OPEN

The European Genome-phenome Archive of human data consented for biomedical research

Ilkka Lappalainen[1], Jeff Almeida-King[1], Vasudev Kumanduri[1], Alexander Senf[1], John Dylan Spalding[1], Saif ur-Rehman[1], Gary Saunders[1], Jag Kandasamy[1], Mario Caccamo[1,5], Rasko Leinonen[1], Brendan Vaughan[1], Thomas Laurent[1], Francis Rowland[1], Pablo Marin-Garcia[1,5], Jonathan Barker[1], Petteri Jokinen[1], Angel Carreño Torres[2], Jordi Rambla de Argila[2], Oscar Martinez Llobet[2], Ignacio Medina[1], Marc Sitges Puy[2], Mario Alberich[2], Sabela de la Torre[2], Arcadi Navarro[2–4], Justin Paschall[1] & Paul Flicek[1]

The European Genome-phenome Archive (EGA) is a permanent archive that promotes the distribution and sharing of genetic and phenotypic data consented for specific approved uses but not fully open, public distribution. The EGA follows strict protocols for information management, data storage, security and dissemination. Authorized access to the data is managed in partnership with the data-providing organizations. The EGA includes major reference data collections for human genetics research.

The technical ability to identify regions of the human genome that harbor variants influencing disease risk is one of the most important recent advances in genomics. Many studies use large disease cohorts, including the Wellcome Trust Case Control Consortium[1] and the UK10K project. At the same time, the International Cancer Genome Consortium (ICGC) is generating the complete genomes of matching tumor and normal samples for a number of cancers in an effort to understand the genomics of the disease. Published genetic variants are collated in fully public resources such as the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies[2] or Ensembl[3]. In addition to public variants, individual-level genetic and phenotypic data or summary statistics from the research projects are often required for replication[4], meta-analysis[5] and many other secondary uses, such as methods development[6] or use as control samples[7]. However, these data must be processed, archived and transferred in a manner that respects the consent agreements signed by the study subjects[8]. This often means that data can only be provided to bona fide researchers and used for specific research aims[9].

The existing public data archives that provide unrestricted access to data are incompatible with these requirements, and the EGA was thus launched in 2008 by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to support the voluntary archiving and dissemination of data requiring secure storage and distribution only to authorized users. Recently, the EGA has expanded from an exclusively EMBL-EBI project to a collaboration with the Centre for Genome Regulation (CRG) in Barcelona, Spain, in what may be a first step toward a larger distributed network of data archiving and dissemination services. Both EMBL-EBI and the CRG are publicly funded organizations, and the former is an intergovernmental organization formed by a collection of mostly European member and associate member states.

Since the launch of the EGA, researchers from around the world have deposited and accessed data from over 700 of its studies of various types (**Fig. 1** and **Table 1**). These studies vary from large-scale array-based genotyping experiments on thousands of samples in case-control[1,10] or population-based[11,12] studies to sequencing-based studies designed to understand changes in the genome, transcriptome or epigenome in both normal tissue[13] and various diseases such as cancer[14–16]. As a result, the EGA has grown from about 50 TB to 1,700 TB during the last 4 years.

Since 2011, the bulk of EGA submissions have transitioned from array-based genotyping to next-generation sequencing studies. Summary-level genetic information is also accepted if such data cannot be publicly released. Phenotype data are currently most often provided at the level of the data set rather than the individual; for example, a group of samples may be reported to have the same disease phenotype. However, submission of individual-level, detailed phenotypes is increasing in frequency and is encouraged.

[1]European Molecular Biology Laboratory–European Bioinformatics Institute, Hinxton, UK. [2]Centre for Genomic Regulation, Barcelona, Spain. [3]Institute of Evolutionary Biology, Universitat Pompeu Fabra–Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Spain. [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [5]Present addresses: Genome Analysis Centre, Norwich, UK (M.C.), and Fundación Investigación Clínico de Valencia (INCLIVA), Valencia, Spain (P.M.-G.). Correspondence should be addressed to P.F. (flicek@ebi.ac.uk), J.P. (paschall@ebi.ac.uk) or A.N. (arcadi.navarro@upf.edu).

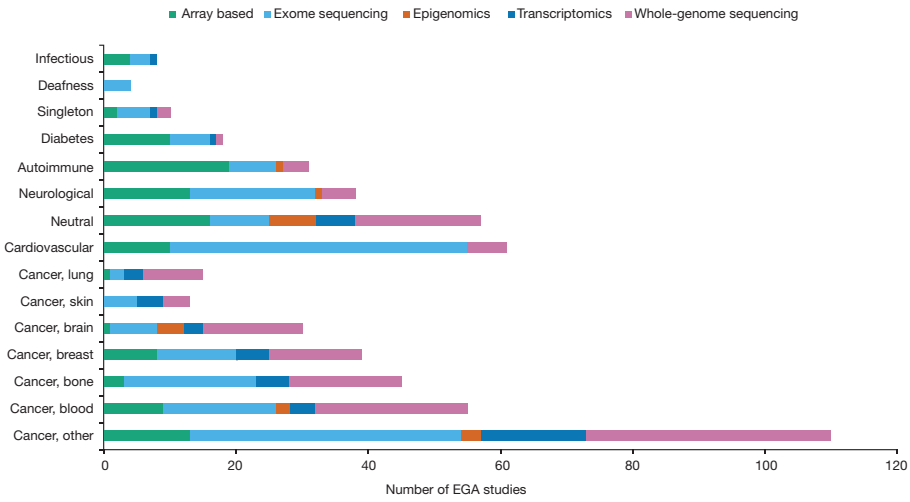

Figure 1 Breakdown of EGA studies by disease topic as of 2014. The majority (57%) of EGA studies investigate cancer of various types. EGA experimental data describe the methodology employed for each study, which is represented by the different colors. Exome sequencing (38%) is the most common methodology, followed by whole-genome sequencing (29%), array-based technologies (20%), transcriptomics (9%) and epigenomics (3%). As one might expect, array-based experiments are typically from older studies, whereas both transcriptomic and epigenomic investigations are more recent. A complete list of the meta-data collated to create this graph is provided on the EGA website.

In this report, we describe the roles and policies of the EGA, provide information on how access decisions are made, outline the methods for data submission and dissemination, and describe the EGA system infrastructure. The EGA has similarities and differences with the database of Genotypes and Phenotypes (dbGaP) provided by NCBI[17]; where appropriate, we describe the features and procedures that distinguish these two databases.

Roles and policies of the EGA

The role of the EGA is to promote the distribution and sharing of biomolecular and phenotypic data collected from human subjects who have consented to them being shared for research uses—but not for full, open public release—by providing a system for the secure archiving and dissemination of such data. Submitters are required to certify that the data they have deposited in the EGA have been produced and made available in a manner that is consistent with the original consent agreements, national laws and applicable regulations. Data sets submitted to the EGA are further required to be made accessible in a timely fashion to all bona fide researchers whose use of the data is consistent with the original consent agreements. As described below, the EGA brokers data access on behalf of the submitting organization and provides data management and distribution services for users of the database. Any security breach or data misuse by users is immediately reported to the relevant Data Access Committee (DAC) upon becoming known to the EGA, following a standard operating procedure.

The EGA supports prepublication data release in accordance with the Toronto agreement for community resource projects[18] and for other research organizations and funding agencies that require or encourage data release. For example, the data from the UK10K project are made available to authorized users as regular data updates during the project. In addition, the ICGC uses EGA to provide access to the raw sequence data and other appropriate data generated by many of the international partner projects[19]. The EGA also archives and distributes a wide range of data sets in support of scientific publications, providing published data sets as a permanent part of the scientific record.

Biomolecular databases and archives are distributed worldwide with significant concentrations at dedicated institutes, including NCBI and EMBL-EBI. Within this landscape, the EGA serves as a secure, authorized-access mechanism for data types that, if consented for fully open release, could otherwise have been deposited in the EMBL-EBI resources tailored to store DNA, RNA, protein or sample data[20–22]. In some cases, the EGA stores sample-level raw data files and detailed phenotype information, whereas aggregated results, such as disease-associated variants, or other non-sensitive data are stored in the public archives with data set linking to enable discovery.

The EGA security policy includes the development of a safe computing facility and a comprehensive suite of protocols for information management.

Table 1 EGA users are distributed throughout the world as of May 2015

Geographical location	Number of submitters	Archived data (TB)	Number of authorized data users
North and South America	53	341	1,995
Europe	137	1,204	2,608
Asia	38	130	556
Oceania	7	110	246
Africa	0	0	5
Total	235	1,785	6,601[a]

[a]The total number of authorized users includes 1,191 from commercial organizations, which have not been separated geographically.

Access to data

The EGA has a distributed access-granting policy in which data access decisions are made by the nominated DAC for a given submission and not by the EGA. The DAC may consist of a dedicated committee formed by the funding or governmental organization that approved and monitored the initial study, an institutional committee or an individual primary investigator. Regardless of the scope or composition of the DAC, the EGA only provides services for studies when access decisions are made exclusively on the basis of scientific and ethical criteria in compliance with the original informed consent agreements. The EGA will withdraw service if data access is being denied selectively because of scientific competitiveness or other reasons not based on the original informed consent.

In a typical case, users wishing to access a specific data set apply directly to the corresponding DAC (**Fig. 2**), following contact instructions on the EGA website. Assuming approval, a Data Access Agreement (DAA) is made directly between the prospective user and the DAC, and it dictates data management policies, security arrangements and other potential limitations on data use. For example, some data may not be used for commercial purposes and users may be subject to a temporary publication embargo for projects participating in prepublication data release. In accordance with accepted practice, the EGA provides data access at the level of granularity that is appropriate for the submitted study. As an example, in a case-control study, the user may separately request to access individual-level data only for the control data set.

Once approved for access, a user will be issued an EGA account, which is subject to a number of conditions, including that the account

information not be shared. EGA accounts can be updated with additional access rights upon each successful application. To ensure that the DAA remains valid, the EGA requires DAC authorization for changes to user details, such as institutional affiliation. The EGA offers support and online tools for the DACs to manage the access rights for their data sets directly within the system.

The policy of distributed access-granting is the most important distinguishing feature of the EGA in comparison to dbGaP. Authorization decisions for dbGaP's data sets are made by the US National Institutes of Health (NIH) institute that sponsored the study in question. In the United States, the NIH serves as both a funding and policy-making agency and, through the NCBI, a mechanism for data distribution. This allows the NIH to specify dbGaP as a required (although non-exclusive) data distribution channel for specific studies. In contrast, the rest of the world has a diversity of funding agencies and national regulations, and these are very often compatible with the distributed data access policy of the EGA for data archiving and distribution. Indeed, this distributed model is an especially good fit for the European research structures that provide support to the EGA.

The EGA and dbGaP share meta-data to improve the discoverability of data deposited in either repository. This sharing of publicly available information such as study name and publication information enables researchers to search for data sets and find the relevant starting point for the access approval process, regardless of whether the data are in the EGA or dbGaP. Data are only disseminated from the archive that accepted the original submission, as actual data files are not exchanged between the EGA and dbGaP.

The EGA websites

Users can access the full EGA service from its instance at either EMBL-EBI or the CRG. Both current EGA websites are arranged around the study concept. A study is typically an experimental investigation of a particular phenomenon, for example, a genome-wide association study or a matched tumor-normal cancer genome project. The EGA study page describes how the study was conducted and all the associated data sets. The page also includes links to other relevant data resources at the EMBL-EBI or NCBI, the primary publication when available and the data provider. Studies, data sets, DACs and data providers are assigned stable identifiers that should be referred to in the publication and are used to link together information within the EGA. These identifiers provide direct access to the relevant EGA webpage through the central EMBL-EBI search engine and serve as stable URLs.

The primary point of entry for accessing the controlled-access data stored in the EGA is provided through the data set page. Each data set includes publicly available information about the technology used to assay the samples and guidelines describing how to apply for data access. Once access has been granted and users have logged into the secure EGA website, the page will show all the associated manufacturer raw data files, processed information such as variants or genotypes, or any associated study summary data. Logging into an EGA account facilitates data requests from the archive and allows users to track their current requests within the system.

All data are encrypted for dissemination and made available to each authorized user through FTP as well as fast Internet transfer protocols such as Aspera and UDT. Data transmission methods for submission and dissemination have evolved as data volumes have increased and now include a custom Java client making use of the UDT protocol and performing automatic MD5-checksum validation and encryption. This automation has increased user-friendliness and reduced error.

Data submission

Complete up-to-date information about submitting data to the EGA is available from its websites. Briefly, submitters first request a private submission account from the EGA to

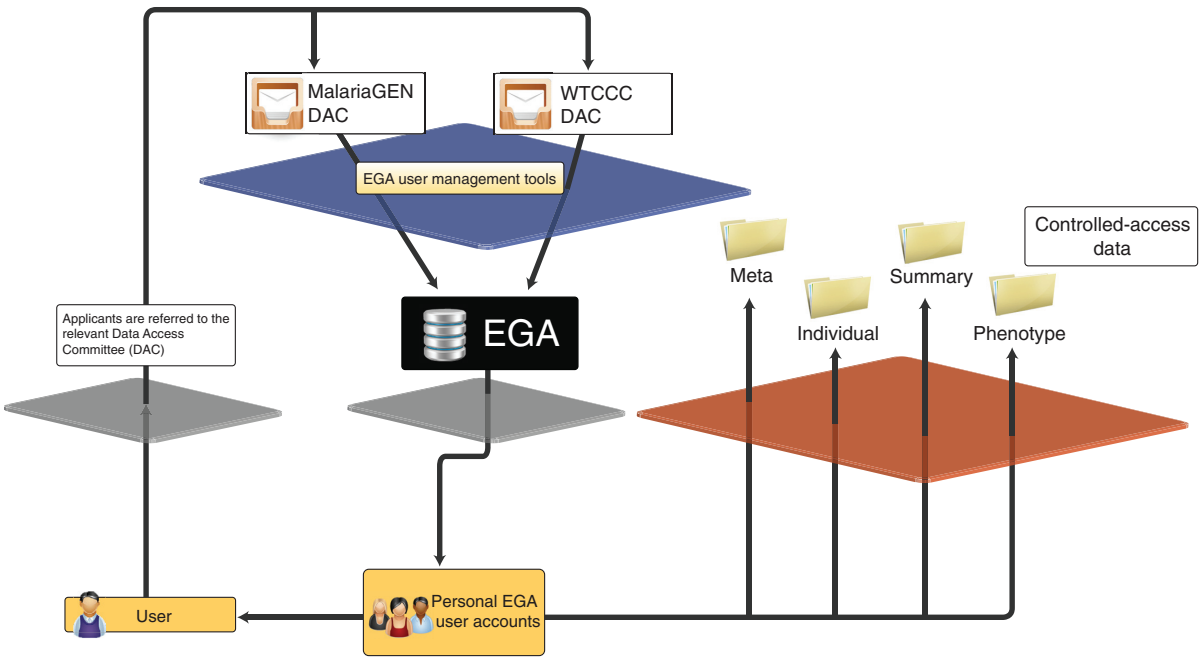

Figure 2 The EGA distributed data access model. The EGA refers applicants to appropriate DACs on the website. Each DAC grants access to its data independently. The EGA will create a personal account for each approved applicant listed on the application. The account holds all approved data access permissions and allows account holders to request services from the EGA team, such as downloading of encrypted files from the archive or support for any technical or data content–related questions. The data downloaded from the EGA website are provided under a DAA, which is a legal agreement between each approved user and the data-governing DAC.

Table 2 Further information available from the EGA website

Guidelines	Description	Example web address[a]
Introduction to EGA	Documents related to EGA processes, stable identifiers and Frequently Asked Questions	https://ega.crg.eu/about/introduction
Tutorial videos	Video library for EGA account holders, data submitters or DAC members	http://www.ebi.ac.uk/ega/about/videos
Submissions to EGA	Submission manual for all experiment types and Frequently Asked Questions	https://ega.crg.eu/submission
DACs	Documentation explaining how to establish and manage a DAC effectively	http://www.ebi.ac.uk/ega/submission/data_access_committee
EGA tools	Collection of EGA tools for data download or the submission process	https://ega.crg.eu/submission/tools
EGA security	EGA security policies	https://www.ebi.ac.uk/ega/about/security

[a]All EGA policies are available from both the CRG website and the EMBL-EBI website by using either http://ega.crg.eu or http://www.ebi.ac.uk/ega, respectively, the listed URL.

access the range of tools available for file and meta-data upload. It is recommended that all primary data files be uploaded using the secure EGA application that automatically provides data encryption and transfer integrity checks. Meta-level information about a study should be submitted using either the Webin online tool[22] for experiments using next-generation sequencing technology or an EGA-provided spreadsheet-based meta-data submission template for other assay types. It is also possible for submitters to connect local information management systems directly to the EGA for automatic submission support. The EGA submission guidelines provide detailed information about each stage (**Table 2**). To ensure that all possible submitters can be served by the EGA, encrypted data will also be accepted on hard drives if data size or submitter bandwidth necessitates this.

Once the submission has been completed, the EGA confirms the integrity of each submitted file, transfers the data into a secure computing area, and decrypts and uploads it into archival databases. The EGA staff work directly with the submitter to make sure that the data are correctly uploaded into the system, pass quality checks and are accurately represented on the website. While data are being collected and analyzed, all uploaded files and the website may be made visible to research collaborators, referees for manuscripts under review provided they are willing to state their identity and make an access application to the appropriate DAC, or any other approved users. The EGA supports a 'hold until publication' (HUP) status for 6–12 months to enable a study to be submitted and verified but kept private until it is released simultaneously with a journal publication. There is no defined maximum time that a data set can remain in HUP status, but extensions beyond 1 year require justification. Although all published data are made available as soon as possible, actual initiation of data dissemination from the EGA requires authorization from the submitting organization.

Future directions

The recent expansion of the EGA to an EMBL-EBI and CRG collaboration will help support major new EGA data sets, including genomic data from the Genome of the Netherlands[23] and Deciphering Developmental Disorders[24] projects, epigenetic and functional data from the Blueprint[25] and HipSci consortia, and data relevant to the genetic basis of rare disease from the UK BRIDGE Project.

The EGA is also working on several new added-value services that will increase the usability of the submitted data. For example, submitted sample phenotypes will be described using ontology-based terms to facilitate better search functionality and assist users looking to merge data across studies. Links are being established with literature databases such as Europe PubMed Central to more closely track secondary publications based on data from the EGA. The EGA will also provide a variant calling and imputation service for limited sets of data submitted to the database. Finally, with the support of the Barcelona Supercomputing Center (BSC) and user-facing EMBL-EBI computational resources, the EGA is exploring cloud-based data analysis options.

URLs. EGA website, http://www.ebi.ac.uk/ega/ or http://ega.crg.eu/; UK10K Project, http://www.uk10k.org/; US NIH Data Sharing Policies, http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html; HipSci Project, http://www.hipsci.org/; UK BRIDGE Project, https://bridgestudy.medschl.cam.ac.uk/; Europe PubMed Central, http://europepmc.org/.

ACKNOWLEDGMENTS

We thank A. Ducanson, R. Banerjee, N. Walker, E. Birney and S. Potter for helpful discussions and comments. The EGA has received support from the European Molecular Biology Laboratory, the European Union ELIXIR Technical Feasibility Study, the Wellcome Trust (grant WT 085475/C/08/Z), the UK Medical Research Council (grant G0800681), the Spanish Instituto de Salud Carlos III Instituto Nacional de Bioinformática (grant PT13/0001/0026), the Spanish Ministerio de Economía y Competitividad (MINECO) and Centro de Excelencia Severo Ochoa (grant SEV-2012-0208), the Fundació La Caixa and the Barcelona Supercomputing Centre. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013 under grant agreements 211601–ELIXIR, 200754–GEN2PHEN, 262055–ESGI, 242006–BASIS, 261376–IHMS and 305444–RD-CONNECT).

AUTHOR CONTRIBUTIONS

I.L., A.N., J.R.d.A., J.P. and P.F. provided project leadership and management. V.K., A.S., J.D.S., S.u.-R., M.C., R.L., P.M.-G., I.M., O.M.L. and S.d.l.T. developed software. J.A.-K., M.S.P. and G.S. provided user support. J.K., B.V., T.L., F.R. and M.A. developed the EGA website. J.B., A.C.T. and P.J. provided systems support. P.F. and I.L. wrote the manuscript with contributions from all other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
2. Welter, D. *et al. Nucleic Acids Res.* **42**, D1001–D1006 (2014).
3. Flicek, P. *et al. Nucleic Acids Res.* **42**, D749–D755 (2014).
4. Ban, M. *et al. Eur. J. Hum. Genet.* **17**, 1309–1313 (2009).
5. Berndt, S.I. *et al. Nat. Genet.* **45**, 501–512 (2013).
6. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. *Nat. Genet.* **44**, 955–959 (2012).
7. Lu, Y. *et al. Hum. Mol. Genet.* **23**, 6112–6118 (2014).
8. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. *Genome Med.* **5**, 100 (2013).
9. Kaye, J. *Annu. Rev. Genomics Hum. Genet.* **13**, 415–431 (2012).
10. Trynka, G. *et al. Nat. Genet.* **43**, 1193–1201 (2011).
11. McEvoy, B.P. *et al. Genome Res.* **19**, 804–814 (2009).
12. Surakka, I. *et al. Genome Res.* **20**, 1344–1351 (2010).
13. Zilbauer, M. *et al. Blood* **122**, e52–e60 (2013).
14. Wiegand, K.C. *et al. N. Engl. J. Med.* **363**, 1532–1543 (2010).
15. Kulis, M. *et al. Nat. Genet.* **44**, 1236–1242 (2012).
16. Sato, Y. *et al. Nat. Genet.* **45**, 860–867 (2013).
17. Mailman, M.D. *et al. Nat. Genet.* **39**, 1181–1186 (2007).
18. Toronto International Data Release Workshop Authors. *Nature* **461**, 168–70 (2009).
19. International Cancer Genome Consortium. *Nature* **464**, 993–998 (2010).
20. Gostev, M. *et al. Nucleic Acids Res.* **40**, D64–D70 (2012).
21. Vizcaíno, J.A. *et al. Nucleic Acids Res.* **41**, D1063–D1069 (2013).
22. Pakseresht, N. *et al. Nucleic Acids Res.* **42**, D38–D43 (2014).
23. Genome of the Netherlands Consortium. *Nat. Genet.* **46**, 818–825 (2014).
24. Firth, H.V., Wright, C.F. & DDD Study. *Dev. Med. Child Neurol.* **53**, 702–703 (2011).
25. Adams, D. *et al. Nat. Biotechnol.* **30**, 224–226 (2012).