Data Storage, Sharing and Organization in the EPIGEN Project

Authors: Maíra Ribeiro Rodrigues
Infrastructure, Informatics, Integration and Enrichment (I3E) Team
maira.r.rodrigues@gmail.com

Created at: 19/08/2013 Updated at: 30/08/13

Contributions to the text:
Eduardo Tarazona and Wagner C. S. Magalhães

1. Introduction

The purpose of this report is to describe the strategies used by the EPIGEN to store, share and organize the data and the results produced by project teams. These issues are important due to the large size of the data handled by project teams and to the distributed location of collaborating teams.

2. Data Storage and Sharing

All the data, tools and documents produced by the EPIGEN project are stored in two places: a Linux server, located at the LDGH/UFMG centre, and a Web portal, hosted by the LCC/UFMG. The project also maintains a storage device for data backup. The LDGH/UFMG centre manages both the server and the web portal.
The server stores large genotyping and sequencing data, as well as all data generated by the analyses processes. The Web portal stores documents, flowcharts and images, for better visualization, and provides an easily accessible graphical interface.

2.1 Data and Analyses Server

The server stores project data in a centralized manner, as illustrated in Figure 1. The goal of the centralised storage is to guarantee data coherence, since all participating research centres access and use the same data and data updates are visible to all centres.

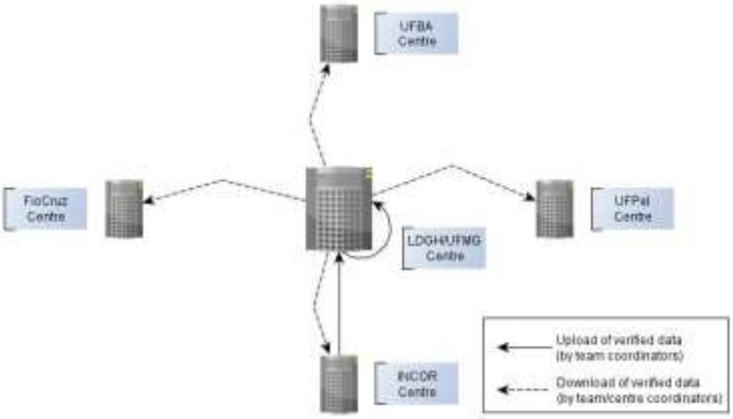

UFBA Centre
FioCruz Centre
UFPal Centre
LDGH/UFMG Centre
INCOR Centre
Upload of verified data (by team coordinators)
Download of verified data (by team/centre coordinators)

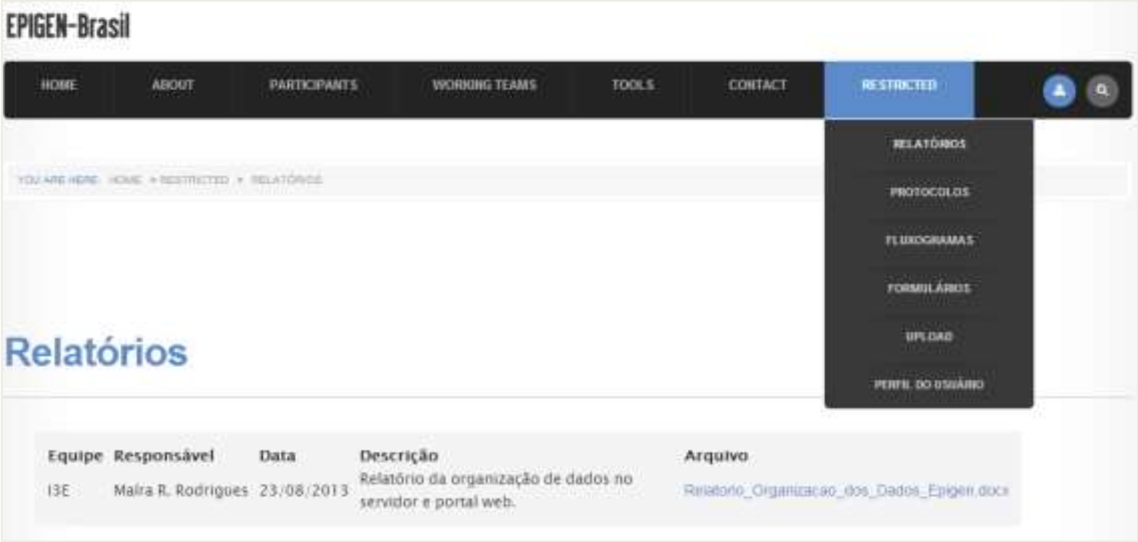
Figure 1: Data storage policy.

Regarding data security, the server is accessed through individual password protected accounts. It is also required that all accounts be authorized by means of a consent form signed by the account user and the head of his research centre (see Attachment I).

The server has a Unix-based operating system and a command line interface, which are ideal for large-scale analyses and for handling large files.

2.2 Web Portal

The web portal (www.epigen.grude.ufmg.br) hosted by the LCC/UFMG stores documents, images and flowcharts to allow a user-friendly visualization of results, protocols and reports. It also provides an easily accessible graphical interface, which is an alternative for users not familiar with Unix-based systems. Access to project documents is available through a restricted, password protected website area. Only the teams' and centres' coordinators are allowed to upload documents to guarantee the quality of available documents.



Relatórios

Equipe	Responsável	Data	Descrição	Arquivo
I3E	Maira R. Rodrigues	23/08/2013	Relatório da organização de dados no servidor e portal web.	Relatorio_Organizacao_dos_Dados_Epigen.docx

Figure 2: Web Portal restricted area - snapshot of the "Reports" page.

The upload system accepts files with maximum size of 20MB and with common document and image extensions (doc, docx, pdf, txt, jpg, png, bmp, gif).

The upload and visualization systems allow all geographically distributed research teams to make their research results available and to have access to the documents produced by each other and, therefore, to collaborate more efficiently.

3. Data Organization

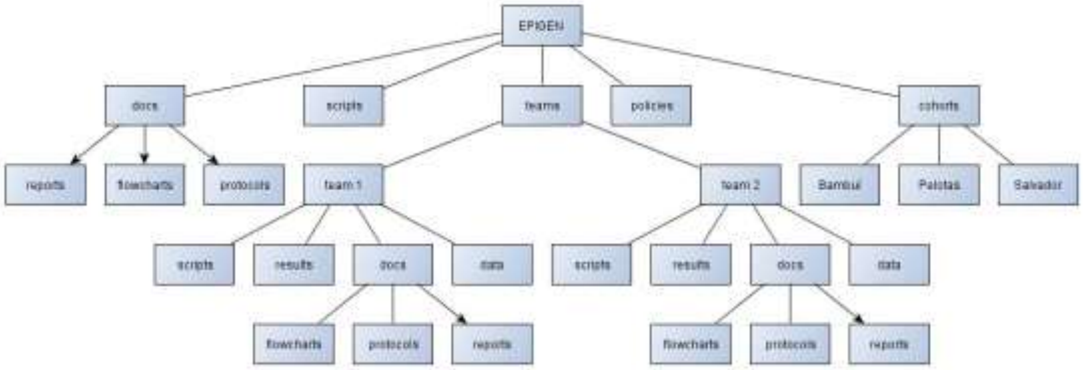


Figure 3: Server directory organization.

To organize the data produced by the different analyses teams we used the directory organization suggested by (Noble 2009). It is illustrated in Figure 3. The folders “scripts”, “policies” and “docs”, at the top hierarchy, are accessible by all members of the EPIGEN project with a server account. The “scripts” folder stores final and tested versions of scripts produced by analyses teams. We established a default code header for all scripts developed in the EPIGEN Project (see Attachment II). Final documents produced by analyses teams are stored in the “docs” folder, and final documents produced by the project board committee are stored in the “policies” folders; files in both folders are also uploaded to the Web portal for easy visualization. The “teams” folder contains one subfolder per analysis team (the EPIGEN project has currently 5 analyses teams), which subfolder is accessible only to member of the respective analysis team. The genetic analyses teams are: Basic analysis; Ancestry; Haplotype Inference and Imputation (I2H); Infrastructure, Informatics, Integration and Enrichment (I3E), and Analysis Pipeline.

All the intermediate data, documents, results and scripts produced by each analysis team are stored in the team’s respective subfolders “data”, “docs”, “scripts” and “results”. When final versions of documents and scripts are produced by each analysis team, they are transferred to the top hierarchy “docs” and “scripts” folders. For files in the “docs” folder, a copy is also uploaded to the Web portal. This guarantees that only revised and verified documents and scripts are made available to all project members. Similarly, when final data files and results stored in the “data” and “results” folders are ready for association studies (that is, data files after being processed by the analyses teams for data cleaning, ancestry analysis, imputation, and others), they are transferred to the “cohorts” folder at the top hierarchy. Each cohort has a private subfolder that is accessed only by members of the respective cohort.

With such directory organization we facilitate the management and access of the large number of (and large size of) files produced by each research team.

A simpler organization structure is applied to the Web portal, since it only stores documents and images. We created four document categories that are: reports, protocols, flowcharts and forms. In each category, documents are listed in a table format, as shown in Figure 2, with the following information: research team in which it was produced, document author (or first author in case there are many), date of creation, brief content description, and file name. This organization complements the server directory organization by providing an alternative, friendlier interface for the visualization of documents and images. Since it is web-based, project members have access to such documents from any computer with an internet connection.

4. Analyses Organization

To organize the data analysis processes we use two strategies: one for data analysis representation, the flowcharts; and one for analysis execution, the masterscripts. Team members or other teams’ coordinators revise both flowcharts and masterscripts before their final version is available. They are described next.

4.1 Flowcharts

We adopted the flowchart representation to describe all the data analyses done by the project’s research teams. A sample flowchart for a simple phase inference analysis is illustrated in Figure 4. This analysis has two steps: first, PED and MAP files with unphased data are converted to a BED file using the process “pedmap2bed”; and second, this BED file is used as input for a “phase inference” program that generates a phased data file. Although in this example we use generic names for files and programs, flowcharts can contain the specific names of programs used in the corresponding analysis.

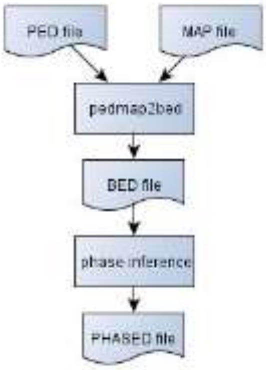


Figure 4: Simple flowchart for a phase inference analysis.

The main requirement of the flowcharts is that each process is represented with clear inputs and outputs. In Figure 4, for example, the "pedmap2bed" process has two inputs, a PED file and a MAP file, and one output, a BED file). The advantage of using flowcharts is that they provide an easy and intuitive visualization of the data analysis as a whole. Therefore, it facilitates the understating of the analysis process by members of different research teams and the reproducibility of the analysis. We are currently using the software YED for flowchart representation. It allows representing flowcharts with different levels of detail, for specification of concepts and commands.

4.2 Masterscripts

A masterscript is a simple set of commands, independent of programming language (it can be written in shell, sed&awk, perl, etc). It must comprise all commands of a given analysis, including check points for data quality control, and the specification of the parameters used for a specific analysis. All commands must have comments briefly describing their purpose. In our data analysis organization, the masterscript contains all the commands necessary to execute the processes in a flowchart.

Table 1: Sample masterscript.

```
#PED/MAP to BED conversion
pedmap2bed file.ped file.map > file.bed
echo "ped/map to bed conversion finished"

#Phase inference program
phaseinference file.bed > file.phased
echo "phase inference program finished"
echo "phase inference analysis finished"
```

A sample masterscript for executing the analysis represented by the flowchart in Figure 4 is shown in Table 1. It has comments for each step of the analysis (lines starting with "#"), and check points that indicate when each step finishes (lines starting with "echo").

Therefore, to repeat a given analysis it is only necessary to run its masterscript again. Also, with the masterscript it is possible to recover all the parameters and inputs used to run a software in a given analysis step. Together, the flowchart and the masterscript representations facilitate and warrant the reproducibility of the analysis results.

Supporting Infrastructure

Although we use only our local server for storing the project's large-size datasets, the EPIGEN project counts with other supporting computational infrastructure for running several genetic analyses. These are: a cluster computer from LCC-CENAPAD (UFMG), a cluster computer from CEBio, and a cluster computer from CESUP-CENAPAD (UFRGS).

References

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424.

ATTACHMENT I

Confidentiality Term that must be signed by all member of the EPIGEN Project with access to the data on the server and web portal.

Timbre do laboratório/instituição

TERMO DE COMPROMISSO E SIGILO

DECLARO, para todos os fins, o compromisso de utilizar dados genômicos das coortes de Bambuí, Pelotas e Salvador somente com a autorização de um dos membros do Conselho Diretor do Epigen-Brasil.

Estou ciente de que, após 12 meses da presente data, devo destruir o banco da minha pesquisa (exceto quando a renovação foi autorizada, o que implicará em assinatura de novo termo), que não posso utilizá-lo para outro fim, nem repassa-lo a qualquer outra pessoa. Declaro, ainda, que a publicação resultante das minhas análises seguirá os critérios adotados pelo Conselho Diretor do Epigen-Brasil.

Local, Data

Assinatura do pesquisador

Assinatura do membro do Conselho Diretor responsável pela autorização

ATTACHMENT II

Code header default for the EPIGEN Project:

```
#==========================================
# EPIGEN PROJECT
#==========================================
#
# (C) Copyright 2013, by <GROUP NAME> and Contributors.
#
#
# -----------------
# PROGRAM NAME
# -----------------
#
# Original Author:
# Contributor(s):
# Updated by (and date):
#
# Command line:
#
# Parameter description:
#
# Description:
#
# Dependencies: Libraries, compilers, packages, etc.
#
# Output:
#
# Sample input files:
#
```