

Dados FAIR na prática: Desafios e soluções na partilha de dados genómicos

Dora Henriques



Data Talks: Formação e competências para a gestão e dados FAIR

Objetivos

- Mostrar como projetos de genómica aplicam princípios FAIR
- Ilustrar boas práticas de gestão de dados no IPB
- Partilhar experiências reais em cumprimento de políticas de dados



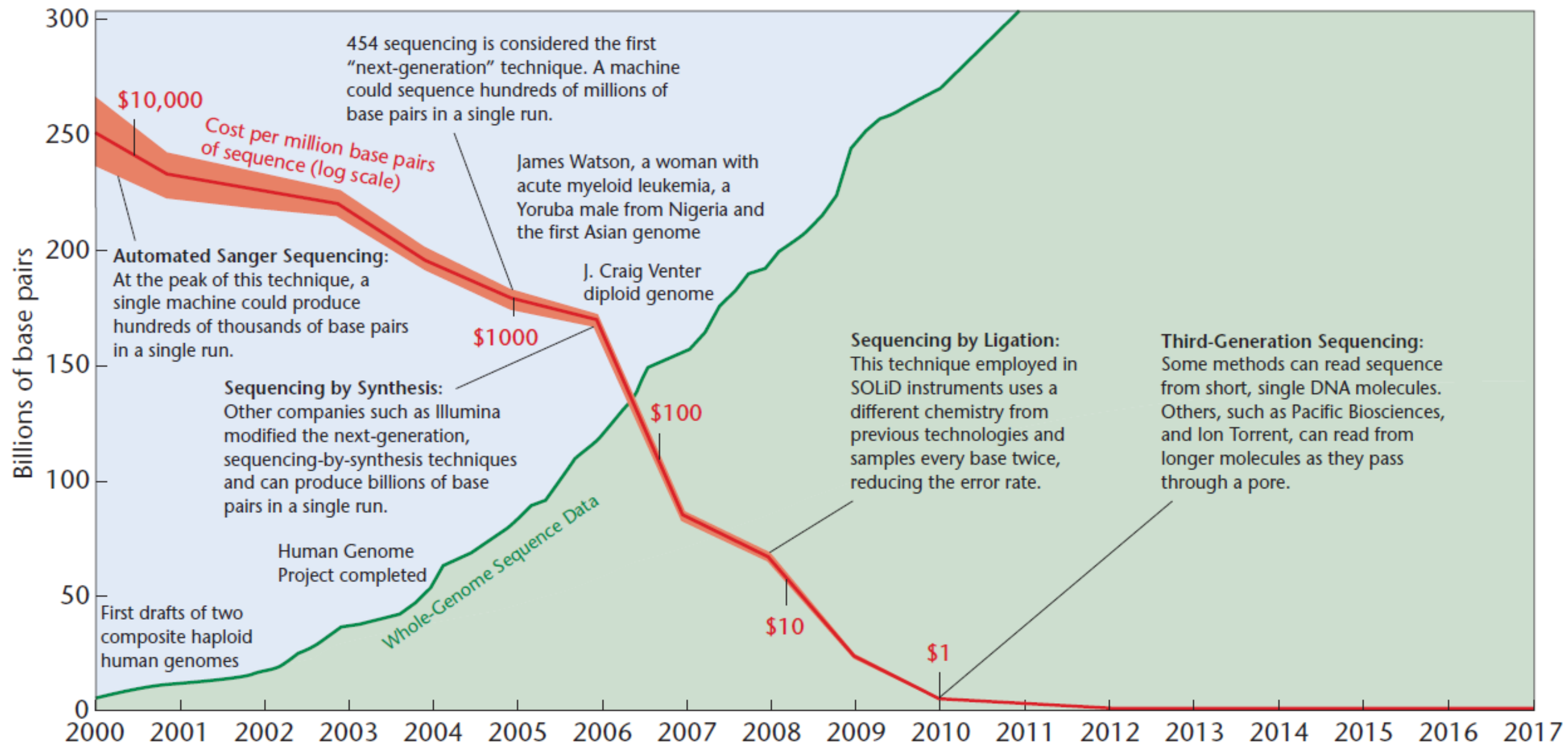
Clarisse Pais



Re.Data

Rede para a Gestão de
Dados de Investigação

A Queda dos Custos e a Explosão da Informação Genômica



Principais desafios na genômica



Our group projects



MITE-PNASA



MEDIBEES
-PRIMA project



Better-B
Horizon project



Bee3Pomics
FCT



Ana Rita Lopes



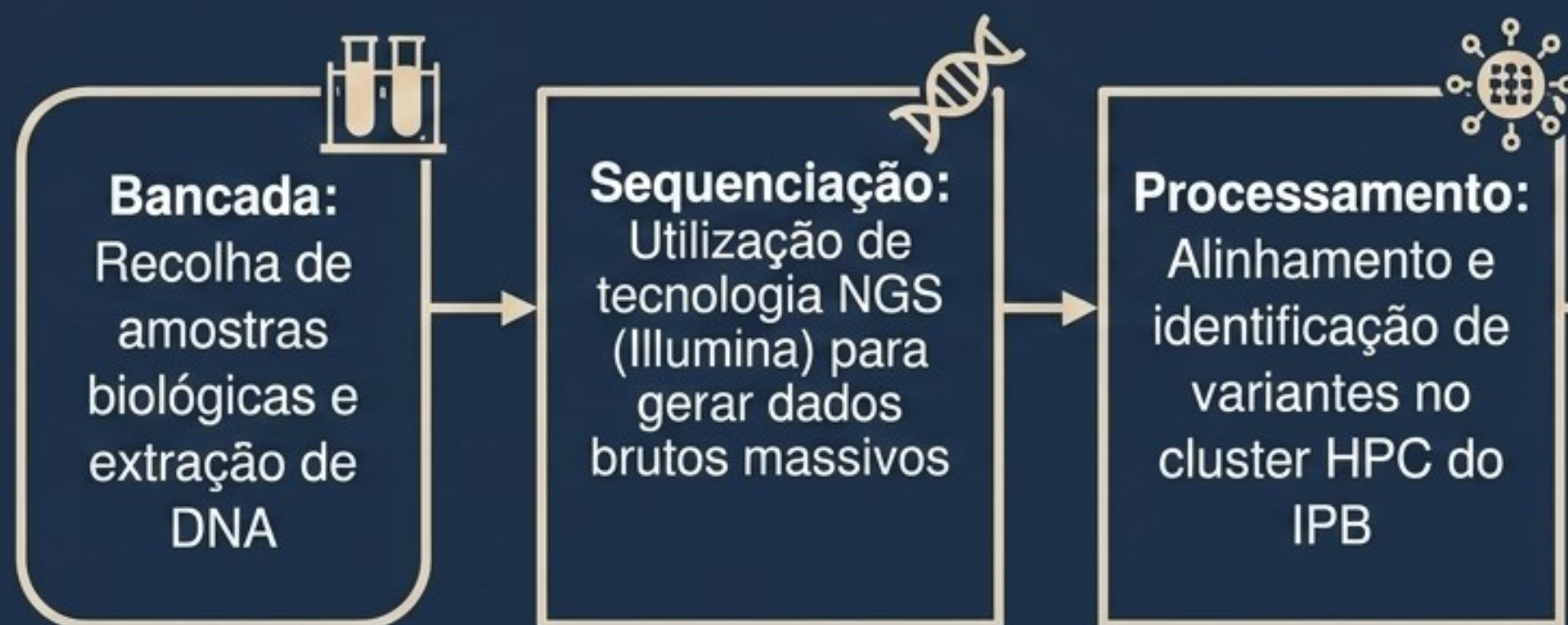
M. Alice Pinto



Dora Henriques

A Linha de Montagem: Da Bancada à Ciência Aberta

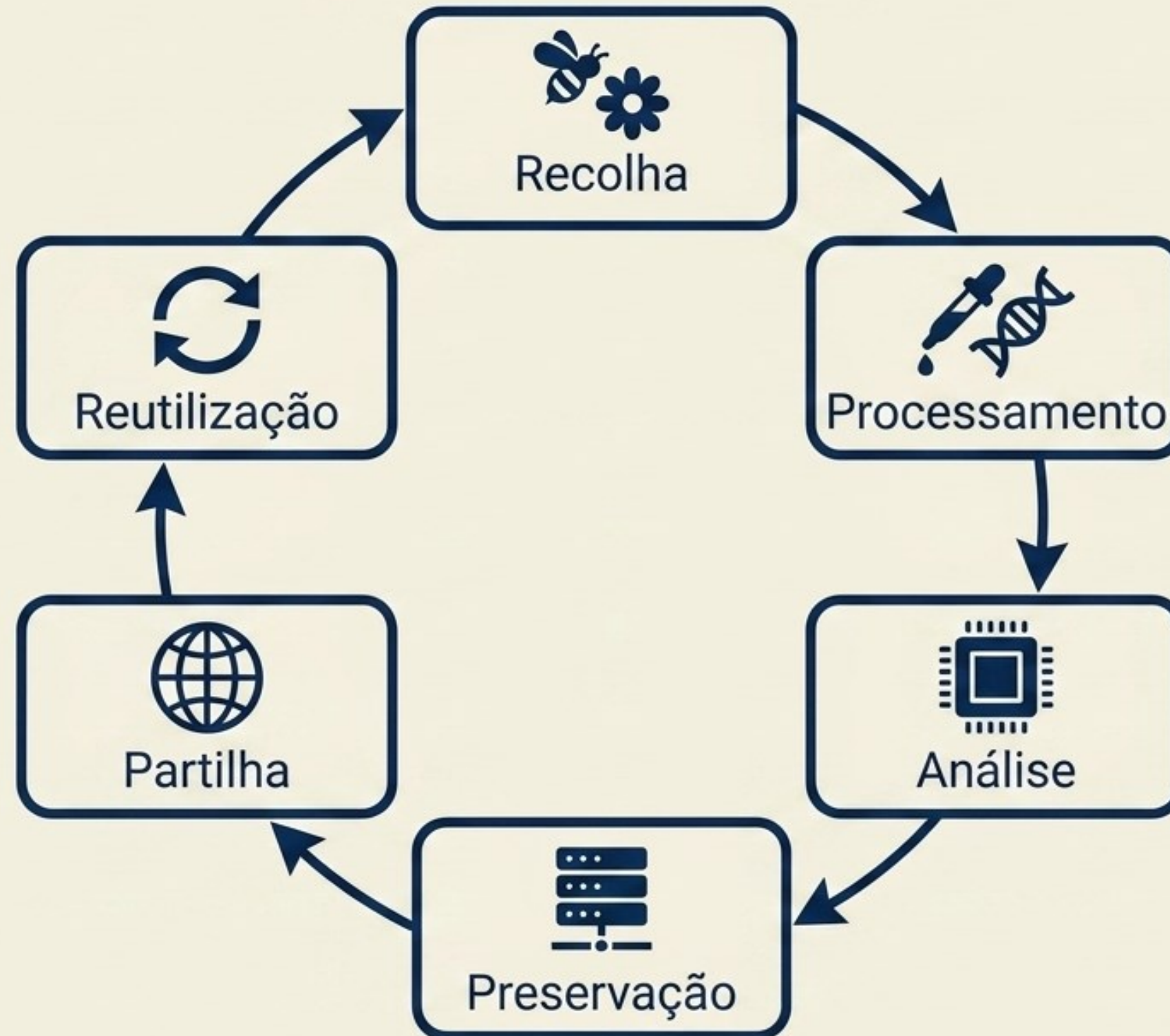
Fluxo de Dados e Processamento

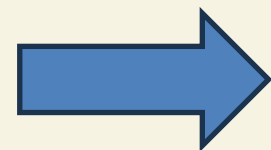


Gestão e Exportação

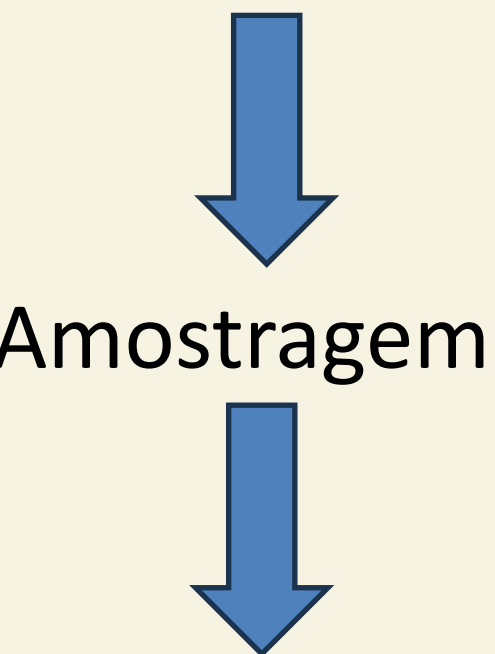


O Ciclo de Vida de Dados: Um Fluxo Contínuo





Repositórios Disciplinares



Amostragem



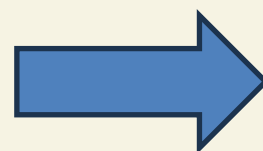
Maíra Costa



Sistema de Caderno Eletrónico
de Laboratório (ELN)
adotado pelo **Instituto
Politécnico de Bragança (IPB)**



Extração de DNA



Sequenciação

**Registrar
Experiências**

**Garantir
Integridade**

Colaboração



Repositórios Disciplinares



SRA SRA Apis mellifera Search Help

Access Public (57,583)

Source DNA (46,070) RNA (11,254)

Type exome (7) genome (13,435)

Library Layout paired (44,228) single (13,383)

Platform ABI SOLiD (179) BGISEQ (704) Capillary (857) Illumina (52,835) Ion Torrent (376) LS454 (229) Oxford Nanopore (1,297) PacBio SMRT (819)

Strategy EpiGenomics (420) Exome (964) Genome (14,273) RNASeq (10) other (41,944)

Data in Cloud GS (57,579) S3 (57,579)

File Type

Summary 20 per page

Search results

Items: 1 to 20 of 57611

1. [swarming_queen3](#)
1 BGISEQ (MGISEQ-2000RS) run: 26.2M spots, 733.8M bases, 871.8Mb downloads
Accession: SRX33041238

2. [swarming_queen2](#)
1 BGISEQ (MGISEQ-2000RS) run: 25.4M spots, 722.2M bases, 851.7Mb downloads
Accession: SRX33041237

3. [swarming_queen1](#)
1 BGISEQ (MGISEQ-2000RS) run: 26.3M spots, 751.5M bases, 867.8Mb downloads
Accession: SRX33041236

4. [non-swarming_queen3](#)
1 BGISEQ (MGISEQ-2000RS) run: 25.4M spots, 714.3M bases, 848.1Mb downloads
Accession: SRX33041235

5. [non-swarming_queen2](#)
1 BGISEQ (MGISEQ-2000RS) run: 26.3M spots, 696M bases, 850.5Mb downloads
Accession: SRX33041234

6. [non-swarming_queen1](#)
1 BGISEQ (MGISEQ-2000RS) run: 25.1M spots, 714.9M bases, 834Mb downloads
Accession: SRX33041233

7. [Apis mellifera LP region Nanopore PCR amplicon sequencing, bee from apiary E008](#)
1 OXFORD NANOPORE (MinION) run: 1.483 spots, 1.1M bases, 1Mb downloads

Results by taxon

Top Organisms [\[Tree\]](#)

Apis mellifera (24289)
insect gut metagenome (10201)
metagenome (5254)
bee metagenome (4060)
gut metagenome (2561)
All other taxa (11246)
[More...](#)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	50,400		50,400
BioProject	792		792
dbGaP			
GEO Datasets	2,825		2,825

Find related data

Database: [Select](#)

[Find items](#)

Search details

"Apis mellifera"[Organism] OR Apis mellifera[All Fields]

SRX30107425: WGS data for sample X_5492_

1 ILLUMINA (Illumina NovaSeq 6000) run: 23.6M spots, 7.1G bases, 2.2Gb downloads

Design: DNA extracted from honey bee drone thorax using Maxwell RSC PureFood GMO and Authentication kit

Submitted by: Centro de Investigacao de Montanha

Study: A. m. sahariensis (Morocco) population structure
[PRJNA1305528](#) • [SRP608911](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample:
[SAMN50609066](#) • [SRS26170704](#) • [All experiments](#) • [All runs](#)
Organism: [Apis mellifera sahariensis](#)

Library:
Name: X_5492_
Instrument: Illumina NovaSeq 6000
Strategy: WGS
Source: GENOMIC
Selection: RANDOM PCR
Layout: PAIRED

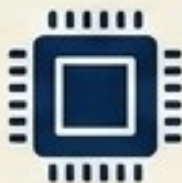
Runs: 1 run, 23.6M spots, 7.1G bases, [2.2Gb](#)

Run	# of Spots	# of Bases	Size	Published
SRR34988712	23,558,030	7.1G	2.2Gb	2025-08-14

ID: 40224572



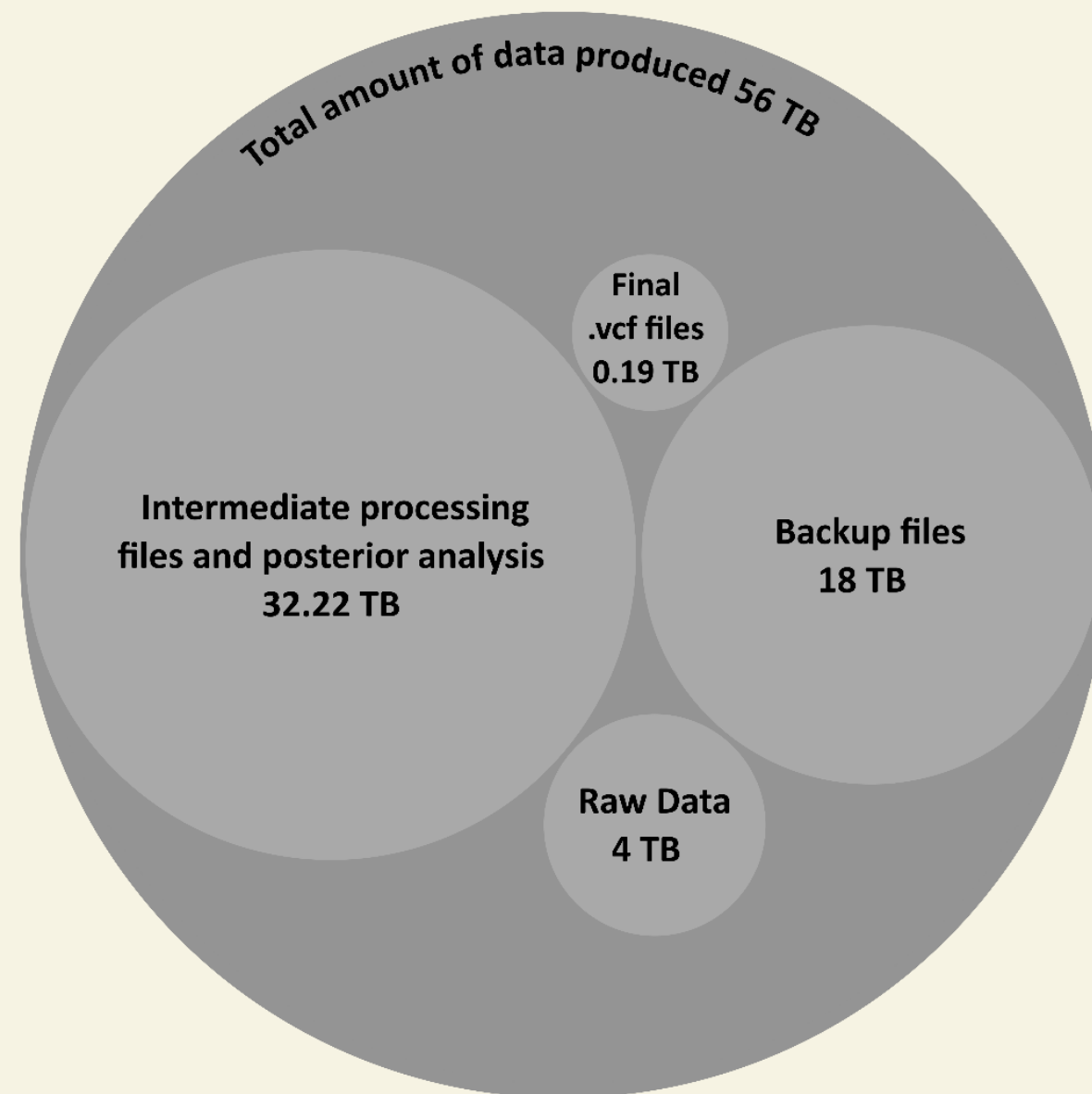
Processamento



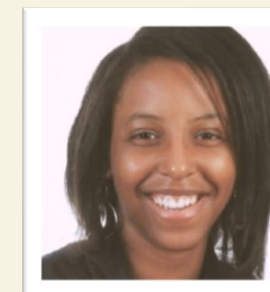
Análise



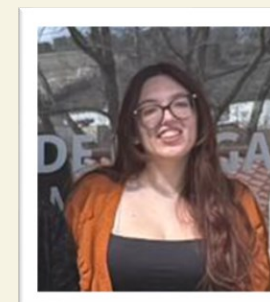
Preservação



Fernanda Li



Daniela Barbosa



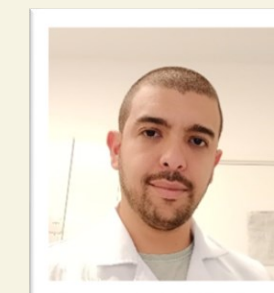
Felipe de Souza



José Rufino



Carlos Yadró



Dora Henriques



<https://cedri.ipb.pt/about/cluster>

José Rufino Amaro

Arquitetura de Colaboração e Segurança no Cluster CeDRI



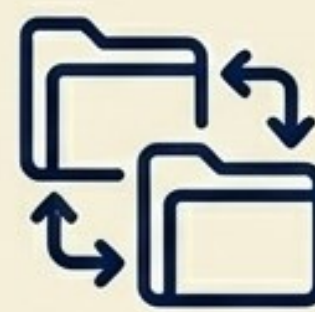
"Cofre" de Dados Brutos (Raw Data)

Ficheiros armazenados em duplicado. Acesso estrito de Leitura (Read-Only) para todos os membros. Nenhuma permissão de modificação para garantir integridade imutável.



Contas de Processamento

Áreas de trabalho individuais isoladas. Onde a equipa executa a análise computacional e consome os recursos do HPC sem colocar os dados originais em risco.



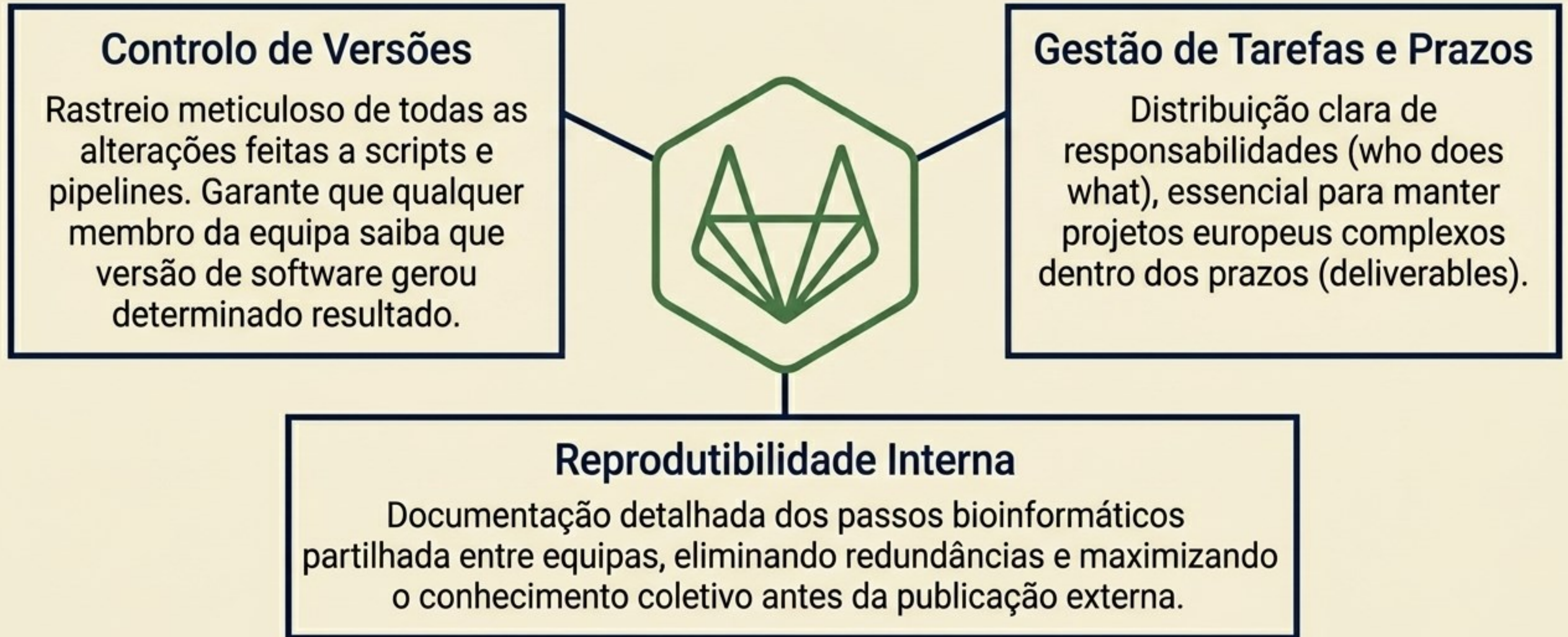
Diretórios Partilhados

O destino dos resultados finais (VCFs). Dados organizados de forma padronizada, com backups ativos, prontos para a redação científica.

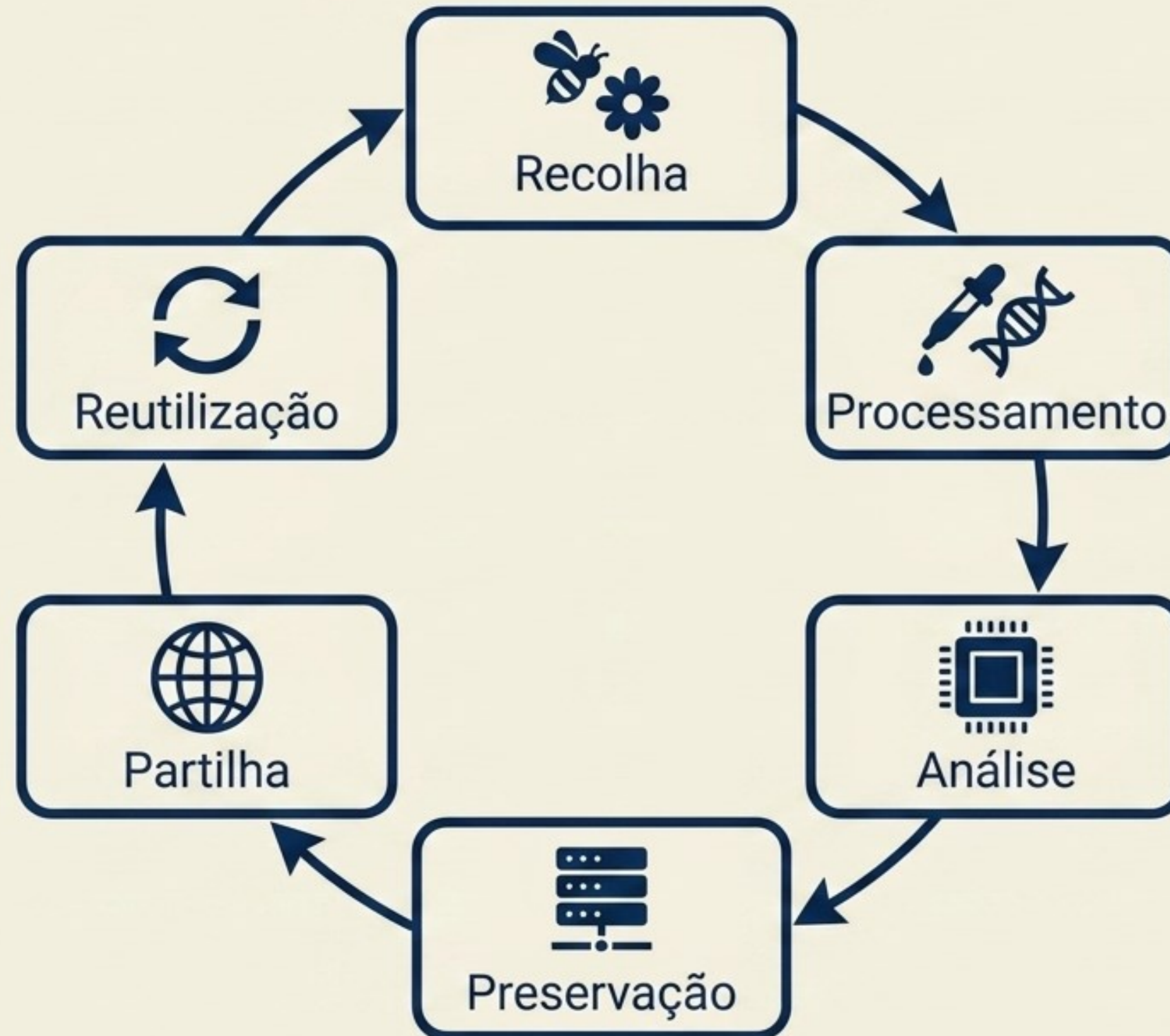
Matriz Tipológica: Como Diferentes Dados Exigem Diferentes Destinos

Tipologia	Estado e Volume	Acesso Interno	Destino Externo
Dados Brutos (Sequências Raw)	Imutável, Massivo (TB)	Read-Only (Protegido)	NCBI Sequence Read Archive (SRA)
Dados Processados (Variantes VCFs)	Refinado, Colaborativo (GB)	Pastas Partilhadas	
Código e Scripts (Pipelines Analíticos)	Dinâmico, Evolutivo (KB/MB)	GitLab Interno	GitHub / Zenodo

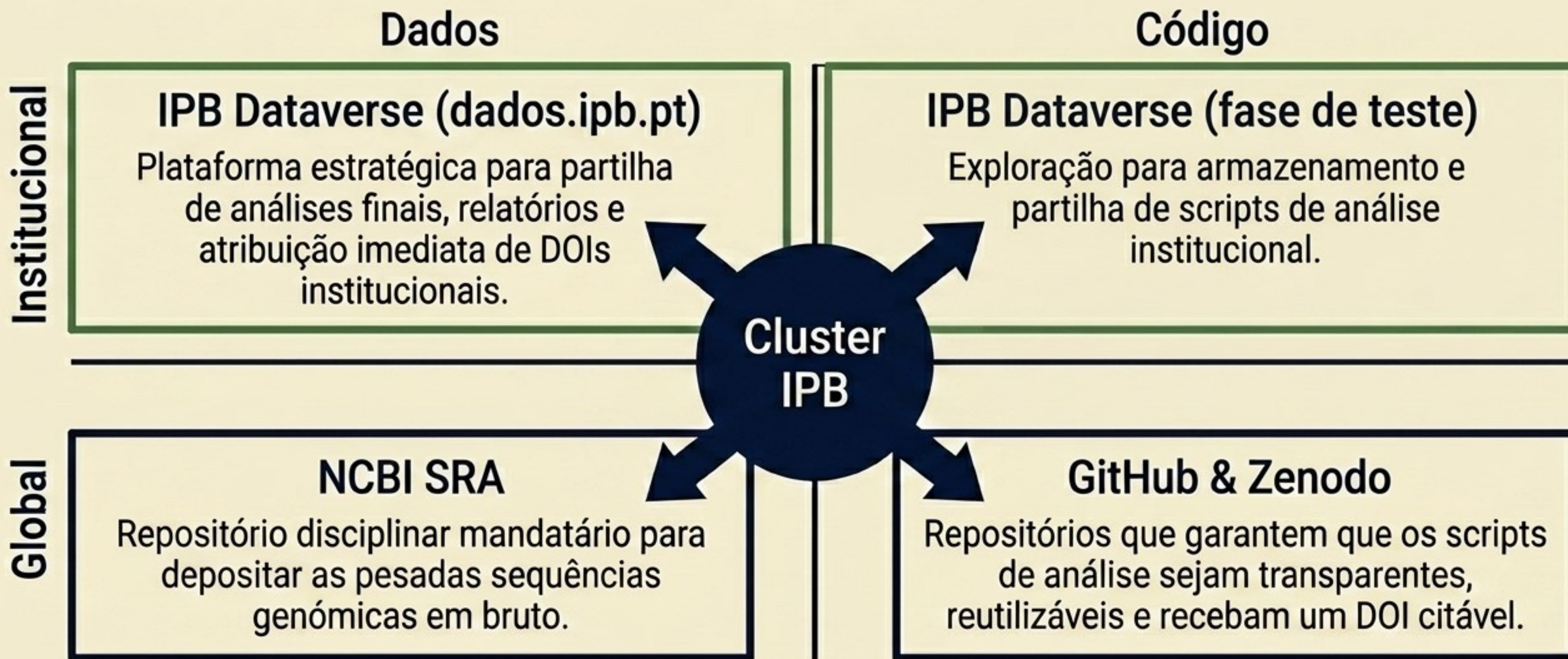
Gestão Interna e Transparência de Processos (GitLab)



O Ciclo de Vida de Dados: Um Fluxo Contínuo



O Mapa de Transição: Rumo aos Repositórios Globais



A obrigatoriedade de publicação de dados

Journal policies

[Data and image processing](#)

[Image screening](#)

[Studies involving humans and animals](#)

[Reporting sex- and gender-based analyses \(SGBA\)](#)

[Clinical trials](#)

[Human embryos and embryo models, gametes, and stem cells](#)

[Chemical compounds](#)

[Structures of biological macromolecules](#)

Resource availability

[Sensitive human data](#)

[Prepublication publicity](#)

[Process for post-publication issues](#)

[Name change policy](#)

[Permissions](#)

- All unique/stable reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

You must be willing to share all data and original code reported in your paper unless there is a legal or ethical prohibition (e.g., the data are confidential medical records). In this section, you must describe the steps you have taken to make your data and original code available. We strongly recommend that you archive all unprocessed data and original code that your paper reports in an online repository that meets the criteria for digital longevity, implementation of FAIR standards, and community support as articulated by [Fairsharing.org](#) and [collaborators](#) [↗]. If issues arise after publication, failing to produce your data might lead to a retraction.

Cell Press has specific requirements for data types that have community-endorsed, data-type-specific repositories. We call these data types “standardized data types.” Examples of standardized data types include nucleic acid sequencing, proteomics, and crystallography. Please refer to our list of [standardized data types and recommended repositories](#) [↘].

[This author's guide](#) [↘] describes requirements and recommendations for standardized datasets and original code at each stage, from submission through publication. Please note that it describes general practices; editors will provide guidance for specific cases. Exceptions to these requirements can be made for individual datasets when sharing data is legally or ethically prohibited. If you need an exception, you must contact your editor and discuss this as early as possible, ideally at the submission stage.

Requirements for publication

All datasets that are composed of standardized data types and reported in Cell Press papers must be deposited in a data-type-specific, Cell Press-recommended repository before a paper is accepted for publication. The datasets' accession number(s) or unique, permanent identifier(s) must be reported in the published paper. Any embargo must be lifted by the paper's publication date, and the deposited datasets must be publicly accessible unless there is a countervailing legal or ethical prohibition and you have received permission from your editor for an exception. Note that if standardized datasets cannot be made accessible, editors might require that summary statistics (e.g., [genomic summary results](#) [↗]) or non-identifiable, post-processed versions of the



Recommended general-purpose repositories

Cell Press encourages the practice of archiving all datasets and original code digitally. Original code and datasets of all types can be archived in general-purpose repositories. Data listed under “[Standardized data types](#)” below must be in a data-type-specific repository but could also be in a general-purpose repository. The general-purpose repositories listed below meet the criteria for digital longevity, implementation of FAIR standards, and community support [articulated by FAIRsharing.org and collaborators](#). They are recommended by Cell Press. We also encourage authors to consider institutional repositories.

- [Dataverse](#)
- [Dryad](#)
- [Figshare](#)
- [Mendeley Data](#)
- [Open Science Framework](#)
- [Science Data Bank](#)
- [Zenodo](#)

Nucleotide, sequencing, and associated datasets

Note: Local laws may prohibit the sharing of raw human sequencing data, identifiable human genetic data, or patient data. Cell Press will make reasonable compromises to general policy in these cases. Please email your editor for more details.

- Raw data generated by next-generation sequencing platforms, with per-base quality scores (e.g., BAM, SFF, and HDF5 files or paired FASTQ/QUAL files)
- Processed next-generation sequencing data, in accordance with the MINSEQE standards, as enforced by recommended, data-type-specific repositories (e.g., GEO)
- Raw and processed microarray data, in accordance with the MIAME standards, as enforced by recommended, data-type-specific repositories (e.g., GEO)
- Partial and complete genomes (e.g., FASTA files) with associated metadata as enforced by recommended, data-type-specific repositories (e.g., GenBank)
- Human genotype-phenotype data (e.g., Matrix, PLINK, and VCF files) that are eligible for deposition at a recommended, data-type-specific repository (e.g., dbGAP)
- Data that describe genetic variation and are eligible for deposition at a recommended, data-type-specific repository (e.g., dbGAP for human data or EVA for non-human data)

Data availability or accessibility

The raw DNA sequence data generated in this study can be found in NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under the SRA study and BioProject accessions PRJNA1305528, PRJNA1305509 and PRJNA1307299 respectively. Codes for Genotype-Environment association analyses are available on GitHub (https://github.com/cyadrogarcia/GEA_ligustica_iberiensis) with an archived snapshot on Zenodo (<https://doi.org/10.5281/zenodo.16726609>).

Desafios Futuros na Gestão da Ciência Aberta

Integração Total da Equipa

O desafio cultural de garantir que 100% dos colaboradores utilizam o ecossistema institucional de forma consistente e integrada.



Monitorização Global de Projetos

A urgência de implementar ferramentas universais que conectem prazos, deliverables e DOIs num único painel de controlo.

Flexibilidade vs. Robustez nos DMPs

A dificuldade atual de criar Planos de Gestão adaptáveis à natureza muito dinâmica da genómica, devido à falta de apoio sistémico global.

OpenDMP
PORTALBERTA
Planos de Gestão e Partilha de Dados

ipb
INSTITUTO POLITÉCNICO
DE BRAGANÇA

Universidade de Minho

Acompanhe os seus dados

Crie planos de gestão e partilha de dados.

Configure modelos específicos da sua área científica.

Associe o plano a serviços do seu financiador.

Partilhe com colegas ou orientadores.

Síntese: A Engenharia dos Princípios FAIR na Prática

F

Findable -
Encontrável

Atribuição de DOIs
persistentes aos
datasets através do
IPB Dataverse.

A

Accessible -
Acessível

Disponibilização
pública e gratuita
dos dados brutos
massivos através do
NCBI SRA.

I

Interoperable -
Interoperável

Adotar Formatos de
Arquivo Standard
Implementação de
Metadados

R

Reusable -
Reutilizável

Transparência total
com pipelines e
scripts fortemente
documentados
exportados para o
GitHub/Zenodo.

Três Lições Aprendidas na Trincheira da Big Data

1

O Planeamento é a Base de Tudo

Planos de Gestão de Dados (DMPs) não são apenas burocracia para financiadores. São a arquitetura que previne caos em projetos de 56 TB.

2

Repositórios Institucionais são Motores Estratégicos

O uso do dados.ipb.pt agiliza drasticamente a redação de artigos científicos e a submissão de relatórios exigidos pelas agências europeias.

3

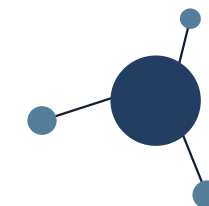
A Filosofia FAIR Eleva o Impacto

Aplicar os princípios FAIR deixa de ser uma obrigação e passa a ser uma ferramenta que exponencia a transparência, a confiança e a visibilidade académica do investigador.



Obrigada

Dora Henriques
dorasmh@ipb.pt



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

Please keep this slide for attribution



Síntese: A Engenharia dos Princípios FAIR na Prática

