E.2 Data Management

E.2.1 Data Management Overview

Data coordination and management have been key strengths of the ICGC to date. ICGC-ARGO is considerably larger in scale than ICGC, and involves a far richer and more complex set of clinical and environmental information, which requires structural changes to the existing Data model to ensure the sound management of ICGC-ARGO data through several operational entities described in this policy. The overarching principles of the data management system and its design are:

- Provide secure and reliable mechanisms for the sequencing centers, clinical data managers, and other ICGC participants to upload their data;
- Track data sets as they are uploaded and processed, to perform basic integrity checks on those sets;
- Allow regular audit of the project in order to provide high-level snapshots of the consortium's status;
- Perform more sophisticated quality control checks of the data itself, such as checks that the expected sequencing coverage was achieved, or that when a somatic mutation is reported in a tumor, the sequence at the reported position differs in the matched normal tissue;
- Enable the distribution of the data to the long-lived public repositories of genome-scale data, including sequence trace repositories and microarray repositories;
- Provide essential meta-data to each public repository that will allow the data to be findable and usable;
- Facilitate the integration of the data with other public resources, by using widely-accepted ontologies, file formats and data models;
    - allow researchers to compute across data from ICGC-ARGO donors that are stored in multiple localities and to return analytic results that span the entire distributed data set.
    - support for hypothesis-driven research: The system should support small-scale queries that involve a single gene at a time, a short list of genes, a single specimen, or a short list of specimens. The system must provide researchers with an interactive system for identifying specimens of interest, finding what data sets are available for those specimens, selecting data slices across those specimens

Each data producer will manage its own data submission and be responsible for primary QC, data integrity and protection of confidential information.

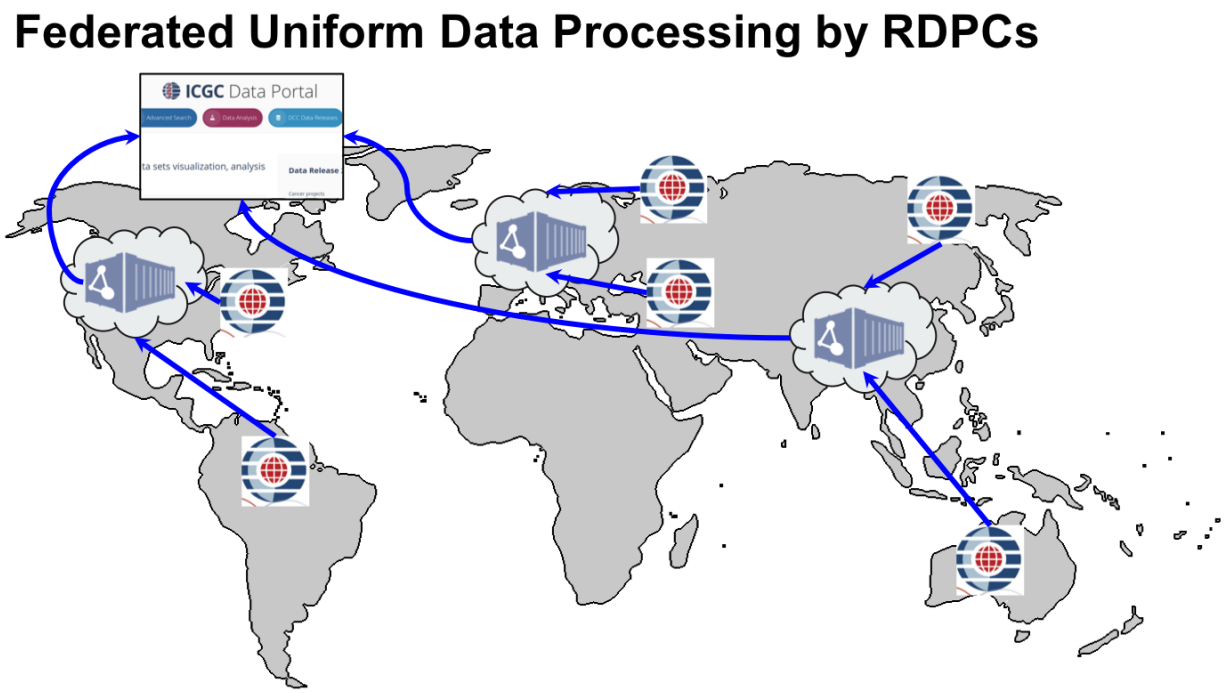
E.2.2 ICGC Data Management Infrastructure

The ICGC-ARGO Data Coordination Centre (DCC) in collaboration with the working groups and consortium members will define the clinical data dictionary and data model. The DCC will provide a submission system for accepting and validating clinical data. The DCC will also coordinate with Regional Data Processing Centres (RDPCs) to accept, validate and uniformly

analyze molecular data submitted by the ICGC-ARGO sequencing centres. The RDPCs will process the sequencing data through a standardized series of data analysis pipelines to identify genomic mutations. The use of RDPCs and cloud compute providers together give ICGC-ARGO considerable flexibility in where the data is physically stored. This will allow the project to successfully navigate the changing landscape of international policies on human genetic data storage and distribution. Carefully written software will allow researchers to compute across data from ICGC-ARGO donors that are stored in multiple localities and to return analytic results that span the entire distributed data set.

The interpreted results, along with quality control metrics, will be sent to the DCC for integration with other ICGC-ARGO data sets and dissemination to the scientific and lay communities. Processed sequencing data will be archived in one or more public sequence archives, and mirrored to several cloud compute providers, where qualified researchers will be granted the right to perform additional analyses on the data in a secure and ethically responsible fashion.

The ICGC-ARGO software engineering group will support the operations of the ICGC DCC and the ICGC-ARGO Regional Data Processing Centres, and will be responsible for developing and distributing the software systems and protocols required for the operation of these centres.



E.2.3 Data Release

POLICY: The members of the International Cancer Genomics Consortium (ICGC) are committed to the principle of rapid data release to the scientific community.

Data producers are recognized to have a responsibility to release data rapidly and to publish initial global analyses in a timely manner. Of equal importance is the responsible use of the data by end-users, which is defined as allowing the data producers the opportunity to publish the initial global analyses of the data within a reasonable period of time, as per the Publication Policy.
The members of the ICGC agree to identify the projects they support and carry out for the comprehensive genomic characterization of human cancers as a set of community resource projects. Data producers, by explicit agreement as members of the ICGC, acknowledge their responsibilities to release data rapidly and to publish initial global analyses in a timely manner. Similarly, funding agencies acknowledge their role in encouraging and facilitating rapid data release from cancer genome projects.
Timing of Data Releases
ICGC ARGO member programs will have privileged access to data from other members of the Consortium based on their level of Membership. Data access is tiered and aimed not to disadvantage Members or Associate Member Data producers, with a framework that encourages data sharing, yet provides data generators with sufficient time to perform analyses:

- Up to 12 months from completion of standardised analyses: Access to Programme submitting data only
- 12 months: Access to Members
- 18 months: Access to Associate Members
- 24 months: Accessible by external parties

Standardized analyses are considered complete when both mandatory clinical data has been submitted by the Program and molecular data has been uniformly analysed by the RDPCs.

E.2.4 Data Access
The nature of the data that will be produced by ICGC-ARGO members; substantial clinical annotation and extensive genomic data, raises important human subject privacy protection issues. The patient/individual protection policies developed for ICGC-ARGO are designed to balance two important goals: to facilitate investigations of genomic changes related to cancer and, at the same time, to respect and protect the patients/individuals whose data and materials have been or will contribute to ICGC-ARGO member programmes. It is technically possible that genomic information generated by ICGC-ARGO could lead to re-identification of an individual if linked or combined with other information or archived data There is also a risk of individual identification by computer-based analysis of the clinical data in conjunction with, for example, third-party demographic and healthcare management databases. This potential identification

could then publicly link the individual to his/her clinical information collected by the participating projects and could lead to social risks such as discrimination or loss of privacy.

ICGC ARGO member programs will have privileged access to data from other members of the Consortium based on their level of Membership. After a 24-month period following standardized analysis ICGC ARGO data will be made available to external parties following established data access processes described below. Data users will be required to consult the ICGC ARGO Publication Policy to be aware of the publication status of data sets and guidelines in place on behalf of data producers.

ICGC-ARGO have carefully considered, based on existing knowledge and best practice, which data types should be publicly accessible and which should be governed by a controlled process. Open access datasets contain the **Core Clinical Data** points (which are mandatory for each member programme to submit) and will be publicly accessible and contain only data that cannot, at present, alone or combined with other data be aggregated to generate a dataset unique to an individual without reasonable efforts.[1] The amount and nature of genetic data that might be associated with an individual from the Open Access Datasets has been carefully considered and will continue to be monitored by ICGC-ARGO. The second category, Controlled Access Datasets, will contain composite genomic and clinical data that are associated to a unique, but not directly identified, person. Controlled access datasets contains **Extended Clinical Data fields** (which are optional for programmes to submit) as summarised below.

ICGC Open Access Datasets	**Controlled Access Datasets**

Cancer pathology · Histologic type or subtype · Histologic nuclear grade · Tumour staging **Patient/person** · Gender · Age (single category for ages over 89) · Vital status · Age at last follow-up (single · category for ages over 89) · Survival time · Cause of death · Relapse type · Relapse interval · Disease status at last follow-up · Interval from primary diagnosis to last follow-up · Treatment type · Treatment duration · Therapeutic intent · Response to therapy · Cumulative drug dosage **Specimen/sample** · Specimen tissue source · Specimen anatomic location Gene expression (normalized) • DNA methylation • Genotype frequencies • Computed copy numbers and loss of heterozygosity · Newly discovered somatic variants	**Detailed Phenotype, treatment and outcome data** • Region of residence • Ethnicity • Risk factors • Post therapy staging • Performance status • Detailed treatment cycle and dose details • Treatment toxicity **Specimen/sample** • Specimen processing Gene Expression (probe-level data) • Raw genotype calls • Gene-sample identifier links • • Genome sequence files

This list will be periodically revised to reflect the continually evolving fields of genomics, bioinformatics, and to comply with ethics and privacy policies and regulations.

ICGC established two bodies to oversee controlled access: The Data Access Compliance Office (DACO) and an International Data Access Committee (IDAC). DACO is responsible for processing access requests from the scientific community and its activities are overseen by IDAC. DACO is required to verify the conformity of users' projects with the goals and policies of

ICGC, including, but not limited to, policies concerning the purpose and relevance of the research, the protection of participants, and the security of participants' data.

DACO, IDAC, and ICGC's Ethics and Governance Committee (AEGC) collaboratively developed the data access application forms (which include an access agreement), as well as the policies to be used by ICGC. The rules and policies of ICGC have influenced the controlled access strategies of several database projects, including the Wellcome Trust Sanger Institute and the Human Epigenome Consortium".

Authorizations to access controlled data will be broad, so that authenticated users will get permission to obtain access to controlled data generated from all samples studied by any participating ICGC ARGO project (as the feasibility of providing permissions to datasets originating from single or partial subsets of participating centres has been determined to be unworkable in the context of the ICGC).

The DACO will also develop guidelines to streamline approaches to providing qualified investigators with access to controlled data. In doing so, it will consider mechanisms and tools that have been already in use by other organizations that distribute controlled datasets to international scientists (for example, GA4GH or the Wellcome Trust Case Control Consortium). Under current processes potential users and their institutions will be required to submit an Access Application Form and sign a Data Access Agreement. Interested users and institutional officials who are authorized to make legally binding agreements for the institution will be required to adhere to the conditions laid out in the Access Agreement. Investigators will need to agree to regular review and renewal requested by the DACO for such authorization and in cases when they move to new institutions.

E.2.5 Data Sharing

ICGC-ARGO is aligned with the GA4GH in being dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. Efficacy and responsibility will be supported through ensuring that the data management processes described use the most up-to-date technologies and software to ensure rapid, secure and high-quality data submission and transfer, overseen by robust ethics and data access frameworks. As a driver project for the GA4GH, ICGC-ARGO will have access to new technologies and expert communities to assist its work, as well as through which it can disseminate its gained knowledge.

Much of the data contributed to ICGC-ARGO will be retrospective in nature. As well, because membership will span many different countries with differing regulatory requirements and cultural norms, there will be limitations in how and with whom some data can be shared. ICGC-ARGO follows the GA4GH belief that members should be encouraged to share as widely as is possible and will work with groups to maximize data sharing to greatest extent possible within accepted legal and ethical boundaries. ICGC-ARGO experts will keep abreast of any

changing laws and regulations that might impact the cross-border sharing of data and will act to ensure ICGC-ARGO respects any changing circumstances.

POLICY: ICGC-ARGO members will be encouraged to work towards ensuring that data sets can be shared to greatest extent possible while recognizing differing legal and ethical requirements

---

[1] Council of Europe, Recommendation Rec (2016) 6 of the Committee of Ministers to member states on research on biological materials of human origin