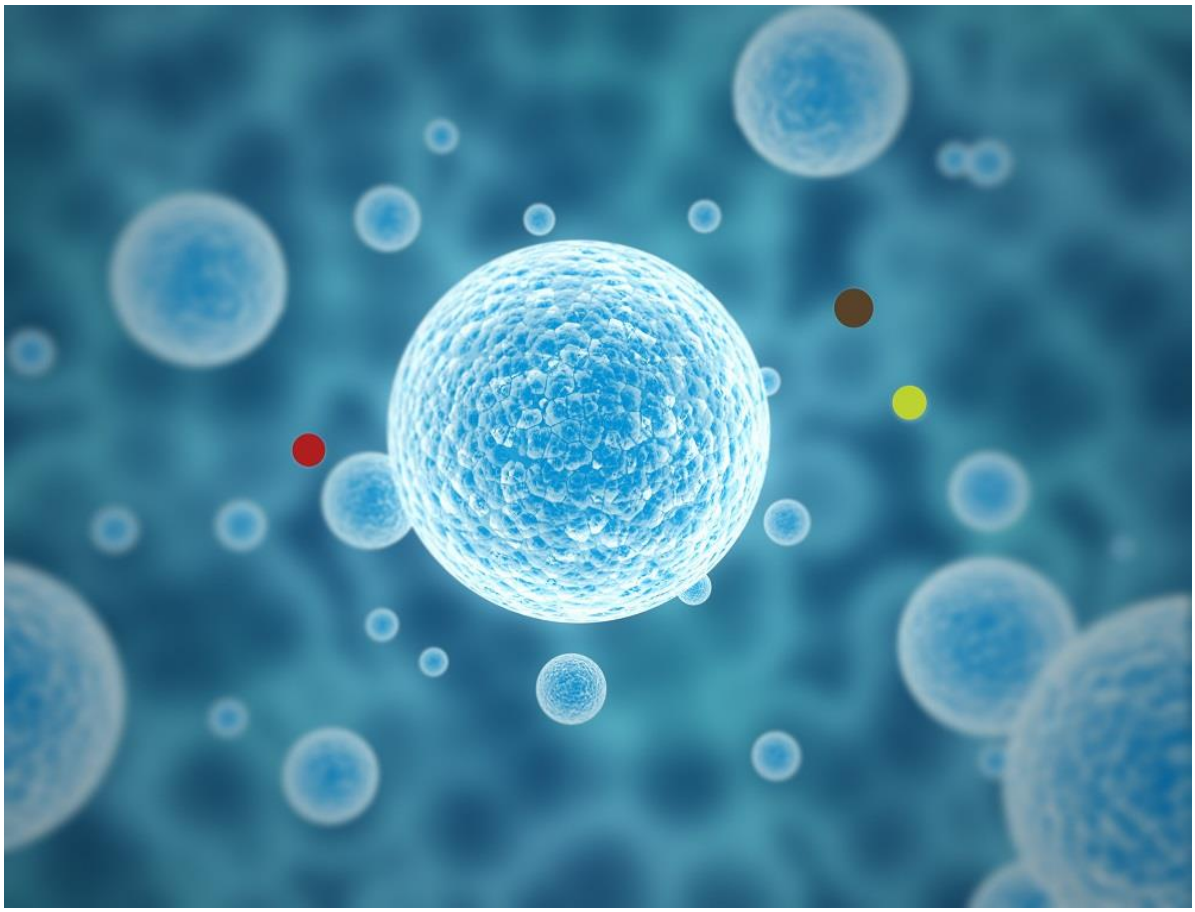

Table of Contents

1 Introduction

The European Bank for induced Pluripotent Stem Cells (EBiSC) is a not-for-profit iPS cell banking and distribution service enabling academic and commercial researchers to access quality-assured, disease relevant, research-grade iPSC lines, data and cell services (https://ebisc.org/). For a quick overview of the EBiSC deposition steps, please refer to the EBiSC webpage "Information for Depositors" (https://cells.ebisc.org/depositors/).

EBiSC is dedicated to supporting research through provision of a high-quality, well-characterised collection of human iPSC lines from a range of genetic backgrounds and reprogramming methods. Lines banked at EBiSC are associated with an extensive cell line data package. Genetic / genomic data are of particular added value to the cell lines. This guidance document describes how genetic / genomic data are captured and made accessible to the research community.

2 Deposition of genetic or genomic data associated to EBiSC cell lines

Genetic or genomic data associated to EBiSC cell lines is stored in two main resources. Low-throughput or low volume genetic data (e.g. STRs, karyotypes) can be entered as part of the cell line data package at the human pluripotent stem cell registry (hPSCreg®; hpscreg.eu), which is the official cell line data registry for all EBiSC lines. Access to low-throughput genetic data is managed according to its risk for re-identification of the donor (see Table 1). High-throughput (-omic) data, which is considered sensitive personal data (https://www.gdpreu.org/the-regulation/key-concepts/personal-data/), should be deposited in a public repository that supports controlled access mechanisms, such as the European Genome-Phenome Archive (EGA; https://ega-archive.org/), hosted by the European Bioinformatics Institute (EMBL-EBI) and the Centre for Genomic Regulation (CRG).

2.1 Low-throughput genetic data

Low-throughput genetic data, such as karyotypes, short tandem repeats (STRs), HLA types and genetic variants (e.g. determined by Sanger sequencing) can be stored in the cell line record at hPSCreg®, the official data registry of all EBiSC cell lines. Low-throughput data is typically "low volume", and can be manually entered in hPSCreg® through the user interface. Section 3 explains how these data are secured for GDPR. For further information on the registration of EBiSC cell line data in hPSCreg®, please refer to the guidance document "EBiSC Cell Line Data Registration Guide" on the EBiSC website (here).

2.2 High-throughput -omic data

High-throughput -omic data, such as whole genome/exome sequencing or SNP array data, should be deposited at the European Genome-Phenome Archive (EGA), which operates in a fully GDPR-compliant manner, an essential pre-requisite for the data management of sensitive data in European projects.

Detailed guidance on how to submit array-based or sequence data can be found on the EGA website (here). To ensure that the EGA datasets are properly linked with the EBiSC cell line data in hPSCreg®, please carry out the highlighted steps A and B at the appropriate stage in the EGA Submission process (Figure 1). For downstream data re-use, it is recommended that the EGA-deposited data be placed under the control of the EBiSC Data Access Committee (EGAC00001000768), which can take over the task of managing data access to these sensitive genomic datasets (**Step A** in Figure 1).

EGA Data Accession numbers (prefixed by "EGAD") of the deposited -omic datasets (**Step B** in Figure 1) should be entered in the cell line data record at hPSCreg® to maintain the link from the cell line to the -omic datasets (please see Figure 2). To maximize data discoverability, please include the EGA Data Accession number in any publications where the data was generated or used. Europe PMC automatically links publications to the deposited datasets(s) in EGA (please see an example here).

Figure 1. Quick guide to EGA submission. Particular attention should be paid to Steps A and B at the appropriate stages of EGA data submission, to ensure that the EGA-deposited data is linked to the EBiSC cell line. Figure adapted from the EGA Quick Guide (https://ega-archive.org/submission/quickguide).

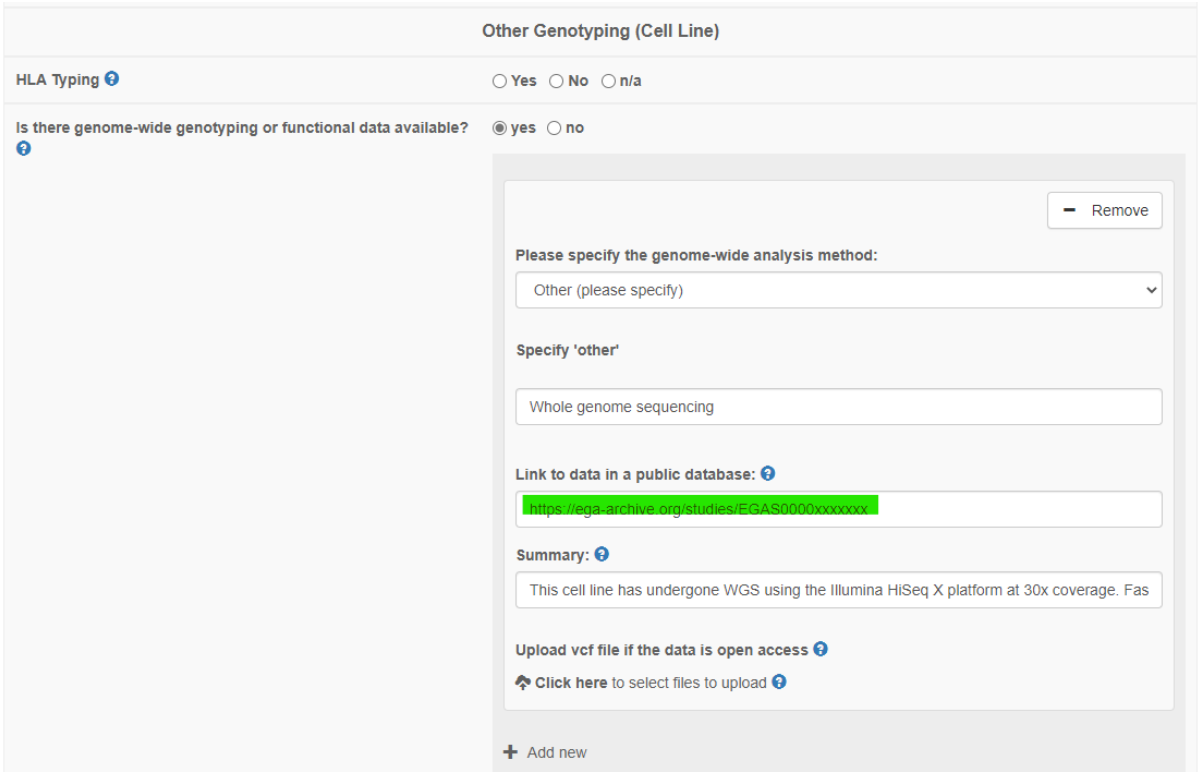

Figure 2. Recording the weblink to the EGA-deposited dataset in hPSCreg®, the official cell line data registry of all EBiSC lines. The weblink should be saved in the "Other Genotyping (Cell Line) section of the "Genotyping" tab in the hPSCreg® User Interface.

3 Access to genetic or genomic data associated to EBiSC cell lines

Genetic / genomic data associated to EBiSC lines is subject to different access levels, depending on risk assessment levels for re-identification of an individual (see Annex). In principle, all data is accessible to scientific researchers, under the accessibility conditions in Table 1.

Table 1. Data Access Provisions

Data Type	Access Provision
Low-throughput genetic variants (e.g. a disease associated variant in a specific gene)	Public
Short Tandem Repeats (STR)	Managed access: subject to approval by the EBiSC Data Access Committee and contractual agreement through the Data Access Agreement
Human Leukocyte Antigen (HLA)	Managed access: subject to approval by the EBiSC Data Access Committee and contractual agreement through the Data Access Agreement
Raw -omic or high-throughput data (e.g. fastq, cram, bam, SNP arrays)	Managed access: subject to approval by the EBiSC Data Access Committee and contractual agreement through the Data Access Agreement
Bulk processed variant data from -omic data (e.g. VCF files)	Managed access: subject to approval by the EBiSC Data Access Committee and contractual agreement through the Data Access Agreement

For access to genomic datasets, please refer to the FAQ on the EBiSC website (here) and contact EBiSC (here).

4 Annex – Risk-based Anonymisation

For the EBiSC bank to remain a sustainable, easily accessible community resource, much of the cell line data is accessible and searchable on the public EBiSC cell line catalogue. EBiSC must strike a balance between providing sufficient cell line data to enable researchers to find and use the cell lines appropriate for their research needs, whilst protecting the cell line donors (Data Subjects) from re-identification attacks.

Within risk-based anonymisation, the re-identification risk is controlled by two factors: data transformation and control measures (El Emam K, Arbuckle L. Anonymizing Health Data: Case Studies and Methods to Get You Started. 1st ed. Sebastopol, CA: O'Reilly Media, Inc.; 2013). Data transformation entails techniques like generalization, suppression, adding noise, and microaggregation, whereas control measures include security, privacy and contractual controls. For the cell line associated genetic / genomic information in Table 2, we summarize the anonymisation measures taken to hinder re-identification, as well as map the potential risks of breach for each piece of information according to the risk matrix in Table 3.

Table 2. Summary of genetic / genomic information associated to EBiSC cell lines and potential risks

No.	**Original Information held by Clinician or Depositor**	**Information stored in hPSCreg®**	**Data Transformation Measure**	**Control Measures**	**Description of Mapped Risk**
1	Genetic Variants	Low-throughput genetic Variants			Re-identification of a Data Subject may occur in combination with other data; higher risk associated with genetic variants with low frequency in a population.
2	Karyotype	Karyotype	Suppression, Generalization*	Contractual control: managed access to high-resolution data via Data Access Committee	In the absence of high-resolution data, re-identification of a Data Subject may occur in combination with other data. Access to high-resolution karyotyping data is subject to approval by DAC and per data access agreement, but end users may break the terms of the contract.
3	Functional datasets, such as genome or transcriptome sequencing	Link to database repository	Suppression	Contractual control: managed access to data via Data Access Committee	Access to data is subject to approval by DAC and per data access agreement, but end users may break the terms of the contract.
4	HLA	HLA	Suppression	Contractual control: managed access to data via Data Access Committee	Access to data is subject to approval by DAC and per data access agreement, but end users may break the terms of the contract.
5	STR	STR	Suppression	Contractual control: managed access to data via Data Access Committee	Access to data is subject to approval by DAC and per data access agreement, but end users may break the terms of the contract.

* In progress

Table 3. Risk matrix for re-identification and assessments for EBiSC cell line genetic / genomic information

	Consequence / Severity					
Likelihood		Negligible	Minor	Moderate	Major	Critical
	Rare	Low	Low (low throughput-genetic variants)	Low	Medium	High
	Unlikely	Low	Low	Medium	Medium	High
	Possible	Low	Medium (karyotype)	Medium	High (-omic datasets, HLA, STR)	High
	Likely	Medium	Medium	High	High	Extreme
	Almost Certain	Medium	Medium	High	Extreme	Extreme

Mitigation of Risks in Tables 2 and 3

The highest ranked risks in Table 3 involve personal data that is especially sensitive, namely genetic data, whose access is managed by a data access control mechanism. Development of a structured monitoring process designed to audit controlled access data users could also provide additional assurances that researchers honour their contractual responsibilities as set out in the data access agreement.

Medium-ranked risks include kinds of data that have potential for re-identification, but only in combination with other data. The resolution of karyotyping depends on the sensitivity of the method. A possible mitigation control would be to make the karyotyping information less specific using domain generalization hierarchies. For example, karyotype information, regardless of the method, could be simplified into “genetic sex confirmed = true/false” and “karyotype = normal/abnormal”, to suppress the karyotype details.

Low-ranked risks involve single or low-throughput genetic variants. This data could be used in a re-identification, but only in combination with many other data points, which might include other data within the cell line record, as well as external data such as publications and social media.

5 Change History

Version	Valid from	Changes compared to previous version
0.1	16-AUG-2022	First draft, with suggestions from EGA and EBISC