ATPBR

PLATFORM DOCUMENTATION

Brazilian Population Tumor Atlas

An initiative dedicated to understanding the biological and genomic heterogeneity of cancer in the Brazilian population.



atpbr.com.br

Brazilian National Cancer Institute (INCA) · Version 0.1 · November 2026

About this booklet

This booklet brings together the documentation of the Brazilian Population Tumor Atlas (ATPBR) platform: what the project is, how the data are governed, how to submit and access data, how the information is organized and how the results are processed and explored.

The chapters follow the life cycle of data on the platform and are grouped into four parts. Each chapter can be read on its own: cross-references give the number of the chapter where each subject is covered in detail.

The content matches the ATPBR online documentation on the date of this edition. The governance and access rules follow INCA's institutional policies; in case of any discrepancy, the institutional documents in force prevail. This English edition is a translation of the Portuguese one; where they differ, the Portuguese edition prevails.

TITLE	Brazilian Population Tumor Atlas (ATPBR): platform documentation
VERSION	Version 0.1 · November 2026
INSTITUTION	Brazilian National Cancer Institute (INCA), Ministry of Health, Brazil
WEBSITE	www.atpbr.com.br
CONTACT	atpbr@inca.gov.br
HOW TO CITE ATPBR	Boroni M, Costa Conteville L, Ribeiro Pinto LF, P B Viola J. *The Brazilian Population Tumor Atlas Project: addressing representation in cancer genomics in Brazil*. Lancet Reg Health Am. 2026;56:101397. doi:10.1016/j.lana.2026.101397



A WORD FROM THE COORDINATOR

Brazilian data for Brazilian oncology

Precision medicine has already changed the way we diagnose and treat many types of cancer. It depends, however, on data, and the data that support it today bear little resemblance to the population served by Brazil's Unified Health System (SUS).

Most knowledge in cancer genomics has been produced with patients from the Northern Hemisphere. In The Cancer Genome Atlas (TCGA), one of the largest repositories in the world, 77% of the samples came from white patients and only 3% from Hispanic or Latino patients. The Brazilian population, one of the most diverse on the planet, with Indigenous, African and European ancestry, remains almost absent from these databases.

This absence has concrete consequences. A study of 2,723 Brazilian genomes found more than 8.7 million genetic variants never described before. Genetic testing criteria used in the country would fail to identify between 17% and 25% of people carrying variants associated with hereditary cancer. Variants that define risk, prognosis and response to treatment do not automatically transfer from one population to another, and clinical decisions based on poorly representative data tend to be less effective for Brazilians.

The Brazilian Population Tumor Atlas (ATPBR) was created to close this gap. Led by INCA and the Brazilian Cancer Genome Consortium, and funded by the Ministry of Health in partnership with CNPq, the project brings together clinical and molecular data from patients in different regions of the country in a national, secure and governed infrastructure that allows them to be integrated, compared and reused.

Increasing the presence of Brazilians in cancer data is not only a matter of scientific equity. It is a condition for discovering biomarkers relevant to our population, developing diagnostic and treatment strategies that work for it and bringing precision medicine to those cared for by SUS.

I invite researchers, health professionals and partner institutions to get to know the platform described in this booklet and to contribute to it. Every case submitted with quality, consent and complete metadata brings Brazilian oncology closer to a faithful portrait of its patients.

João P.B. Viola
Research and Innovation Coordinator
Brazilian National Cancer Institute (INCA)

Contents

PART 1 **Getting to know ATPBR** 5

01 About ATPBR 6
02 Start here 10
03 Platform architecture 12
04 Data governance 14

PART 2 **Submitting and sharing data** 25

05 Data submission 26
06 Data access 30

PART 3 **Data organization** 34

07 Data model 35
08 Data dictionary 37
09 Identification and barcode system 42
10 Data types 53

PART 4 **Data analysis and use** 55

11 Analysis platform 56
12 Results browser 60
13 Usage policies 64

Quick reference

14 Frequently asked questions 66

PART 1

Getting to know ATPBR

The project, how this documentation is organized, the platform architecture and the governance rules.

01 About ATPBR

02 Start here

03 Platform architecture

04 Data governance

CHAPTER 01

About ATPBR

What is ATPBR?

The **Brazilian Population Tumor Atlas (ATPBR)** is a national initiative dedicated to mapping the molecular characteristics of cancer in the Brazilian population and to sharing these data responsibly. Funded by the **Ministry of Health** in partnership with **CNPq** (National Council for Scientific and Technological Development) since December 2023, and led by the Brazilian National Cancer Institute (INCA) and the Brazilian Cancer Genome Consortium, the project gathers, organizes and makes available clinical and **multi-omics** data (genomics, transcriptomics and epigenomics) from cancer patients in different regions of the country.

In practice, ATPBR works as a **centralized and secure infrastructure** that integrates clinical metadata and molecular results, so that tumor omics data produced in different studies, institutions and technologies can be integrated and compared consistently.

The problem ATPBR helps to solve

Most of the genomic data that guide **precision medicine** today were generated in Northern Hemisphere populations.

- In one of the world's largest repositories, The Cancer Genome Atlas (TCGA), **77% of the samples analyzed came from white patients** and only **3% from Hispanic/Latino patients** (Spratt et al., 2016).
- **PCAWG/ICGC-TCGA:** 76.2% of individuals had European ancestry and only 1.0% were classified as Admixed American/Amerindian/Latino (Campbell et al., 2020).
- **AACR Project GENIE:** 72% of patients were classified as white; Hispanic patients had only 57% of the expected representation relative to the US cancer population (Pugh et al., 2022; Cheung et al., 2023).
- **NCI Patient-Derived Models Repository:** 75% of the samples are of European ancestry and less than 1% of Native/Latin American ancestry (Molina-Aguilar and Robles-Espinoza, 2023).

These data reveal a persistent gap in the representation of Latin American populations in the main genomic resources used for biomarker discovery, the development of predictive models and the implementation of precision medicine.

This is a problem because the genetic characteristics of one population do not automatically apply to another: variants associated with **risk**, **prognosis** and **treatment response** may differ. Clinical decisions based on poorly representative data tend to be less effective for Brazilians.

Why does Brazilian diversity matter?

The Brazilian population has **great genetic diversity**, the result of a demographic history marked by **Indigenous, African, European and other ancestral contributions**. This diversity represents a scientific opportunity that is still little explored:

- A study of **2,723 Brazilian genomes** identified more than **8.7 million previously undocumented genetic variants** (Nunes et al., 2025).
- The eligibility criteria for genetic testing evaluated in a Brazilian study would fail to identify between **17% and 25%** of carriers of **pathogenic or likely pathogenic variants** associated with hereditary cancer predisposition syndromes (de Oliveira et al., 2022).
- Some pathogenic germline variants associated with hereditary cancer are **more frequent in the Brazilian population** than in international reference databases (Boroni et al., 2026).
- The **Brazilian Genome Map Oncology Subproject (BGMO)** demonstrated the feasibility of generating tumor-germline WGS in a multicenter setting in Brazil, with 275 patients recruited in the country's five regions and a population composition distinct from that of the large international databases (Ferreira et al., 2026).

These data show that the challenge is not only to **generate more Brazilian genomic data**, but to ensure that they can be **standardized, integrated, governed, shared and reused**. **ATPBR** was designed to fill this gap, creating a national infrastructure capable of turning data that are currently scattered into an integrated resource for research and precision medicine.

Our mission

To increase the representation of the Brazilian population in cancer research, providing resources that contribute to biomarker discovery, the development of new diagnostic and treatment strategies and, ultimately, the delivery of **precision medicine in SUS, Brazil's Unified Health System**.

Who is ATPBR for?

The platform was developed to serve different user profiles, including:

- Researchers interested in accessing and analyzing cancer data.
- Institutions and researchers who wish to share data in a standardized and secure way.
- Bioinformaticians and data scientists looking for data for their multi-omics analyses.
- Project managers and coordinators who need to follow the data life cycle.
- Collaborators involved in national cancer research initiatives.

The **Start here** chapter points out the best reading path for each profile.

What does ATPBR offer?

ATPBR brings together, in a single environment, resources that cover the entire data life cycle. Each of them is detailed in a chapter of this documentation:

- **Secure storage:** national infrastructure for storing and managing clinical and molecular data, with protection and access-control mechanisms (Platform architecture).
- **Data governance and protection:** policies that set out responsibilities, access criteria and conditions of use, ensuring the protection of participants and compliance with ethical and legal standards (Data governance).
- **Standardized submission:** mechanisms for registering projects and for the structured submission of data from different institutions and studies (Data submission).
- **Integration and harmonization:** standardization of clinical, sociodemographic and multi-omics data, enabling the integration and comparison of information produced by different technologies and institutions (Data model and Data dictionary).
- **Reproducible processing and analysis:** use of standardized, automated and versioned bioinformatics pipelines, ensuring the consistency, traceability and reproducibility of analyses (Analysis platform).
- **User-friendly, interactive exploration and visualization:** tools to query, explore and compare molecular and clinical data, making it easier to identify patterns and generate new scientific hypotheses (Results browser).
- **National sovereignty and responsible sharing:** data kept under Brazilian governance, with public access to non-sensitive information and controlled-access mechanisms for sensitive data, respecting usage authorizations, consents and current legislation (Data access).

Commitment to quality and trust

To ensure reliable and reusable data, ATPBR adopts the **FAIR** principles (Findable, Accessible, Interoperable and Reusable) and guides its development by the **international standards of the Global Alliance for Genomics and Health (GA4GH),** which promote the interoperability, governance and responsible sharing of genomic and clinical data. The platform sets three minimum criteria for including samples: (i) high-quality sequencing data, (ii) informed consent from participants/patients (or a waiver approved by the competent ethics body), in compliance with current ethical and regulatory standards, and (iii) a minimum set of clinical and sociodemographic metadata.

In compliance with the **Brazilian General Data Protection Law** (LGPD, Law No. 13,709/2018), ATPBR receives only data that have been **de-identified** beforehand by the participating institutions. The platform does not collect direct identifiers, nor certain variables that could increase the risk of participant re-identification, and the keys that link the codes to individuals remain at the institution of origin. Because the records of the same participant remain linked to one another, these data are **pseudonymized**, not anonymous: for this reason, they receive the protections that the LGPD provides for sensitive personal data. Data are stored and processed following information security and data governance practices appropriate to scientific research. These principles are revisited throughout the documentation and detailed in **Data governance** (ch. 4) and **Usage policies** (ch. 13).

In this way, ATPBR seeks to reconcile the protection of participants' privacy with the **promotion of open**, **transparent** and **reproducible** scientific research.

Learn more

- Official website: www.atpbr.com.br
- Contact: atpbr@inca.gov.br
- Reference article: Boroni M, Costa Conteville L, Ribeiro Pinto LF, P B Viola J. *The Brazilian Population Tumor Atlas Project: addressing representation in cancer genomics in Brazil.* Lancet Reg Health Am. 2026;56:101397. doi:10.1016/j.lana.2026.101397

CHAPTER 02

Start here

Welcome to ATPBR!

This section was created to help new users quickly understand the structure of the platform and find the resources they need for their activities. Whatever your profile, we recommend reading this chapter before the others.

How the documentation is organized

The chapters follow the **data life cycle** in ATPBR, from submission to the exploration of results. Each chapter builds on the previous one and points, at the end, to the next step:

1. **About ATPBR**: the project's objectives, mission and principles.
2. **Start here**: this chapter.
3. **Platform architecture**: an overview of the path that data take.
4. **Data governance**: who is responsible for each stage, how participants are protected and which authorizations each use requires.
5. **Data submission**: the step-by-step guide to registering a project and sending your data.
6. **Data access**: what is open, what is controlled and how to request access.
7. **Data model**: how the registered information is organized and interrelated.
8. **Data dictionary**: the fields of each module and the rules for filling them in.
9. **Identification and barcode system**: how the platform identifies patients, samples and assays.
10. **Data types**: the formats accepted for submission and the files generated by the platform.
11. **Analysis platform**: how submitted data are processed.
12. **Results browser**: how to explore the results in the ATPBR cBioPortal.
13. **Usage policies**: the responsibilities of those who use the data.
14. **Frequently asked questions**: quick answers to the most common questions.

Choose your profile

If you do not intend to read the documentation in order, start with the path closest to your needs:

I want to submit data

Read **Data governance**, which explains the responsibilities of those who submit data and the ethics documentation required. Then continue from **Data submission** through **Data types**. These chapters cover everything you need for registration.

I want to explore results

Start with the **Results browser**. To understand how these results were generated, see the **Analysis platform**.

I want to access controlled-access data

Read **Data governance**, **Data access** and **Usage policies**.

Need help?

See the **Frequently asked questions** section.
If you need further support, contact the ATPBR team by email at atpbr@inca.gov.br.

CHAPTER 03

Platform architecture

The ATPBR architecture was designed to support the entire life cycle of cancer data, from submission to analysis and the release of results to the scientific community. The platform consists of interoperable modules that work together to ensure traceability, standardization, scalability and reproducibility.

This chapter presents the **big picture**. Each stage described here is detailed in a chapter of its own, indicated throughout the text.

Architecture overview

In simplified form, the data flow in ATPBR follows these stages:





At every stage, the participant-protection principles presented in About ATPBR (ch. 1) are applied through role- and permission-based access control, activity logging and auditing, and the management of access requests for controlled-access data. The rules for each stage, and who is responsible for it, are set out in Data governance (ch. 4).

Data ingestion

Ingestion is the receipt of the metadata and sequencing files submitted by users. The step-by-step process is described in Data submission (ch. 5). Data are organized according to the Data model (ch. 7) and receive standardized identifiers, described in Identification and barcode system (ch. 9).

Validation

After ingestion, data undergo automatic and manual checks of integrity, format, required fields and consistency between metadata and files, as well as a review of the ethics documentation and usage permissions. Being accepted by the system is not the same as being admitted: while there are pending issues, the dataset remains blocked from release and from new uses. The checks and how to correct errors are described in **Data submission** (ch. 5), and the admission rules in **Data governance** (ch. 4).

Data storage

The architecture separates storage into **two complementary layers**, linked by a metadata catalog:

- **Raw data layer:** holds the large-volume files (the submitted FASTQ and the CRAM generated during processing), preserving the originals for traceability and future reanalysis.
- **Structured data layer:** holds the data processed and organized for querying, such as harmonized clinical data, genomic alterations and gene expression matrices.
- **Metadata and cataloging:** the catalog records projects, investigations, studies, patients, samples, assays, experiments, files and the processing runs performed. It is what makes it possible to locate a dataset, understand its origin and reconstruct the complete history of how the information was generated.

Data processing

Validated FASTQ files are processed by standardized pipelines, which identify somatic variants and quantify gene expression. The pipelines, their versions and the files generated are described in **Analysis platform** (ch. 11).

Harmonization and integration

Processing results are standardized, annotated and integrated with clinical data, which makes it possible to compare different studies. Non-sensitive harmonized data are loaded into the ATPBR cBioPortal.

Exploration and sharing

Results can be explored in the **Results browser** (ch. 12) (ATPBR cBioPortal). The projects' studies and datasets can be consulted in the platform's Data catalog, provided that their release has been authorized by those responsible for them. To learn more about the release criteria, the access levels (open or controlled) and the procedures for requesting data access, see the **Data access** (ch. 6) section.

CHAPTER 04

Data governance

In Platform architecture (ch. 3), we saw the technical path that data follows in ATPBR. This chapter covers the **rules** that apply along that path: who is responsible for each step, which conditions a dataset must meet to enter the platform and be shared, and how research participants are protected.

Read this chapter **before submitting data or requesting access to controlled-access data**. The following chapters apply these rules to each situation: Data submission (ch. 5) for those who send data, Data access (ch. 6) for those who request access and Usage policies (ch. 13) for those who receive the data.

ATPBR governance documents and instruments

The governance of the Brazilian Population Tumor Atlas (ATPBR) rests on institutional policies, participant protection instruments and terms that formalize responsibilities. Together, they cover the entire data life cycle, from submission to reuse, in compliance with Brazilian law and international good practice. The Status column indicates whether each document is approved, under review or in preparation; documents that have not yet been approved **must not be read as rules in force**.

Institutional policies

Document	What it establishes	Intended for	Status
Data Governance Policy	General governance principles; the role of the Brazilian National Cancer Institute (INCA) as the coordinating and maintaining institution; governance bodies (CGN, CCP, CAD, CESP and SAB) and the responsibilities of participating institutions; constitution and management of the database; participant protection; information security and oversight mechanisms.	All institutions, researchers and professionals involved in ATPBR.	Under review. It will be published in full after approval.
Data Flow and Access Policy	Procedures for submitting, verifying, classifying and making data available; access modalities (public, registered, controlled through data visiting, and internal); criteria for requesting and authorizing access, including pre-authorization through DUO codes; duties of the Data Access Committee; monitoring and termination of granted access.	Submitting researchers, access requesters, the Data Access Committee and platform operators.	Under review. It will be published in full after approval.

Data protection and participants' rights

Document	What it establishes	Intended for	Status
Personal Data Protection Impact Assessment (RIPD)	Identifies and assesses the risks, both individual and collective, of processing personal and sensitive data on the platform, and describes the security measures, safeguards and mechanisms adopted to prevent or mitigate them, in compliance with the LGPD (Brazilian General Data Protection Law).	Institutions responsible for processing the data, data protection officers, governance and information security teams and the competent supervisory authorities.	Under review. Internal document, because it contains information on risks and security controls; an institutional summary will be published.
Layered informed consent form (TCLE) template	Reference template that presents, in a clear and structured way, the purposes of collecting, storing, sharing and reusing the data in future research, with choices per layer (future use, international sharing, commercial use, recontact and return of findings) recorded as DUO codes. **It is not a general authorization:** each study must assess the compatibility between its consent, its ethics approval and the intended uses.	Researchers responsible for data collection, participating institutions, Research Ethics Committees and research participants.	In preparation. The template will be published; each study adopts the version approved by its Research Ethics Committee (CEP).

Responsibility and sharing terms

Document	What it establishes	Intended for	Status
Institutional Agreement	Formalizes each institution's membership in the ATPBR consortium and the prior agreements required by CNS (National Health Council) Resolution No. 738/2024 for databases built with the participation of more than one institution: data contribution, access and sharing, responsibilities, participation in the collegiate bodies and credit to the centers.	Participating institutions and INCA.	In preparation. The template will be published.
Submitting Researcher's Responsibility Agreement	Declarations and responsibilities of those who submit data to ATPBR, accepted on the platform **before the first submission**: legitimacy of the submission, compliance with ethics approvals, compatibility with the applicable consents and authorizations, quality of the information and compliance with the platform's policies.	Researchers responsible for submitting data.	Under review. It will be published in full after approval.
Controlled-Access Data Use Agreement (DAA)	Conditions for using authorized data in the ATPBR secure instance, through data visiting and without download: approved purpose, confidentiality, prohibition of attempts at re-identification and of extraction of individual-level data, export of verified results only, security measures, incident reporting, processing costs where applicable and conditions for terminating access.	Researchers and institutions authorized to access controlled-access data.	In preparation. The template will be published.
Data Transfer Agreement (DTA)	Legal, institutional and technical conditions for sending data from the source institutions to ATPBR: authorized purposes, responsibilities, protection measures, use restrictions and storage and retention conditions.	Source institutions that contribute data, and INCA.	In preparation. The template will be published.

The operating rules of the collegiate bodies are set out in their own bylaws, which are in preparation: the **CAD Bylaws** (composition, deadlines and appeals) and the **CCP Bylaws** (embargo period, authorship and credit to the centers).

Commitment to the responsible use of data

ATPBR adopts a responsible access approach that takes into account the nature of the data, its conditions of use and the rights of research participants.
The use permissions of each dataset are classified with the **Data Use Ontology (DUO)** of the Global Alliance for Genomics and Health (GA4GH), which standardizes use restrictions and how they are interpreted in access requests. The DUO classification helps in reading the permissions, but it **does not replace the verification of the consent and ethics approval** of each study.
Data sharing complies with the ethics approvals, the applicable legal bases and the relevant institutional instruments, ensuring that each dataset is used exclusively within the authorized conditions. Controlled-access data are analyzed **through data visiting**, in the ATPBR secure instance, **without download**.

These documents follow the legislation and ethical standards applicable to research involving human participants, including:

- Brazilian General Data Protection Law (LGPD, Law No. 13.709/2018).

- Law No. 14.874/2024, which governs research involving human participants, and its implementing regulations.
- CNS Resolutions No. 466/2012 (research involving human participants), No. 340/2004 (human genetics) and No. 738/2024 (databases for research).
- CD/ANPD Resolution No. 19/2024 (international transfer of personal data), issued by the ANPD (Brazilian Data Protection Authority).
- INCA's Governance, Cybersecurity and Crisis Management Policies.

ATPBR's policies **complement, but do not replace**, the ethics approval of each research project, the consent forms and the agreements signed between the institutions. Approved documents are published in full in this documentation, so that they can be consulted before accepting the Responsibility Agreement or submitting an access request; internal documents, such as the RIPD, have a public summary. Questions can be sent to atpbr@inca.gov.br.

Principles

Data management in ATPBR follows these principles, among others:

- **Scientific purpose:** data are used for legitimate, defined research objectives that are compatible with the conditions under which the data were obtained.
- **Participant protection:** respect for the dignity, autonomy, confidentiality and choices expressed by each participant, and prevention of collective risks, such as the stigmatization of population, ethnic, community or territorial groups.
- **Necessity and proportionality:** each use receives only the data appropriate to the authorized purpose.
- **Ethical compliance:** ethics approvals, consent forms and relevant ethics decisions are always observed.
- **Quality and scientific integrity:** data are kept intact, contextualized and traceable to their origin.
- **Security and prevention:** safeguards are proportional to the risks of improper access, re-identification and incompatible use.
- **Segregation of duties:** those who prepare, those who validate, those who decide and those who execute are distinct people or bodies.
- **Accountability:** decisions are recorded, so that the compliance of each operation can be demonstrated.
- **Transparency and interoperability:** the platform's rules are public, and the data follow standardized metadata, controlled vocabularies and identifiers.

Different authorizations that do not replace one another

A single dataset may depend on several authorizations. Each has its own purpose, and none of them waives the others:

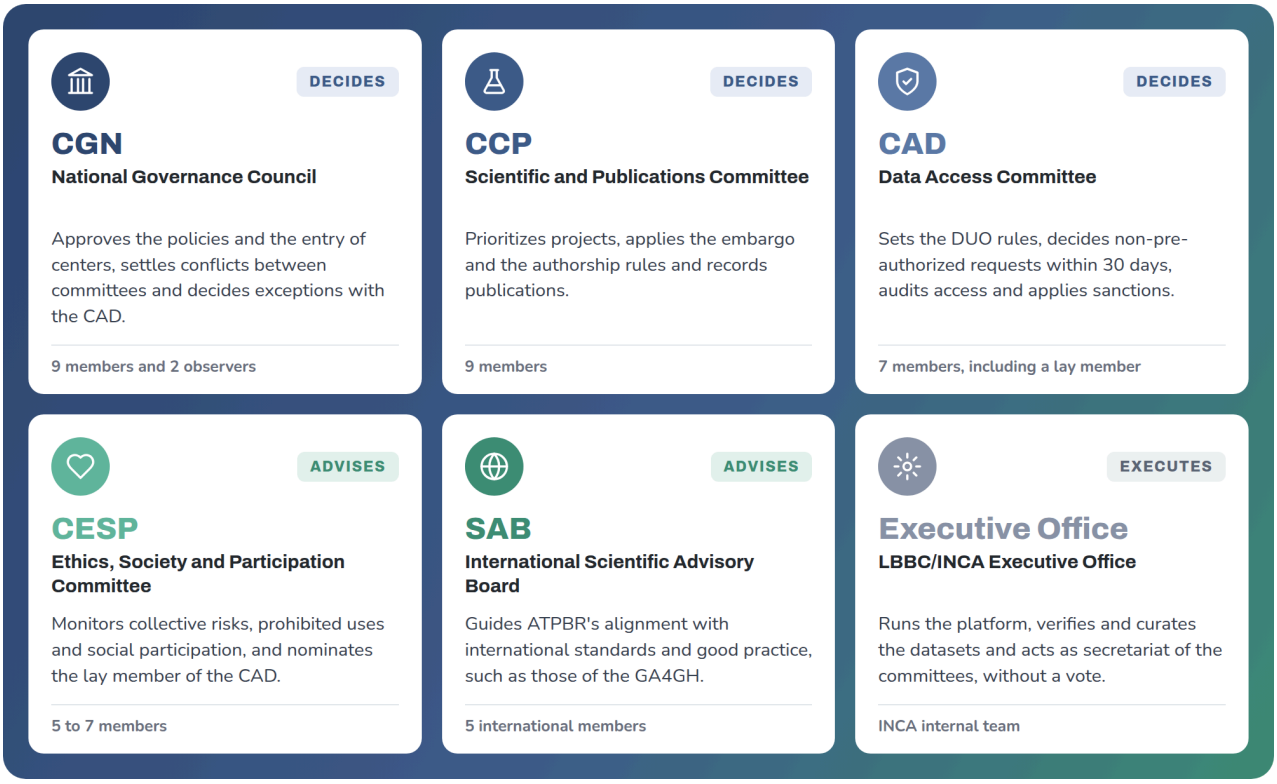
Authorization	Who grants it	What it guarantees
Ethics approval	Research Ethics Committee (CEP) or National Research Ethics Commission (Conep)	That the research, whether the original study or a new one, is ethically compliant
Participant consent (or its waiver)	The participant; the waiver can only be granted by the competent ethics body	That the participant authorized the use of their data, under the conditions of the signed form
Consent of the provider institution	An authorized representative of the institution that provided the data	That the source institution agrees to that particular sharing
Access authorization	ATPBR Data Access Committee (CAD)	That a specific controlled-access request is compatible with the conditions of the dataset

Two practical consequences:

- The ethics approval of the original study **does not, by itself, authorize** any new use or sharing of the data.
- Being accepted by the system **is not the same as** being admitted. Filling in the fields and uploading the files do not, by themselves, constitute authorization for incorporation, publication or scientific use.

Who's who

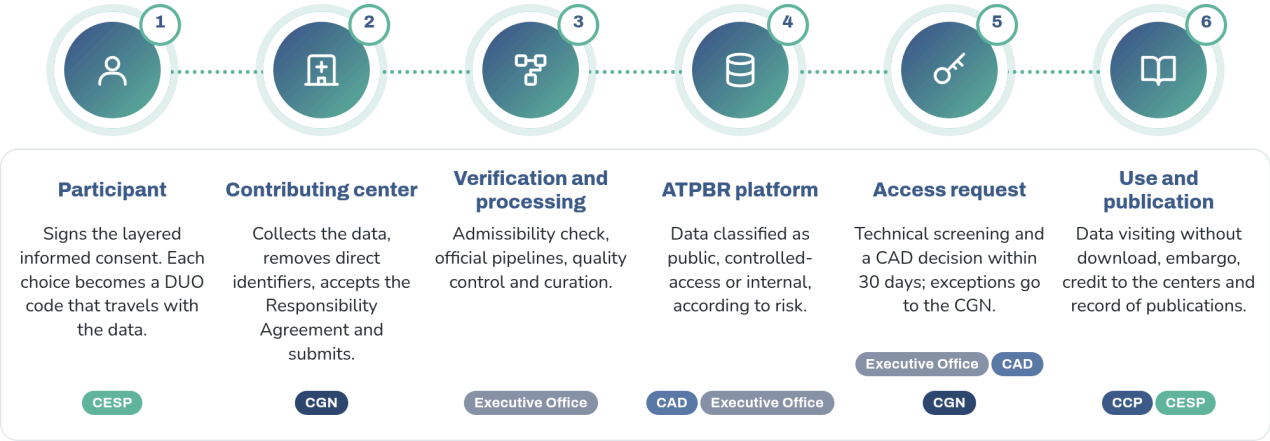
The consortium's governance rests on six bodies with complementary roles: three decide, two advise and one executes.



Participant	Main responsibilities
INCA: coordinating and maintaining institution	Developed ATPBR and maintains its infrastructure, operations and executive team (Executive Office/LBBC). Decisions on science, access and ethics are made by the consortium's collegiate bodies (CGN, CCP, CAD, CESP and SAB), in which the participating centers have a seat and a vote. For LGPD purposes, INCA acts as **controller** of the data processed on the platform, jointly with the participating centers, which remain controllers of the data they originate.
Source (provider) institutions	Ensure the compliance of collection and consent; remove direct identifiers and keep the **linkage keys** separately; report permissions and restrictions; communicate changes to and withdrawals of consent; cooperate in responding to participants and to incidents. For LGPD purposes, they remain controllers of the data they originate.
Submitting researchers	Prepare and send the data according to the platform's standards, provide the ethics documentation and the permissions of each dataset and accept the Submitting Researcher's Responsibility Agreement.
Executive Office and Operational Curation (LBBC/INCA)	Internal team that runs the operations: verifies the integrity, completeness, quality, provenance and ethics documentation of each dataset, records outstanding issues, runs the pipelines, curates data and metadata (including the DUO codes), supports the centers in submission, monitors and audits the access granted and acts as secretariat to the committees, **without a vote**. Reports to the CGN.
Technical and security teams	Implement the platform's controls: access profiles, authentication, encryption, audit logs, backups and vulnerability management.
National Governance Council (CGN)	**Highest decision-making body**, with 9 voting members and 2 observers (Decit/MS, the Department of Science and Technology of the Ministry of Health, and CNPq, the National Council for Scientific and Technological Development). Defines the strategy, approves the admission of centers and the general policies, resolves conflicts between the committees and decides, together with the CAD, exceptional cases such as commercial use and broad international transfer.
Scientific and Publications Committee (CCP)	**Decision-making**, with 9 members. Evaluates analysis proposals, defines priority projects, avoids overlap between groups, applies the embargo and authorship rules and registers manuscripts and presentations.
Data Access Committee (CAD)	**Decision-making**, with 7 members, including a lay member nominated by the CESP and a representative of the National Tumor Bank (BNT), who links the committee with the BNT and with the clinicopathological data. Defines the compatibility rules for DUO codes, decides controlled-access requests that are not pre-authorized **within 30 days** and may **suspend or revoke** access in case of non-compliance.
Ethics, Society and Participation Committee (CESP)	**Advisory**, with 5 to 7 members, including representatives of patients and of ethnic and racial minorities. Monitors collective risks, prohibited uses and social participation, and nominates the lay member of the CAD.
International Scientific Advisory Board (SAB)	**Advisory**, with 5 international members. Guides the alignment of ATPBR with international standards and good practice, such as those of the GA4GH.
Recipient researchers and institutions	Use the data only for the authorized purpose, maintain confidentiality and security, **do not attempt to re-identify participants** and comply with the signed use instrument; respect embargo periods and cite ATPBR and the source centers.

The roles are performed by people formally designated by INCA, with defined substitutes, and organized in a **responsibility matrix** that separates who provides the data, who verifies them, who decides on their incorporation or release and who executes and oversees. Members of the collegiate bodies serve 2-year terms, renewable once, declare **conflicts of interest** every year and for each agenda, and **whoever submits a dataset does not vote on it**. No member holds a seat on the CCP and the CAD at the same time, and there are no seats for companies.

The data path and who acts at each step



The tags below each step indicate the bodies that act on it.

Participant protection

The data received by ATPBR **do not include direct identifying information** about participants.

Each participant and sample receives a code, which makes it possible to integrate clinical and genomic information without directly identifying people. The linkage between codes and identities is protected and remains exclusively under the responsibility of the **source institution**.

All data are treated as **sensitive personal information**, in accordance with the General Data Protection Law (LGPD), with access controls and additional safeguards for genetic data.

Data protection rules

Before sending data to ATPBR, all institutions follow standardized rules:

- Removal of **personal identifiers** (name, CPF (Brazilian individual taxpayer number), National Health Card number, RG (identity card number), medical record number, registration number, address, telephone and email).
- Use of the patient and sample **codes** defined by ATPBR (see Identification and barcode system (ch. 9)).
- Replacement of clinical event dates with the **number of days** from diagnosis (see Data dictionary (ch. 8)).
- Checking of **free-text fields, file names, headers and technical metadata**, where identifiers often appear unnoticed.

Life cycle of a dataset

Step	What happens
1. Registration and upload	The submitting researcher accepts the Responsibility Agreement, registers the Project, Investigation, Study, Patients, Samples, Assay, Files and Experiments, and provides the ethics documentation and the use permissions (see Data submission (ch. 5)).
2. Admissibility check	Three distinct and complementary checks: (i) **automated**, of required fields, structure and vocabularies; (ii) **technical**, of integrity, consistency and correspondence between patients, samples, experiments and files; and (iii) **ethics and documentation**, of ethics opinions, consent and source permissions.
3. Validation result	The validation team records the dataset as **eligible**, **pending completion** or **not admitted**, with the reasoning and the person responsible for the decision.
4. Processing and curation	Eligible data are processed by the Analysis platform (ch. 11). Provenance is preserved, and source restrictions carry over to derived data.
5. Classification and release	Each item is classified as public, controlled-access or internal, and anything public undergoes a re-identification risk assessment before release (see Data access (ch. 6)). Before opening, the dataset may be subject to an **embargo period** (publication moratorium), during which first publication is reserved for the original project; the embargo and authorship rules are applied by the Scientific and Publications Committee (CCP).
6. Controlled access	Researchers request access, and the CAD decides case by case (see Data access (ch. 6)).
7. Retention and disposal	The need to retain data and documents is reviewed periodically, according to the scientific purposes, the legal and ethical obligations and the source conditions.

Pending or **not admitted** datasets **remain blocked**: they are not visible to third parties, are not published and cannot be used in new research until they are brought into compliance. Relevant changes after validation, such as a new document, a new restriction or new patients, trigger a **new verification** of the dataset.

Ethics documentation and permissions of each dataset

Each dataset is linked to the documents proving that it may enter the platform and be used. This set of records forms the dataset's **dossier**.

Level	What is reported and checked
Project	Title, scientific lead, participating institutions and ethics approval, with the supporting documents, their version and their validity
Study	Relevant ethics opinion, cohort characteristics, conditions for display in the catalog and study restrictions
Patients and Samples	Consent type, use permissions and restrictions, and identification codes
Assay, Experiments and Files	The conditions of use carry over to the data produced

The expected documents include the CEP or Conep opinion, the version of the informed consent form (TCLE) used (or the consent waiver decision), the data transfer agreement (DTA) and the record of use permissions. **A filled-in field without the corresponding document does not meet the proof requirement.**

The permissions recorded at each level (Study, Patient, Sample, Assay and File) are considered **together**. **The most protective restriction always prevails**, and a discrepancy between levels, for example a study released for general use with a patient who refused future use, **prevents release** until it is resolved and documented.

Permissions and restrictions follow the GA4GH **Data Use Ontology (DUO)**, which standardizes conditions such as general research use (GRU), health/medical/biomedical research (HMB), disease-specific research (DS), non-commercial use only (NCU), ethics approval required (IRB), collaboration with the provider researcher required (COL), geographic restriction (GS) and publication moratorium (MOR).

Future use of data

Using the data in new research requires respecting the conditions of the original study and the choices of each participant:

Status of the original consent	What is required
Future use was authorized	The new research may use the data under the authorized conditions, with its own ethics review
Future use was not foreseen	New consent or a waiver granted by the competent ethics body
Future use was expressly refused	Only with new consent from the participant or their legal representative

The CAD does not grant consent waivers: that power lies exclusively with the ethics body.

Data access and the Data Access Committee

What is public, what is controlled-access and what is for internal use, as well as the steps of a controlled-access request and how the CAD decides, are described in Data access (ch. 6). Two points apply in all cases:

- Working on software development or infrastructure maintenance **does not grant access** to research data. Internal permissions follow a profile matrix and the principle of least privilege.
- The CAD, whose composition is described in Who's who, meets monthly, and its members do not take part in deliberations from which they are disqualified. Quorum and other rules are set out in the **CAD Bylaws**.

Biological samples (outside the scope of ATPBR)

ATPBR **manages data only**. The governance of biological samples (collection, storage, access, use and transfer) is the responsibility of INCA's National Tumor Bank (BNT) and of the institutions that hold them, and is not regulated by ATPBR's policies; the same applies to instruments concerning samples, such as the Material Transfer Agreement (MTA). Authorization to access data **does not include access to the corresponding samples**.

The permissions and restrictions that apply to a sample **continue to apply to the data produced from it**.

Participants' rights and consent withdrawal

Participants exercise their rights (access, correction, restriction and withdrawal of their data) through the **source institution**, which holds the linkage keys. Requests are handled in coordination with INCA's data protection officer, and the platform locates the records **by their codes**, without needing to receive the participant's name or identity document.

When there is a consent withdrawal, a new restriction or another request from the participant:

- the affected datasets are identified by their source codes;
- new incompatible uses are blocked;
- researchers who have already received the data are notified, when necessary;
- the measures taken are documented, subject to the cases of retention provided for by law.

The source institution and the submitting researcher must **promptly notify** ATPBR of any change affecting the submitted data, such as a change in the ethics opinion, consent withdrawal, a new restriction or a security incident.

Security, incidents and traceability

The platform follows the INCA Cybersecurity Policy, with additional controls proportional to its risks: profile management, access limited to what each role needs, strong authentication for privileged access, encryption, backups and business continuity.

Records are kept of requests, decisions, consents, enablements, accesses, downloads, suspensions, revocations and terminations, **protected against improper alteration**. These records make it possible to reconstruct who accessed which dataset, when and on the basis of which authorization.

Security incidents are handled through INCA's privacy and security procedures, with the participation of the ATPBR team and the cooperation of the source institutions. Researchers who suspect an incident, re-identification or misuse must **notify INCA immediately**.

The risks of the platform's operations are assessed in ATPBR's **Personal Data Protection Impact Assessment (RIPD)**.

Monitoring and review

ATPBR's coordination tracks indicators on the compliance of datasets, release decisions, participant requests, incidents and corrective measures.

Relevant changes to the platform architecture, data categories, access features, participating institutions or conditions of use trigger an assessment of the impact on governance, the ethics protocol, the legal instruments and the security controls.

Controls are implemented progressively, but **each operation starts only after the controls it requires are in place**. For this reason, public release and controlled access are enabled only after their respective procedures have been completed.

PART 2

Submitting and sharing data

How to register projects and data on the platform and how they can be shared.

05 Data submission

06 Data access

CHAPTER 05

Data submission

Data submission is the process through which research centers, hospitals, laboratories and research groups bring projects and their data into ATPBR. It is the **ingestion** stage presented in Platform architecture (ch. 3): from there, the data move on to validation, processing and sharing.

This chapter describes **what to do and in what order**. After Data access (ch. 6), which explains how submitted data are shared, the following chapters detail how the information is organized (Data model (ch. 7)), how each field must be filled in (Data dictionary (ch. 8)), how identifiers are generated (Identification and barcode system (ch. 9)) and which formats are accepted (Data types (ch. 10)).

Who Can Submit Data

Data can be submitted by researchers, institutions and research centers that wish to bring studies into ATPBR, provided they meet the technical, ethical and governance requirements set by the platform (see the minimum criteria in About ATPBR (ch. 1) and the full rules in Data governance (ch. 4)).

Before sending data, the submitting researcher accepts the **Submitting Researcher's Responsibility Agreement**, shown in the submission area. By accepting it, the researcher declares, among other things, that:

- the data are authentic and the information provided is true, complete and up to date;
- the data have been de-identified according to ATPBR standards, and the keys linking the codes to the participants remain separate and protected at the originating institution;
- they will report the ethics approvals, the version of the informed consent form (TCLE) (or the consent waiver) and the use permissions and restrictions of each dataset, and provide the supporting documents;
- they will notify ATPBR of any changes to ethics approvals, consent withdrawals, new restrictions and security incidents they become aware of.

Also have at hand the ethics documentation for the project and the studies: the approval opinion of the Research Ethics Committee (CEP) or of the National Research Ethics Commission (Conep), the version of the TCLE used or the consent waiver decision, the data transfer agreement (DTA), if any, and the data use permissions and restrictions. The ATPBR team may request these documents during validation.

Step by Step

Submission is done through the platform portal. Each step depends on the previous one, and the project dashboard shows this order with numbers and arrows.

Step	What to register	How
1	**Project**, including the acronym and the associated institutions	Form
2	**Investigation** (research line)	Form
3	**Study** (one cancer type)	Form
4	**Patients** of the Study	Form or template spreadsheet
5	**Samples** from each patient	Form or template spreadsheet
6	**Assay** (one sequencing run)	Form
7	**Data Files** (FASTQ) of the Assay	Upload through the portal **or** path on the INCA Multi-user Bioinformatics Platform
8	**Experiments** (technical metadata of each sequenced sample)	Form or template spreadsheet

Some points to watch at each step:

- **Project:** the associated institutions entered here are the ones that will later appear as **collector center** options when registering patients.
- **Study:** within a single project, each cancer type can have only one Study.
- **Patients and Samples:** a sample can only be registered for a patient who is already registered. Clinical events are entered as a **number of days** from diagnosis, not as dates (see Data dictionary (ch. 8)).
- **Assay:** each sequencing run is a separate Assay. When registering it, you also indicate the desired **sharing conditions** for the Assay. This indication does not publish the data by itself: release depends on the validation of the dataset and on a separate approval (see Data access (ch. 6)).
- **Data Files:** the FASTQ files must be available **before** the Experiments are registered. There are two ways to do this:
  - **Upload through the portal:** in the *Data Files* area, select the Assay and upload the FASTQ files. Each file is linked to the Assay and inherits its sharing permissions.
  - **Files already stored at INCA:** if the FASTQ files are already in the repository of the **INCA Multi-user Bioinformatics Platform**, you do not need to upload them again. Just enter the **absolute path of the directory** containing the files in the Assay's *Associated Repositories* field.
- **Experiments:** an experiment can only be registered for a sample that is already registered. The `fastq_1` and `fastq_2` fields must contain the **exact name** of the Assay's FASTQ files. The platform warns you when a name you enter does not match any uploaded file.

Before Submitting

We recommend checking that:

- The FASTQ files are complete and intact, and in the accepted format (see **Data types** (ch. 10)).
- The required fields have been filled in and the controlled vocabularies have been used (see **Data dictionary** (ch. 8)).
- The patient and sample codes follow the de-identification rules (see **Identification and barcode system** (ch. 9)).
- There is no personal information or direct identifier of participants (name, CPF (Brazilian taxpayer ID), National Health Card number, medical record number, contact details) in the metadata, **in free-text fields such as descriptions and notes, in file names or in file headers**.
- The dates of clinical events have been converted into a number of days from diagnosis.
- The ethics documentation and the use permissions of each dataset are available (see **Data governance** (ch. 4)).
- The spreadsheets were filled in from the official templates, without changing their structure.

Meeting these requirements reduces errors and speeds up the validation process.

Data Validation

After submission, each dataset goes through three distinct and complementary checks, carried out by different people:

- **Automated check:** required fields, structure and use of the controlled vocabularies.
- **Technical check:** file format and integrity, consistency between patients, samples, experiments and files, and compatibility with the ATPBR data model.
- **Ethics and documentation check:** correspondence between the registered study, the protocol and the ethics approval, the consent or waiver conditions, and the use permissions and restrictions.

When registering through a spreadsheet, a preview shows the rows read and the errors found before saving, so you can review and confirm.

At the end, the validation team records the dataset as **eligible**, **pending completion** or **not admitted**, with the reasons. The fact that the data were accepted by the system **does not mean** that they were admitted: pending or not admitted datasets are blocked from release and from new uses until the issues are resolved.

Correcting Errors

If problems are detected during validation, the system will display messages indicating the errors found and the actions needed to correct them.

Depending on the type of inconsistency, you may need to:

- Update metadata.
- Correct naming.

- Upload files again.
- Complete required information.
- Adjust relationships between samples and experiments.
- Provide ethics documents or clarify use permissions.

After the corrections, the data can be resubmitted for a new validation. Submissions that do not meet the requirements of this documentation may be considered not admitted.

If, after submission, there is any change that affects the data (a new ethics opinion, a participant's consent withdrawal, a new use restriction or a security incident), notify the ATPBR team at atpbr@inca.gov.br. Relevant changes send the dataset to a new check.

CHAPTER 06

Data access

When you submit data to ATPBR (see **Data submission** (ch. 5)), you indicate how you would like them to be shared. This chapter explains the access levels, which balance scientific sharing, participant privacy and ethical and regulatory requirements, and describes how to request access to controlled-access data. The principles and responsibilities behind these rules are in **Data governance** (ch. 4).

Access levels

Level	What it includes	Who has access
Public	Descriptive metadata of studies and assays, catalog information and results approved for release, including somatic mutations and a minimum clinical set with grouped variables (age in bands, rounded survival, grouped race/color)	Anyone
Registered	Lower-risk datasets, classified for this level according to criteria approved by the CAD (Data Access Committee)	Researchers with verified identity and institutional affiliation, after accepting the standard terms of use; each access is logged, without individual deliberation by the CAD
Controlled	Sequencing files, intermediate results, germline variants, individual-level genomic data, complete omics matrices and clinical data that require additional protection	Researchers authorized by the Data Access Committee, under the approved conditions, or by pre-authorization, when the purpose is fully compatible with the DUO codes of all the datasets (rules defined by the CAD).
Internal operational	Datasets under validation or processing and administrative information	Authorized ATPBR teams, according to their roles

Who decides what is shared

When registering each **Assay**, the person responsible for the project indicates whether that assay can be made available in ATPBR (see **Data submission** (ch. 5)). This indication is the starting point, but it is not enough: release only happens after the dataset is validated as **eligible** and the publication is approved, with an assessment of re-identification risks. The permissions recorded on the Study, the Patients, the Samples, the Assay and the Files are considered together, and the most restrictive one always prevails.

When the participants' consent, the source permissions and ATPBR policies allow it:

- The assay **metadata** (title, description and omic type) are included in the platform's **Data catalog**.
- The **analysis results** approved for release can be published in the ATPBR cBioPortal.
- The **FASTQ and CRAM files** remain under controlled access.

Public-access data

Results approved for public release in the ATPBR cBioPortal (see **Results browser** (ch. 12)) can be viewed and downloaded without any additional authorization.

The platform's **Data catalog** lists the studies and datasets of all projects, with the cancer type, the omic type and the number of patients, samples and experiments. For each dataset, the catalog shows whether you already have access (because you are part of the project or were granted access) or whether authorization is required.

All public release follows these rules:

- Before publication, small cohorts, rare attributes, combinations of clinical, demographic and genomic filters, and successive queries that could allow individual information to be reconstructed are assessed.
- When the risk is not acceptable, the information is aggregated, shown in bands or omitted, or the dataset moves to controlled access. By default, on the public portal, age at diagnosis appears in 5-year bands, survival in rounded months, race/color in grouped categories, and counts below the minimum threshold are suppressed. Germline variants, raw reads and dates are never public. The thresholds and techniques are defined by the ATPBR coordination, with technical validation.
- Each approval records the person responsible, the dataset version, the publication parameters, the tests performed and the date. Relevant changes to the data, filters or indicators require a new assessment.
- Approving the release of the descriptive record does **not** release the sequencing files, the individual clinical records or the associated omics matrices.

Controlled-access data

Some datasets have access restrictions because of their potential for identification, their sensitivity or the terms of consent. This category includes, among others:

- Raw or aligned sequencing files (FASTQ and CRAM).
- Intermediate processing results.
- Germline variants and other individual-level genomic data.
- Complete omics matrices or matrices not approved for release (processed and approved matrices appear in the Results browser).
- Individual clinical data beyond the variables published in the Results browser.
- Clinical information whose granularity or combination of attributes requires additional protection.
- Other data defined by the ATPBR governance policies.

Access to these resources requires a formal request, reviewed by the ATPBR **Data Access Committee (CAD)**. CAD authorization does not replace the ethics approval of the new research nor the consent of the institution that provided the data, when required.

Internal operational access

Datasets under validation or processing and administrative information are accessed only by authorized ATPBR teams, according to their roles and the principle of least privilege. Working on software development or infrastructure maintenance does not grant access to research data. Privileged or exceptional access requires justification, authorization, a limited time frame and logging, and is reviewed afterwards.

How to request access to controlled-access data

Researchers who want to access controlled-access data must register the request in the access request module of the ATPBR platform, after logging in with an institutional account. Questions about the process can be sent to atpbr@inca.gov.br. The team will provide guidance on the request process and the documentation needed, according to the nature of the data and the access policies in force.

What the request must contain

- Identification and institutional contact details of the principal researcher and of their institution.
- Title, objectives, scientific justification, methodology and expected duration of the research.
- The datasets, variables, files and level of detail requested, with the justification of the need for each data category.
- The protocol and the ethics approval of the new research, when required, and the consent waiver decision, if applicable.
- The team that will have access, the processing and storage environment, the security measures, the country of remote access, if any, and whether copying or extraction is possible.
- The commitment to confidentiality, to no re-identification, to no onward transfer and to respecting the source conditions.

Each collaborator must be individually identified, be linked to the project and accept the same obligations. Merely registering as a user on the platform does not grant access to restricted materials.

Request steps

Step	Responsible	Outcome
1. Request	Requesting researcher	Request registered, with an identification number, or returned for completion
2. Screening	Team designated by ATPBR	Request prepared for review, follow-up query (request for additional information) or block
3. Compatibility	Validation team and coordination	Opinion on the compatibility of the request with the informed consent form (TCLE), the protocol and the permissions of each dataset

Step	Responsible	Outcome
4. Source consent	Authorized representative of the provider institution	Statement recorded or proof of prior consent
5. Deliberation	CAD	Reasoned decision, with conditions and time frame
6. Formalization	Researcher, their institution and INCA	Data use agreement signed
7. Enablement	Authorized technical team	Access activated and logged, per user
8. Monitoring	Coordination and technical team	Review, continuation or suspension of access
9. Closure	Technical team and coordination	Access deactivated and disposition of the data recorded

How the CAD decides

The CAD compares, **for each dataset**, the declared purpose with the permissions and restrictions recorded at the source (see Data governance (ch. 4)), without extending to one dataset the permissions of another. Automated comparison of DUO codes provides support but does not decide: inconclusive cases, conflicts and exceptional requests are always assessed by people.

The CAD may approve the request in full; approve it in part, limiting datasets, variables, people, purpose or time frame; approve it under conditions; issue a follow-up query requesting additional information; or reject it, with reasons. The decision is communicated to the requester through a traceable channel.

The source instrument defines whether the provider institution gives its position **case by case** or has given **prior consent** for compatible uses. In the first case, the position must be explicit: **silence does not count as consent**. In the second, it is checked whether the request is fully covered by the permissions already formalized.

After the decision

- Before release, the researcher and, when applicable, their institution sign the **data use agreement** approved by INCA, which cannot extend the source authorizations.
- Access is **individual**, linked to the project and to the authorized datasets, with an expiration date (as a rule, 12 months, with a notice before expiry) and without shared accounts.
- **Renewal is not automatic**: it depends on new proof of the eligibility conditions.
- Access may be **suspended as a precaution** in the face of signs of misuse, re-identification risk or a change in the source permissions, and **revoked** by the CAD in case of non-compliance.
- At **closure**, the credentials and the processing instance are deactivated. The data remain in ATPBR; the researcher keeps only the results whose export has been checked and authorized.
- Requests for reconsideration may be submitted on the grounds of documentary correction, new facts or material error, but they do not grant provisional access.

Access to data by a researcher or institution abroad depends, in addition to the other conditions, on the source permissions for international sharing and on a transfer mechanism permitted under the LGPD (Brazilian General Data Protection Law).

The obligations of those who receive the data are in Usage policies (ch. 13).

PART 3

Data organization

The data model, the dictionary, the identification system and the accepted formats.

07 Data model

08 Data dictionary

09 Identification and barcode system

10 Data types

CHAPTER 07

Data model

In Data submission (ch. 5), we saw the order in which information is registered. The data model explains **why** this order exists: it defines how information is organized and related on the platform, ensuring traceability, interoperability and consistency throughout the entire data life cycle.

Overview

ATPBR uses a hierarchical structure made up of four main modules:

Project → Investigation → Study → Assay

This organization connects the scientific context of the research to the patients, the biological samples, the experimental procedures and the data files generated.

The **Study** and **Assay** modules also contain specific elements:

- **Study:** Patients and Samples.
- **Assay:** Experiments and Data Files.

Main Modules

Project

Represents the overall scientific context of the research and is the highest level of the hierarchy. It brings together one or more related research lines and defines the scientific objectives, the scope and the participating institutions.

Investigation

Represents a research line within a Project. It groups studies related by a theme, scientific objective or hypothesis, and may cover different cancer types.

Study

Represents a scientific unit dedicated to characterizing a single cancer type within an Investigation. It defines the study design, the cohorts, the selection criteria and the clinical characteristics of the participants.

It contains two elements:

- **Patients:** study participants, each with a single record on the platform, regardless of the number of associated samples.
- **Samples:** biological materials collected from the patients. A patient may be associated with several samples.

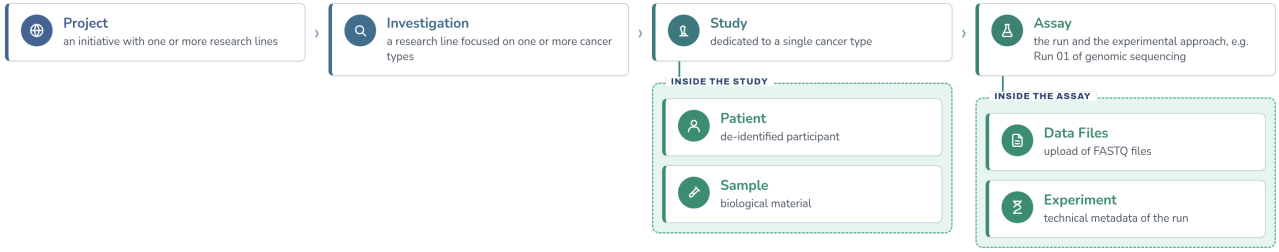
Assay

Represents an experimental run used to generate omics data from the samples of a Study. Distinct runs, even when they use the same omics approach, must be registered as separate Assays. A Study may contain several Assays.

It contains two elements:

- **Data Files:** raw files produced in the experimental run, such as FASTQ (see **Data types**), used as input for processing on the platform.
- **Experiments:** records of the technical metadata associated with each sample processed in the run, including sequencing platform, library preparation and read layout. Each Experiment links a Sample to its Data Files.

How the Elements Relate



The figure shows the hierarchy of the modules and the relationships between their elements.

Besides belonging to a module, certain elements depend on previously registered records:

- A **Sample** must be associated with a **Patient**.
- An **Experiment** must be associated with a **Sample** and with the corresponding **Data Files** available in the Assay.

These dependencies determine the order of registration on the platform and make it possible to trace experimental data back to the sample and patient of origin.

The fields of each module and element are described in the next chapter. Some of them also make up the unique identifier of patients, samples and assays, presented in **Identification and barcode system** (ch. 9).

CHAPTER 08

Data dictionary

Now that you know the modules and elements of the **Data model** (ch. 7), this chapter describes **what to fill in for each one**. The data dictionary defines the structure, meaning and completion rules of the data stored on the platform, ensuring consistency and comparability across the data submitted by the different participating centers.

Each field has a formal definition, an expected type and format, a required or optional status and, where applicable, controlled vocabularies or reference ontologies. Some of these fields also make up the ATPBR **Identification and barcode system** (ch. 9).

Organization of the Data Dictionary

The data dictionary is organized according to the platform's four main modules:

- Project:
  - Required fields:
    - Project Title, Acronym, Description, Principal Investigator (PI), Associated Institutions, Ethics Approval, Funding
  - Optional fields:
    - Co-PIs, Scientific Committee, Start Date, End Date
  - Fields linked to ontologies:
    - Associated Institutions (linked to the ROR ID)
  - Fields used to compose the barcode:
    - Acronym (5 to 7 uppercase letters), Associated Institutions (reference to the center that collected the samples)
- Investigation:
  - Required fields:
    - Investigation Title
  - Optional fields:
    - Description
- Study:
  - Required fields:

    - Study Title, Associated Investigation, Cancer Type, Description, Specific Ethics Approval
  - Optional fields:
    - Publications
  - Fields with reference ontologies:
    - Cancer Type (linked to OncoTree)
  - Within a single project, each cancer type can have only one Study.
- Assays:
  - Required fields:
    - Assay Title, Omic Type
  - Optional fields:
    - Description, Associated Repositories
  - Fields used to compose the barcode:
    - Omic Type
  - Field for files already stored at INCA:
    - Associated Repositories: if the FASTQ files are already in the repository of the INCA Multi-user Bioinformatics Platform, enter the absolute path of the directory here (see **Data submission** (ch. 5)).
  - The Assay also states the sharing conditions requested for its data, which are subject to validation and approval (see **Data access** (ch. 6)).

Following the three main elements of the data model of the data ingestion platform, each element has a specific set of clinical, biological and experimental attributes. The full description of all attributes is available in the spreadsheet ATPBR_planilha_metadados.xlsx

Where applicable, each attribute described in this spreadsheet includes:

- Attribute name
- Description
- Data type
- Required or optional status
- Allowed values
- Controlled vocabulary or ontology

Only the main required fields of each element are presented below. The full description is in the metadata spreadsheet.

- **Patients**
  - Main required fields:
    - `patient_id`: Participant code, 4 digits long, defined by the collector center and unrelated to personal data
    - `collector_center`: Collector center (institution), via the "Collector Center" vocabulary
    - `disease_type`: The patient's cancer type, chosen from the OncoTree classification among the types of the tissue defined in the Study

    - `primary_site`: Organ where the tumor originated
    - `consent_type`: Type of consent obtained for participation in the study
    - `data_use_permission`: Primary permission for the use of this patient's data/samples
    - `data_use_restriction`: Additional restriction on data use according to the GA4GH Data Use Ontology (DUO). A single value; use "NRES" when there is no additional restriction
  - Optional field related to the diagnosis: `disease_subtype`: OncoTree subtype, below the cancer type, when known
  - Automatic field: `cid10_category`: ICD-10 (CID-10) category calculated by the platform from `disease_type`, `disease_subtype` and `primary_site`. The user does not fill in this field.
  - Fields used to compose the barcode: `patient_id`, `collector_center`, `cid10_category`
- **Samples**
  - Main required fields:
    - `sample_id`: Sample code, 4 digits long, defined by the collector center and unrelated to personal data
    - `patient_id`: Identifier of the already registered Patient to whom this sample belongs
    - `sample_type`: Biological material or biological derivative corresponding to the collected sample.
    - `sample_status`: Pathological classification of the sample at the time of collection.
    - `sample_preservation`: Method used to preserve the sample after collection.
  - Fields used to compose the barcode: `sample_id`, `sample_type`, `sample_status`, `sample_preservation`
  - A Sample requires at least one staging system to be filled in (see the section *Mandatory Sample Staging*, below).
- **Experiments**
  - Main required fields:
    - `sample_id`: Identifier of the already registered Sample from which this experiment was generated
    - `analyte_type`: Type of analyte used to generate the sequencing data
    - `library_kit`: Identification of the kit used for library construction
    - `sequencing_center`: Center where the sequencing was performed
    - `sequencing_platform`: Identification of the platform used to sequence the samples
    - `date_of_sequencing`: Date on which the sequencing was performed, in DD/MM/YYYY format
    - `library_layout`: Type of read used
    - `lane_number`: Sequencing lane number
    - `fastq_1`: Exact name of the read 1 FASTQ file, as listed in the Assay's Data Files
  - Conditional field: `fastq_2`: Exact name of the read 2 FASTQ file. Required when `library_layout` = PAIRED.
  - Fields used to compose the barcode: `lane_number`. The combination of `date_of_sequencing`, `sequencing_center` and `sequencing_platform` defines the run number (see **Identification and barcode system** (ch. 9)).

Mandatory Sample Staging

Every sample registered in ATPBR must provide **at least one staging system**. The platform rejects the registration of a sample that has **none** of the following fields filled in:

- **FIGO** (gynecological tumors)
- **TNM** (general system)
- **Rai** (chronic lymphocytic leukemia)
- **Binet** (chronic lymphocytic leukemia)
- **LMC Phase** (chronic myeloid leukemia)
- **AML/ALL Stage** (acute leukemias)
- **CIN** (cervical intraepithelial neoplasia)

You do not need to fill in all of them: the staging system that applies to the case is enough. The AJCC staging fields are optional and do not count toward this rule. If none of the systems above is provided, the sample is highlighted in the registration table with the message: *"Provide at least one staging system (FIGO, TNM, Rai, Binet, LMC Phase, AML/ALL Stage or CIN)."*

Dates and time: information as a number of days

To protect the identity of participants, ATPBR **does not accept calendar dates** for clinical events, such as date of birth, diagnosis, collection, treatment or death. Combined with one another, these dates would make it possible to re-identify the patient.

Instead, temporal fields take the **number of days** counted from a common reference point, the **diagnosis**, which corresponds to **day 0**. The day of diagnosis is the date of the first exam that concluded malignancy.

Examples:

Field	What to enter
`diagnosis_age_in_days`	Patient's age at diagnosis, in days
`days_to_sample`	Days between diagnosis and sample collection
`days_to_treatment_first_course`	Days between diagnosis and the start of the first treatment
`days_to_last_follow_up`	Days between diagnosis and the patient's last record (last follow-up or death)

For example, if the sample was collected 15 days after diagnosis, enter `15` in `days_to_sample`.

To convert the dates in your institution's medical records, use the **date-to-days converter**, available in the platform's registration forms. The conversion must be done before filling in the forms, and the original dates must not be sent to ATPBR.

The only calendar date accepted is the **sequencing date** (`date_of_sequencing`), because it does not refer to the patient. It must be in DD/MM/YYYY format, with a 4-digit year, be a valid date and cannot be in the future.

Versioning

The ATPBR data dictionary is kept under version control and updated whenever necessary to accommodate new technologies, omic modalities, clinical fields, reference ontologies and regulatory requirements.

All changes are recorded and documented, ensuring the traceability of changes and backward compatibility whenever applicable.

CHAPTER 09

Identification and barcode system

Several fields described in the **Data dictionary** (ch. 8), such as the project acronym, the collector center, the patient and sample codes and the omics type, do more than describe the data: they combine to form the **barcode**, the unique identifier of each patient, sample and assay. Users do not type the barcode; the platform generates it automatically from these fields.

ATPBR uses a standardized identification system to ensure the traceability of participants, biological samples, experiments and generated data throughout the entire life cycle of a project or study. This system is based on structured barcodes, which work as unique and persistent identifiers within the platform.

Inspired by the The Cancer Genome Atlas (TCGA) barcode, the ATPBR identification system was adapted to meet the specific needs of Brazilian cancer research, incorporating requirements for governance, de-identification, multi-omics integration and biobank management.

The barcode is the main mechanism linking the biological material stored in the National Tumor Bank (BNT), the metadata recorded on the platform and the files generated during the analytical processes. In this way, the same identifier follows the sample from its collection to the production of molecular results and their release for research.

Principles of the Identification System

The ATPBR barcode system was developed on the basis of four fundamental principles:

Unique Identification

Each participant, sample and assay receives a unique combination of codes, eliminating ambiguity and allowing the unequivocal identification of the objects recorded on the platform.

Traceability

All data produced in ATPBR can be linked to their biological and experimental origin. The system makes it possible to reconstruct the complete history of a sample, from its collection, storage and processing to the generation of the associated omics files.

De-identification

The identifiers contain no personal information or data that would allow the direct identification of participants, and the correspondence between the codes and the individuals is kept only at the institution of origin. Because the barcode deliberately links all records of the same participant, the data are **pseudonymized**, not anonymous: for this reason, they remain protected as sensitive personal data (see Data governance (ch. 4)).

Interoperability

The barcode structure was designed to facilitate integration between the different ATPBR modules and to promote compatibility with standards widely used in biomedical data repositories and platforms.

Biological Material and Digital Objects

The same barcode is used to connect different components of the ATPBR infrastructure.

In the physical environment, it identifies materials stored in the National Tumor Bank, including tissue, blood, plasma and other biological samples.

In the digital environment, the identifier is used to associate:

- Clinical metadata.
- Experimental information.
- FASTQ files.
- Processed files.
- Analytical results.
- Processing and provenance records.

This strategy ensures consistency between the different layers of the platform and makes it possible to trace the origin of any data available in ATPBR.

Barcode Structure

The ATPBR identifier **is not a single code**: it is built **incrementally and in a nested way**, following the hierarchy of the Data model (ch. 7) (**Patient → Sample → Assay**). Each level fully reuses the barcode of the previous level and adds its own fields.

Level	What is identified	Composition	Example
Patient	the study participant	`PROJECT-CANCER-CENTER-PARTICIPANT`	`ATPBR-56-001-1234`
Sample	the collected biological material	`‹patient barcode›-SAMPLE-TYPE`	`ATPBR-56-001-1234-5678-01AB`
Assay	the omics experiment performed on the sample	`‹sample barcode›-ASSAY`	`ATPBR-56-001-1234-5678-01AB-G0101`

For this reason, the "complete" **seven-field** barcode exists only at the **assay** level: it is the result of stacking the three levels. Read from left to right, separated by hyphens (`-`), the complete structure is:

`PROJECT-CANCER-CENTER-PARTICIPANT-SAMPLE-TYPE-ASSAY`

Example:

`ATPBR-56-001-1234-5678-01AB-G0101`

Each component of the identifier is described in the following sections.

Interpreting the Barcode

Each segment of the ATPBR barcode has a specific meaning and represents a piece of information associated with the project, participant, sample or experimental assay.

Taking the example:

`ATPBR-56-001-1234-5678-01AB-G0101`

The interpretation is as follows:

Field	Value	Meaning
Project	ATPBR	Brazilian Population Tumor Atlas
Cancer (ICD-10)	56	Malignant neoplasm of ovary (C56)
Participating Center	001	Brazilian National Cancer Institute (INCA)
Participant	1234	Participant code
Sample	5678	Biological sample identifier
Sample Type	01AB	Frozen malignant tumor tissue
Assay	G0101	Whole genome sequencing (WGS), run 01, lane 01

Therefore, the barcode above identifies a sample of frozen malignant tumor tissue from a patient with ovarian cancer, collected at INCA in the context of the ATPBR project and submitted to a whole genome sequencing assay.

Meaning of the Codes

- **Project Code**

This is the **project acronym**, defined by the user when registering the Project. It must have **5 to 7 uppercase letters** and is unique on the platform: each project has its own.

Example:

Code	Project
ATPBR	Brazilian Population Tumor Atlas

- **Cancer Code (ICD-10)**

Identifies the participant's cancer diagnosis by the **category of the International Classification of Diseases (ICD-10, CID-10 in Portuguese)**, represented by the 2 digits of the code, without the letter "C". For example, category C56 (malignant neoplasm of ovary) enters the barcode as `56`.

> **How this code is defined:** the user **does not type** the code. In the **Patient registration**, the user provides the cancer type (OncoTree classification, among the types of the tissue defined in the Study), the subtype, when known, and the organ of origin of the tumor. From these answers, the platform automatically calculates the corresponding ICD-10 category, and that value is what enters the barcode. For hematologic neoplasms (leukemias, lymphomas, myeloma), the category is defined by the subtype, regardless of the organ.

Examples of codes:

Code	ICD-10 category	Description
15	C15	Esophagus
16	C16	Stomach
18	C18	Colon
20	C20	Rectum
34	C34	Bronchus and lung
43	C43	Melanoma
44	C44	Other neoplasms of skin
50	C50	Breast
53	C53	Cervix uteri
56	C56	Ovary
61	C61	Prostate
91	C91	Lymphoid leukemia
92	C92	Myeloid leukemia
99	(not applicable)	Non-malignant neoplasm (benign, in situ or of uncertain behavior)

The ICD-10 malignant neoplasm categories (C00 to C97) are accepted, plus code `99` for non-malignant neoplasms. The complete list is available in the metadata spreadsheet. The codes are stable: codes already in use are never changed.

- **Participating Center Code**

Identifies the **collector center**, the institution responsible for collecting or submitting the sample. Each center receives a **3-digit** code.

How a center enters the system:

1. The institution is **registered by the user when creating the Project**, together with its **ROR** identifier.
2. From then on, it appears in the **"Collector Center" controlled vocabulary** and becomes available for selection **when registering a patient**.
3. The code recorded in the barcode is that of the institution selected in this vocabulary.

The codes are global and permanent:

- Each institution has **a single fixed code across the entire platform**, uniquely identified by its **ROR id**, so that the same real center always resolves to the same code.
- A **new code is generated only when the institution has not been selected on the platform before**; centers already in use reuse the existing code.
- Codes are **never reused or removed**, so that the barcodes of samples already registered remain valid.

Predefined centers:

Code	Institution	ROR id
001	Brazilian National Cancer Institute (INCA)	055n68305
002	Cancer Institute of the State of São Paulo (ICESP)	005vqqr19
003	Ceará Cancer Institute (ICC)	01pg9bv14

New centers receive the next available sequential code, in the order in which they are first used.

- **Participant Code**

A **4-digit** code that identifies the participant. It is **defined by the collector center** and entered in the `patient_id` field of the patient registration.

Example:

Code	Meaning
1234	Participant code

The code must have no relation whatsoever to the participant's personal data: for example, do not use the medical record number or parts of the date of birth.

- **Sample Code**

A **4-digit** code that identifies the biological sample linked to a participant. It is also **defined by the collector center** and entered in the `sample_id` field of the sample registration.

Example:

Code	Meaning
5678	Specific biological sample

The same participant may have multiple samples. The platform does not accept two records with the same barcode and warns you when it detects a possible duplicate registration.

- **Sample Type Code**

The Sample Type field is made up of four characters and describes, at the same time, the biological material, the pathological status and the preservation method of the sample.

Structure:

`TTSP`

Where:

- **TT** = Material type (2 digits)
- **S** = Sample status (1 letter)
- **P** = Sample preservation (1 letter)

For example, `01AB` corresponds to:

- **01** = *Tissue*
- **A** = *Malignant Tumor*
- **B** = *Frozen*

Result: **Frozen malignant tumor tissue**.

> **Language of the values:** on the platform, the labels of these vocabularies appear **in English**, to keep all projects consistent. The tables below show the official label. What enters the barcode is always the **code** (the left-hand column), not the text.

Material Type Codes

Code	Sample Type
01	Tissue
02	Blood
03	Plasma
04	Serum
05	Bone Marrow
06	Peritoneal Fluid
07	Synovial Fluid
08	Pericardial Fluid
09	Cerebrospinal Fluid
10	Bronchial Lavage
12	Urine
13	Saliva
14	Sputum

Code	Sample Type
15	Feces
16	Skin
17	Cells
18	Breast Secretion
19	Pleural Fluid
20	Extracellular Vesicles
21	Aspiration Biopsy
22	Buffy Coat

Code 11 is not used.

Sample Status Codes

Code	Sample Status
A	Malignant Tumor
B	Benign Tumor
C	Lesion
D	Metastasis
E	Normal - Adjacent
F	Normal - Distant
G	Lymph Node - Metastatic
H	Lymph Node - Reactive
I	Inflammation
J	Cell Line
K	Normal Non Tumoral

Sample Preservation Codes

Code	Sample Preservation
A	Fresh
B	Frozen
C	FFPE (Formalin-Fixed, Paraffin-Embedded)
D	Paraformaldehyde (PFA)

Code	Sample Preservation
E	RNAlater
F	Culture Medium
G	Saline Buffer
L	Saline Solution

> **Anticoagulants:** the collection tube additive (for example, EDTA or citrate) is entered in a separate field, `collection_tube_additive`, which is optional and **does not enter the barcode**.

Interpretation Examples

Code	Interpretation
01AB	Frozen malignant tumor tissue
01AC	Malignant tumor tissue preserved in FFPE
01EB	Frozen adjacent normal tissue
02KB	Frozen normal non-tumoral blood
22KB	Frozen normal non-tumoral buffy coat
17JF	Cell line maintained in culture medium
20AB	Frozen malignant tumor extracellular vesicles

The combination of these codes makes it possible to describe, in a compact and standardized way, the main biological and operational characteristics of each sample, while maintaining traceability throughout the entire ATPBR workflow.

- **Assay Code**

The **Assay** field identifies the omics modality used and information associated with the experimental run. This field makes it possible to distinguish different experiments performed on the same sample and to maintain the traceability of the generated data.

Structure:

`OCCLL`

Where:

- **O** = Omics modality (1 character)
- **CC** = Experimental run number (2 digits)
- **LL** = Lane number (2 digits)

For example:

`G0101`

corresponds to:

- **G** = Whole Genome Sequencing (WGS)
- **01** = First run
- **01** = Lane 1

Result:

Whole Genome Sequencing (WGS), run 01, lane 01

Omics Modality Codes

Code	Omics Modality
G	Whole Genome Sequencing (WGS)
X	Whole Exome Sequencing (WES)
R	RNA-Seq
P	Gene Panel Sequencing

Codes M (DNA methylation) and B (5-base DNA methylation) are reserved for future versions (see the accepted modalities in **Data types** (ch. 10)).

Run Number

A sequential number **assigned automatically by the platform**. The user **does not set** this number manually.

The run is identified by the combination of three fields entered in each Experiment: **sequencing date**, **sequencing center** and **sequencing platform**. Numbering is done within each **Study**:

- The first combination recorded in the Study receives `01`, the second receives `02`, and so on.
- Experiments with the same date, the same center and the same platform belong to the same run and receive the same number.
- Once assigned, the run number does not change.

For this reason, these three fields are required and must be filled in carefully. Besides distinguishing technical replicates and new runs of the same experiment, this identification **makes later batch-effect analyses easier**, where it is important to know which data were generated in different runs.

Code	Meaning
01	First run
02	Second run
03	Third run
...	Subsequent runs

Lane Number

Represents the sequencing lane used during the run.

Code	Meaning
01	Lane 1
02	Lane 2
03	Lane 3
04	Lane 4
...	Other lanes

Identifying the lane supports technical traceability and the investigation of possible sequencing-related problems.

Interpretation Examples

Code	Interpretation
G0101	WGS, run 01, lane 01
X0102	WES, run 01, lane 02
R0201	RNA-Seq, run 02, lane 01
P0101	Gene panel, run 01, lane 01

Examples of Complete Barcodes

Barcode	Interpretation
ATPBR-56-001-1234-5678-01AB-G0101	Frozen malignant tumor tissue sample from ovarian cancer submitted to WGS in run 01, lane 01
ATPBR-50-002-3456-7890-01AC-R0101	FFPE-preserved malignant tumor tissue sample from breast cancer submitted to RNA-Seq in run 01, lane 01
ATPBR-61-003-4321-8765-02KB-X0201	Frozen normal non-tumoral blood sample from prostate cancer submitted to WES in run 02, lane 01

The combination of the assay code with the other barcode fields makes it possible to uniquely identify the biological and experimental context associated with the data produced in ATPBR, ensuring traceability from the physical sample to the files generated by sequencing.

References

- National Cancer Institute. Genomic Data Commons. *TCGA Barcode*. GDC Docs. Available at: https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

CHAPTER 10

Data types

The previous chapters dealt with **metadata**: how they are organized (Data model (ch. 7)), how they are filled in (Data dictionary (ch. 8)) and how they generate the identifiers (Identification and barcode system (ch. 9)). This chapter deals with **formats**: what ATPBR accepts at submission and what the platform generates from it.

Overview

ATPBR accepts two types of data for submission:

Category	Accepted format
Metadata	Portal forms or Excel template spreadsheets (.xlsx or .xls)
Raw sequencing data	gzip-compressed FASTQ (.fastq.gz or .fq.gz)

All other files are generated internally by the platform and **must not be submitted**.

Metadata

Metadata can be recorded in two ways:

- **Portal forms:** registration of one record at a time (one patient, one sample or one experiment). Projects, Investigations, Studies and Assays are always registered through forms.
- **Template spreadsheets:** registration of several patients, samples or experiments at once. Download the template from the registration page itself, fill it in and upload the file. Before saving, a preview shows the rows read and the errors found.

When using spreadsheets, use only the templates provided by ATPBR and keep their original structure (do not rename, remove or reorder columns). The filling rules for each field are the same as in the forms and are described in the Data dictionary (ch. 8).

Raw Sequencing Data

Sequencing data must correspond to the primary files generated by the sequencing instruments.

Accepted modalities

In this first version, the following omics modalities are accepted:

- Whole genome sequencing (**WGS**).
- Whole exome sequencing (**WES**).
- **Gene panel** sequencing.
- RNA sequencing (**RNA-seq**).

Other modalities, such as DNA methylation, will be added in future versions.

FASTQ file requirements

- Files compressed in gzip format (.fastq.gz or .fq.gz).
- For paired-end reads, both files of each sample: read 1 and read 2.
- Linked to an Assay: uploaded through the portal or located on the INCA Multi-user Bioinformatics Platform (see **Data submission** (ch. 5)).
- Names identical to those entered in the `fastq_1` and `fastq_2` fields of the Experiments.
- File names free of personal data, such as the patient's name or medical record number.
- Integrity verified by checksum.

Data Generated by ATPBR

From the FASTQ files, the platform internally generates:

- Alignment files (CRAM).
- Somatic variants (VCF and MAF).
- Gene expression matrices.
- Copy number data.
- Quality control metrics.

The pipelines that generate these files are described in the next chapter.

PART 4

Data analysis and use

How the data are processed, how the results are explored and the responsibilities of those who use them.

11 Analysis platform

12 Results browser

13 Usage policies

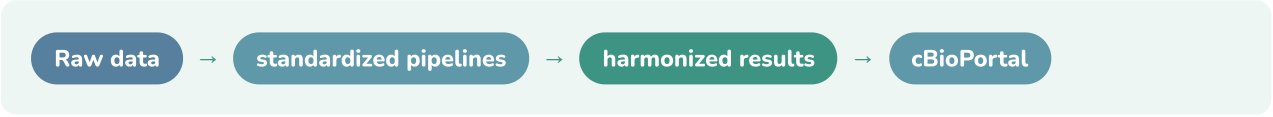
CHAPTER 11

Analysis platform

The FASTQ files accepted at submission (see **Data types** (ch. 10)) are processed by the **ATPBR Analysis Platform (ATPBR-AP)**. This chapter describes which pipelines are used, how they are configured and what results they produce.

Overview

Pipeline execution is **centrally managed by the ATPBR team**: users do not need to configure or run any analysis. This ensures that all data are processed in the same way, which makes it possible to compare different studies.



The pipelines are community workflows from nf-core, built in Nextflow, run in containers and with fixed versions and parameters.

Data	Pipeline	Main result
WGS, WES and gene panel	`nf-core/sarek` 3.10.0	Annotated and filtered somatic variants
RNA-seq	`nf-core/rnaseq` 3.18.0	Gene expression matrices

Quality Control

Quality control makes it possible to assess the characteristics of the sequencing reads and to identify potential problems that could compromise downstream analyses.

The tools used include:

- FastQC: assesses the quality of the raw reads, including per-base quality, GC content, duplication and presence of adapters.
- fastp: filters and preprocesses the reads, including removal of adapters and low-quality sequences, according to the configured parameters.
- MultiQC: consolidates metrics and reports from different tools, making it easier to assess the samples together.

The results of the two tools are combined in a **MultiQC** report. The main metrics assessed are:

Metric	What is expected
Per-base quality (Phred)	Most bases in the green zone (Q ≥ 30). A small drop in the first and last bases of the read is normal.
GC content	Bell-shaped distribution, close to 41% for DNA and 50% for mRNA (Piovesan et al., 2019; Palma et al., 2025). Multiple peaks suggest contamination.
Unidentified bases (N)	Less than 5%. Higher values indicate a problem with sequencing quality.
Duplication	Interpretation depends on the assay type, depth and library complexity. High levels may be expected in WES, gene panels and RNA-seq.
Adapter content	Less than 0.1% after trimming by fastp.
Filtered reads (fastp)	fastp discards reads with more than 40% of bases with Q < 15, with more than 5 N bases, shorter than 15 bp after trimming, or made up of repetitive sequences (for example, poly-G tails).

WGS, WES and Gene Panel Analysis

DNA sequencing data are processed with `nf-core/sarek` , version 3.10.0, with ATPBR-specific settings. The workflow comprises the following steps:

1. **Quality control and preprocessing** with FastQC and fastp.
2. **Alignment** to the GRCh38 reference genome with BWA-MEM. Reads from different lanes are merged per sample and saved in CRAM format.
3. **Preprocessing** of the alignments with GATK: duplicate marking and base quality score recalibration.
4. **Somatic variant calling**: identification of single nucleotide variants (SNVs) and small insertions and deletions (indels) with Mutect2, under two strategies:
    - **Paired samples (tumor-normal):** uses data from the tumor and from a matched normal sample from the same patient, which makes it possible to distinguish somatic from germline variants with greater precision.
    - **Unpaired samples (tumor-only):** uses tumor data only, without a matched normal sample. In this case, the identification of somatic variants relies more heavily on population filters and artifact detection resources, and germline variants may remain among the results.
5. **Annotation** of the effects of each variant on genes, transcripts and proteins with Ensembl VEP.
6. **Filtering** of artifacts and of potentially germline variants, according to the analysis strategy:
    - **Paired samples:** the tumor-normal comparison makes it possible to identify and exclude variants present in the matched normal sample, complemented by the Mutect2 quality filters.
    - **Unpaired samples:** the filter removes variants present in the 1000 Genomes panel of normals (PoN), variants with a population frequency above 1% in gnomAD and variants with a frequency above 0.1% in the Brazilian population, according to ABraOM. ABraOM filtering is performed by an in-house pipeline of the ATPBR team, after `nf-core/sarek` .

Configuration

`nf-core/sarek`	Version 3.10.0
Reference genome	*Homo sapiens* iGenomes GATK GRCh38
Trimming	fastp
Aligner	BWA-MEM
Variant caller	Mutect2 (tumor-normal and tumor-only)
Functional annotation	Ensembl VEP
Filters	PoN, gnomAD, 1000 Genomes and ABraOM, depending on the step and the strategy

Post-processing of somatic variants

After `nf-core/sarek` has run, scripts developed by the ATPBR team process the annotated VCF files to select variants, organize genomic information and add complementary population data.

The documented processing comprises the following steps:

1. **Selection of PASS variants:** Variants whose FILTER field in the VCF has the value PASS are selected, indicating that they met the filtering criteria applied earlier.
2. **Conversion to a table:** the VEP annotation fields and the tumor read counts (depth DP, reads per allele AD and allele frequency AF) are extracted into one table per sample. The 1000 Genomes frequency reported by VEP is renamed `GMAF`.
3. **Annotation with ABraOM:** each variant receives the allele frequency observed in the Brazilian population.

RNA-seq Analysis

The `nf-core/rnaseq` pipeline processes short-read RNA-seq (total RNA, poly(A)-enriched or rRNA-depleted) in the following steps:

1. **Concatenation** of the FASTQ files of a single sample sequenced across several lanes.
2. **Quality control and trimming** with FastQC and fastp.
3. **Pseudoalignment and quantification** with Salmon, which assigns reads directly to the known transcripts of the reference transcriptome. Expression is estimated per transcript and summed per gene.
4. **Normalization** to Transcripts Per Million (TPM), which corrects for differences in library size and transcript length.

By default, no additional filtering steps are applied. If quality control indicates problems, the team may adjust specific parameters (see *Limitations*, below).

Configuration

`nf-core/rnaseq`	Version 3.18.0
Reference genome	*Homo sapiens* Ensembl Release 113 GRCh38
Trimming	fastp
Pseudoalignment and quantification	Salmon

Limitations

- **Modalities:** only the modalities listed in **Data types** (ch. 10) are processed.
- **Organism:** the pipelines are configured for human data only. Samples from non-human or mixed organisms, such as patient-derived xenografts, are not supported.
- **Germline variants:** there is no germline-only analysis, because Mutect2 is optimized for somatic variants. In addition, germline variants are controlled-access data (see **Data access** (ch. 6)).
- **Data heterogeneity:** the data come from several centers, with differences in sample handling, library kits and sequencing protocols. To include rare samples of compromised quality, the team may adjust the pipeline parameters. These adjustments, as well as differences between runs, can introduce **batch effects**. That is why the run number is part of the barcode (see **Identification and barcode system** (ch. 9)).

CHAPTER 12

Results browser

ATPBR cBioPortal

The results generated by the Analysis platform (ch. 11) reach users through the **ATPBR cBioPortal**, an interactive environment for browsing, visualizing and analyzing cancer genomic and clinical data without requiring advanced bioinformatics knowledge. This chapter describes what is available in the portal and how to explore it.

ATPBR uses **cBioPortal for Cancer Genomics, version 7.2**, an open-source platform widely used by the international scientific community for the integrated exploration of molecular and clinical cancer data.

The portal is aimed at translational researchers, molecular biologists, clinicians, bioinformaticians and data scientists interested in the molecular landscape of Brazilian tumors.

1. Available data

cBioPortal brings together, in a single study, the data layers generated along the ATPBR workflow:

Layer	Source	What it lets you explore
Somatic mutations	Variants filtered and annotated by `nf-core/sarek` (WGS, WES, gene panel)	Recurrently altered genes, mutation types and the distribution of alterations across cases.
Copy number	Segmented copy-number data, when available	Genomic gains and losses, broad or focal.
Gene expression	TPM matrices from `nf-core/rnaseq` (RNA-seq)	Expression differences between genes, samples and groups, and the transcriptional effect of genomic alterations.
Clinical and sample data	Metadata registered at submission (see Data dictionary (ch. 8))	Cohort filtering, stratification of analyses and clinical interpretation of findings.

Not every study will have all of these layers: availability depends on the assays performed, the quality of the samples and the release stage. Check the description of each study in the portal.

2. How to navigate

The typical exploration path has four steps:

1. **Choose a study** from the list of studies. Each study corresponds to a curated dataset, usually one tumor type.
2. **Get to know the cohort in the Study View**, which summarizes the number of patients and samples, the available data types and the distribution of the main clinical and molecular variables.
3. **Filter the cohort** by the available attributes, such as tumor type, histology, sample type, staging or sequencing availability.
4. **Query genes** of interest and examine the results in the visualizations described below.

The studies and profiles you can see vary according to each user's permissions and the release stage of the data (see Data access (ch. 6)).

3. Main features

Feature	What it is for
Study View	Overview of the study and starting point for applying filters to the cohort.
Query By Gene	Search for one or more genes in the selected cohort and see how they are altered. By default, it includes non-synonymous mutations, fusions, amplifications and deep deletions.
OncoPrint	Sample-by-sample visual summary of the alterations in the queried genes. Helps identify recurrent events, co-occurring and mutually exclusive alterations.
Mutations tab	Details of each variant: type, position in the protein and affected samples. Useful for identifying hotspots.
Copy number and expression	Explore amplifications, deletions and expression levels, when these data exist in the study.
Plots tab	Compare molecular measurements with clinical annotations, for example the expression of a gene by tumor subtype.
Group Comparison	Compare molecular and clinical characteristics between user-defined subcohorts.
Survival analysis	Outcome plots, when the study includes vital status and follow-up time.
Patient View / Sample View	All clinical and molecular information for a case, including multiple samples from the same patient.
Onco Query Language (OQL)	More precise queries, specifying mutation classes, exact variants, copy-number states or expression thresholds.
Download and sharing	Export tables and results for analysis in R or Python, and save or share queries. Depends on the user's permissions.

4. Use cases

The portal is especially useful for:

- Exploring the frequency and distribution of somatic alterations in selected genes in the Brazilian cohorts.
- Comparing molecular patterns between clinical subgroups, such as subtype, staging or histology.

- Investigating co-occurring or mutually exclusive alterations that suggest shared biological pathways.
- Generating hypotheses about biomarkers and molecular signatures for future validation.
- Checking, before an external analysis, whether a cohort has the required molecular profiles.
- Reviewing individual cases in their molecular and clinical context.
- Preparing figures and summaries for meetings, reports and publication planning.

5. How data reach the portal

Studies are prepared and loaded exclusively by the ATPBR data team. Pipeline results and registered metadata are harmonized, converted to the formats required by cBioPortal and validated before loading. When new assays or corrections become available, the study is reviewed, revalidated and reloaded.

Each study in the portal is made up of the following groups of files:

File group	Role in the portal
Study metadata	Study identifier, name, description, citation and visibility.
Patient clinical data	Patient-level attributes (Table 1).
Sample clinical data	Sample-level attributes (Table 2).
Mutation profile	Somatic mutations.
Expression profile	mRNA and transcript expression.
Copy number	Segmented copy-number data.
Case lists	Samples available for each type of analysis.

Table 1. Patient-level fields.

Attribute	Type	Description
PATIENT_ID	Text	Patient identifier, which links the patient's clinical data to the samples and molecular data.
SAMPLES_PER_PATIENT	Number	Number of samples from the patient in the study.
SEX	Text	Sex of the patient.
RACE	Text	Race/color of the patient, in grouped categories.
COUNTRY	Text	Country associated with the patient, as defined by the study.
DIAGNOSIS_AGE	Number	Age at diagnosis, in 5-year bands.
OS_STATUS	Text	Overall survival status (living or deceased).
OS_MONTHS	Number	Overall survival time in rounded months, from diagnosis.

Table 2. Sample-level fields.

Attribute	Type	Description
PATIENT_ID	Text	Patient to whom the sample belongs.
SAMPLE_ID	Text	Sample identifier, which links it to the molecular data.
CANCER_TYPE	Text	Cancer type.
CANCER_TYPE_DETAILED	Text	Detailed cancer type.
ONCOTREE_CODE	Text	OncoTree code of the disease.
SAMPLE_TYPE	Text	Specimen type, such as primary tumor or metastasis.
TUMOR_STAGE_2009	Text	Tumor stage.
TREATMENT_DRUG_NAME	Text	Drug or treatment regimen, when available.
TREATMENT_RESPONSE	Text	Response to treatment.
WGS_DATA	Text	Indicates whether the sample has whole genome sequencing (WGS) data.
WES_DATA	Text	Indicates whether the sample has whole exome sequencing (WES) data.
PANEL_DATA	Text	Indicates whether the sample has gene panel data.
RNASEQ_DATA	Text	Indicates whether the sample has RNA-seq data.
METHYLATION_DATA	Text	Indicates whether the sample has methylome data.

6. Privacy

All data in the portal are de-identified before release, in line with the principles presented in **About ATPBR** (ch. 1). De-identified does not mean anonymous: that is why only studies validated and approved for release are published, after an assessment of re-identification risks (see **Data governance** (ch. 4)). Use of the data obtained from the portal, including the prohibition of re-identification attempts and the citation rules, is governed by the **Usage policies** (ch. 13). Results should be interpreted taking into account the scope and limitations of each cohort.

CHAPTER 13

Usage policies

Access to ATPBR data, whether open through the **Results browser** (ch. 12) or controlled (see **Data access** (ch. 6)), is subject to the policies described below. They complement the principles of de-identification and compliance with the LGPD (Brazilian General Data Protection Law) presented in **About ATPBR** (ch. 1) and the rules of **Data governance** (ch. 4).

User responsibilities

Authorized users undertake to:

- Use the data exclusively for the approved research, within the approved timeframe and under the approved conditions;
- Not attempt to identify participants, their relatives or the institutions of origin, nor combine ATPBR data with other sources for that purpose;
- Not contact participants on the basis of the records received;
- Not share credentials, not make copies available to unauthorized persons and not retransmit the data to third parties;
- Not use the data for unauthorized purposes, including clinical care or diagnostic purposes. Commercial use is only permitted when participants' consent allows it (datasets without the DUO codes NCU or NPU), under the same access rules, with no exclusivity or patent over the data and with benefit sharing; exceptional cases are decided by the Data Access Committee (CAD) together with the National Governance Council (CGN);
- Not use the data for forensic or police identification, for risk selection or pricing by insurers or employers, for targeted advertising, or to support racial, ethnic or ancestry hierarchies;
- Not remove security controls or transfer the data to unauthorized environments;
- Adopt security measures proportionate to the risks, such as individual authentication, limited privileges, protection of credentials, use of the data exclusively within the ATPBR secure instance and no attempt to extract individual-level data;
- Respect the restrictions set at origin, which continue to apply to derived data, as well as any publication moratoria;
- Not publish results that allow participants to be identified, directly or indirectly. Authorization to analyze the data is not authorization to publish individual records;
- Cite ATPBR and the source studies, acknowledging the centers that contributed samples and data, in publications and products resulting from the use of the data (the reference article is in **About**

ATPBR (ch. 1)).

Users of controlled-access data also undertake to:

- Report to INCA, through the channel and within the deadline set out in the data use agreement, security incidents and any suspected re-identification or misuse;
- Give prior notice of relevant changes to the purpose, methods, team, participating institutions, processing environment or requested datasets, which may require a new decision by the Data Access Committee;
- At the end of the access period or upon revocation of access, delete or return the data, or keep them only in the authorized form, and confirm this to ATPBR.

Failure to comply with these conditions may result in the suspension or revocation of access and in other measures provided for in the data use agreement and in ATPBR policies. Applying these measures is the responsibility of the Data Access Committee (CAD), and the decisions are recorded.

Access to data does not include access to the corresponding biological samples, whose governance falls under the National Tumor Bank (BNT) and is not regulated by ATPBR.

Transparency and traceability

Requests, decisions, institutional sign-offs, enablements, accesses, suspensions, revocations and closures are recorded, with protection against unauthorized alteration, for the purposes of auditing, traceability and compliance with ATPBR governance policies (see **Data governance** (ch. 4)). A public list of approved requests, with the project title, the institution and the authorized datasets, is published periodically on the ATPBR website.

CHAPTER 14

Frequently asked questions

This page gathers the most common questions about using ATPBR. If your question is not here, contact the ATPBR team by e-mail at atpbr@inca.gov.br.

About ATPBR

? Does ATPBR receive data that identifies patients?

Not directly. ATPBR only receives data that has been de-identified beforehand by the participating institutions: names, identity documents, medical record numbers, contact details and dates of clinical events must not be sent, and the link between the codes and the individuals is kept only at the institution of origin. Even so, the data is pseudonymized, not anonymized, and is therefore protected as sensitive personal data. Learn more in **Data governance** (ch. 4) and **Usage policies** (ch. 13).

? Which types of omics data are accepted?

WGS, WES, gene panel and RNA-seq. Other modalities, such as DNA methylation, will be added in future versions. See **Data types** (ch. 10).

Registration and submission

? In what order should I register the information?

Project → Investigation → Study → Patients → Samples → Assay → Data Files → Experiments. Each step depends on the previous one. The step-by-step guide is in **Data submission** (ch. 5), and the reason for this order is in **Data model** (ch. 7).

? Can I register several patients, samples or experiments at once?

Yes, using the template spreadsheets (Excel, `.xlsx` or `.xls`), which you can download from the registration page itself. See **Data types** (ch. 10).

? My collector center does not appear in the list on the patient registration form. What should I do?

The collector center must be among the Project's **Associated Institutions**. Edit the Project and add the institution; from then on, it appears in the "Collector Center" list on the patient registration form. See **Identification and barcode system** (ch. 9).

Why can't I create a second Study with the same cancer type?

Within a single project, each cancer type can have only one Study. New patients with that cancer type must be registered in the existing Study.

Why doesn't the cancer type I am looking for appear on the patient registration form?

The patient's cancer type options are limited to the tissue defined in the Study. If the type you are looking for belongs to another tissue, it must be registered in the corresponding Study.

Dates and time

Why can't I enter dates, such as the date of diagnosis or collection?

Combined with one another, the dates of clinical events make it possible to re-identify the patient. That is why temporal fields take the **number of days** counted from the diagnosis. See Data dictionary (ch. 8).

What is "day 0"?

It is the date of diagnosis, that is, of the first exam that concluded malignancy. All other events are counted from it. For example, a sample collected 15 days after diagnosis must be entered with the value `15` in `days_to_sample`.

How do I convert the dates in the medical record into a number of days?

Use the **date-to-days converter**, available in the platform's registration forms. The conversion is done on your own computer, and the original dates must not be sent to ATPBR.

Is there any date that must be entered as a date?

Yes, only the **sequencing date** (`date_of_sequencing`), in DD/MM/YYYY format, because it does not refer to the patient.

Identifiers and barcode

Who defines the `patient_id` and the `sample_id`?

The collector center. Both are 4-digit codes and must be unrelated to personal data: for example, do not use the medical record number or parts of the date of birth.

Do I need to fill in the barcode?

No. The barcode is generated automatically by the platform from the fields entered when registering the patient, the sample and the experiment. See Identification and barcode system (ch. 9).

Why does "XXX" appear in the barcode?

"XXX" indicates that one of the fields used to compose the barcode has not been filled in yet or could not be identified. Check the record's required fields.

Why does the run number appear as "00"?

The run number is assigned from the combination of sequencing date, sequencing center and sequencing platform. As long as any of these three fields is not filled in, the run appears as `00`.

I received a duplicate registration warning. What does it mean?

The platform does not accept two records with the same barcode. The warning indicates that a patient, sample or experiment with the same identification data already exists. Check whether the record was registered earlier or whether the code you entered is correct.

Samples

My sample was rejected for missing staging. Why?

Every sample must provide at least one staging system (FIGO, TNM, Rai, Binet, LMC Phase, AML/ALL Stage or CIN). The system that applies to the case is enough; the AJCC fields do not meet this requirement. See Data dictionary (ch. 8).

Where do I enter the collection tube anticoagulant (EDTA, citrate)?

In the `collection_tube_additive` field (collection tube additive), which is optional. This field is not part of sample preservation and is not included in the barcode.

Sequencing files

In what format should I submit the sequencing files?

Gzip-compressed FASTQ (`.fastq.gz` or `.fq.gz`). For paired-end reads, submit both files for each sample. See Data types (ch. 10).

Why can't I register the experiments before submitting the FASTQ files?

Each experiment indicates, in the `fastq_1` and `fastq_2` fields, which files belong to that sample. That is why the files must be available in the Assay before the experiments are registered. See Data submission (ch. 5).

My FASTQ files are already on the INCA Multi-user Bioinformatics Platform. Do I need to submit them again?

No. Enter the absolute path of the directory where the files are stored in the Assay's *Associated Repositories* field. See Data submission (ch. 5).

I received a warning that a FASTQ file was not found. What does it mean?

The name entered in `fastq_1` or `fastq_2` does not match any file in the Assay. Check that the file was submitted to the correct Assay and that the name was typed exactly the same, including the extension.

Can the FASTQ file name contain the patient's name or the medical record number?

No. File names must not contain personal data either. Rename the files before submitting them, if necessary.

Do I need to submit BAM, CRAM or VCF files?

No. These files are generated by the ATPBR pipelines from the submitted FASTQ files. See Analysis platform (ch. 11).

Do I need to run any pipeline?

No. The pipelines are run centrally by the ATPBR team. See **Analysis platform** (ch. 11).

Governance and ethics

What is the Submitting Researcher's Responsibility Agreement?

It is the agreement the researcher accepts before submitting data to ATPBR. In it, the researcher declares that the data is authentic and has been de-identified, that they will provide the ethics documentation and the use permissions of each dataset, and that they will report changes, such as consent withdrawals or new restrictions. See **Data submission** (ch. 5).

My study has ethics approval. Is that enough to share the data in ATPBR?

Not necessarily. The ethics approval of the original study does not, by itself, authorize new uses or sharing. You must check whether the informed consent form (TCLE) and the protocol allow inclusion in ATPBR and future use of the data, and whether the institution of origin authorizes the sharing. See **Data governance** (ch. 4).

The system accepted my data. Is it already available to other researchers?

No. Being accepted by the system is not the same as being admitted. Each dataset goes through technical and ethical-documentary checks and only becomes available after it is recorded as eligible. Pending datasets remain blocked until the issues are resolved. See **Data submission** (ch. 5).

I filled in the ethics approval field. Do I need to send any document?

Yes. The completed field does not replace the supporting document. The ATPBR team may request the approval opinion of the Research Ethics Committee (CEP) or of the National Research Ethics Commission (Conep), the version of the TCLE (or the consent waiver decision) and the data transfer agreement. See **Data governance** (ch. 4).

A participant has withdrawn consent. What should I do?

Notify the ATPBR team by e-mail at atpbr@inca.gov.br, identifying the participant **only by the code** used on the platform (`patient_id` or barcode), never by name or identity document. ATPBR blocks new uses of the data and, when necessary, notifies the researchers who have already received it. See **Data governance** (ch. 4).

Does data access include the biological samples?

No. Access to data and access to or transfer of samples are separate authorizations. Decisions about samples rest with the institutions that hold them. See **Data governance** (ch. 4).

Data access

Which data can I consult without requesting authorization?

The information approved for public release: the metadata in the Data catalog and the results published in the **Results browser** (ch. 12) (ATPBR cBioPortal).

Who decides whether my project's data appears in cBioPortal?

The person responsible for the project indicates, when registering each Assay, whether the data can be made available. Publication, however, only happens after the dataset is validated and the release is approved, with an assessment of re-identification risks and in compliance with the participants' consent and the original permissions. FASTQ and CRAM files always remain under controlled access. See **Data access** (ch. 6).

How do I obtain access to controlled-access data, such as FASTQ, CRAM or germline variants?

Submit an access request in the access request module of the ATPBR platform, using your institutional login (questions: atpbr@inca.gov.br). The access request is reviewed by the Data Access Committee (CAD) and, if approved, release depends on signing the data use agreement. CAD authorization does not replace the ethics approval of your research. See **Data access** (ch. 6).

Can my collaborators use my access?

No. Access is individual and accounts cannot be shared. Each collaborator must be identified in the access request and accept the same obligations. See **Usage policies** (ch. 13).

How should I cite ATPBR?

Cite ATPBR and the studies from which the data originates. The reference article is in **About ATPBR** (ch. 1), and the usage rules are in **Usage policies** (ch. 13).

Brazilian Population Tumor Atlas

Expanding the representation of the Brazilian population in cancer research by providing resources that contribute to the discovery of biomarkers, the development of new diagnostic and treatment strategies and the delivery of precision medicine in SUS, Brazil's Unified Health System.

WEBSITE **www.atpbr.com.br**

CONTACT **atpbr@inca.gov.br**

Funding: Ministry of Health, Brazil, in partnership with CNPq

GENOMAS BRASIL INCA SUS MINISTÉRIO DA SAÚDE