“Best practices in research data management and stewardship” workshop - Example scenario

You are a postdoctoral researcher, who has recently joined University of Luxembourg’s Life Sciences Research unit (LSRU). You will be working with Prof John Black on understanding the gene regulatory interactions underlying cell differentiation and identity. Specifically, you will work on temporal data on gene expression and context-specific open chromatin states for improved identification of key transcription factors and the gene regulatory networks.

Your project is part of an ongoing collaboration, a research programme, between the LSRU, the Centre Hospitalier de Luxembourg (CHL), the Luxembourg Centre for Systems Biomedicine (LCSB) and the Integrated Biobank of Luxembourg (IBBL). These institutes have already obtained an ethics approval for the programme and have signed an agreement that, among other provisions, allows for exchange of bio-samples and data with each other.

For your research project the planned dataflow is illustrated in Figure 1 and described below:

- A healthy volunteer will be recruited by Dr Igor Green, your collaborator at the CHL. The volunteer will fill out a consent form at the hospital, to confirm whether he/she agrees to the use of donated bio-samples and any derived data for cancer research. The donor will also confirm whether or not the data can be transferred internationally, including countries outside the EU. After consent, the volunteer will donate blood, from which CHL will isolate bone marrow-derived stromal cell lines. CHL will submit a batch of these cell cultures to the IBBL for long term storage, and another batch to the LCSB, for immediate analysis for your project.

- The cell lines will be received by the Sequencing Platform at the LCSB, led by Prof Alice White. The LCSB will generate whole-genome sequencing and ChiP-seq time course data from samples. They will share with you the data on their WebDAV server.

- You will download sequencing data to your laptop and also keep a copy on the university’s HPC platform for analysis. You will perform analysis on the data using scripts/pipelines developed by Postdoctoral researcher Dr Roy Blue, who is also a member of your lab. In addition to LCSB generated data your analysis will use human reference genome data for the identification of transcription factors.

- Upon completion of analysis, you will publish your findings in a peer-reviewed journal, following the journal’s requirements of data and code availability.”

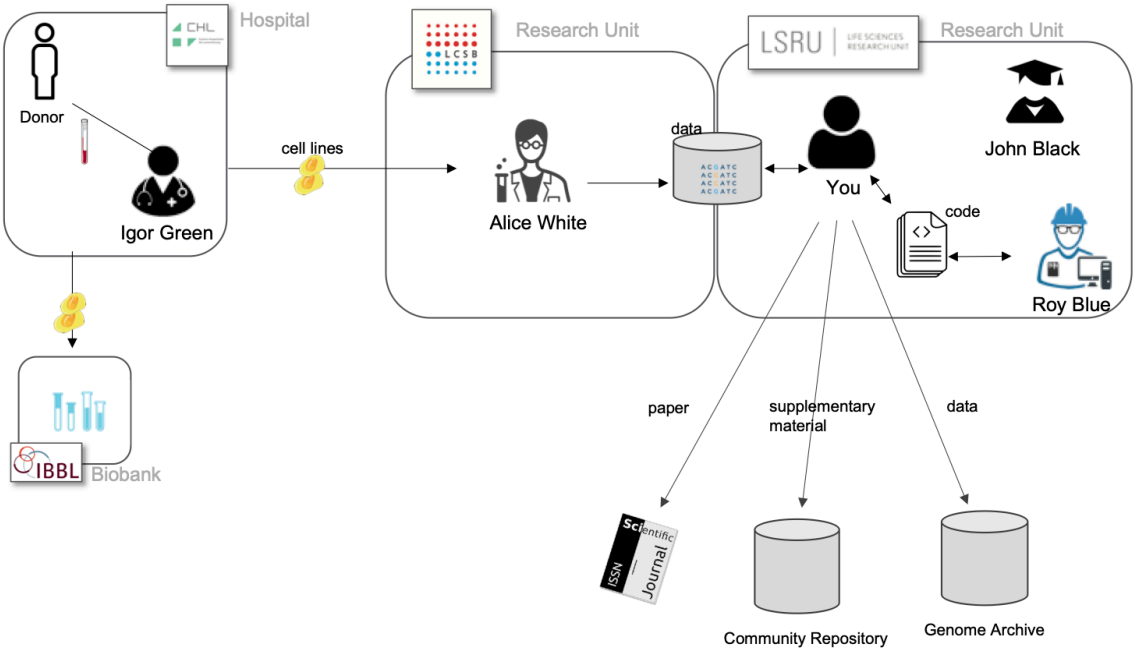

Figure 1Graphical depiction of example scenario

Our workshop contains four practical sessions, which will follow this scenario:

- **Data Management Planning:** Your postdoc project is funded by the FNR. The funder requires the submission of a Data Management Plan (DMP) no later than 4 Months after the start of your project. In our first practical, you will use an online tool called the Data Stewardship Wizard to generate a DMP document for your research project.

- **Data Inventory, GDPR Accountability:** You are using human-subject data in your work, therefore, to meet data protection requirements for such data under the GDPR, you will need to document your data processing. In our second practical, you will use an online tool called Data Information System (DAISY) to register details of your project and the data involved.

- **Reproducible Data Analysis:** Your PI John Black has recently adopted the policy of full computational reproducibility for any analysis devised in his lab. As a result, you will need to encode your analysis as an end-to-end workflow, which can be re-configured and re-run at any time. In our third practical you will use an analytical pipeline built with the Snakemake workflow system for your analysis of ChIP-Seq data.

- **Transparent Research Publishing:** For many journals, data and code availability is a prerequisite for publication. Therefore, you will be expected to publish these in support of your paper.

    - As the data you use is sensitive sequencing data from a human subject and it carries conditions to its use, you would need to deposit this data to a dedicated genome data repository with a controlled access process. An example of such

a repository is the European Genome-phenome Archive. Controlled access would allow you and your collaborators to act as controllers foreseeing the data's careful dissemination in a manner that honours data use conditions. In addition, publishing to dedicated data repositories is a lengthy process, which requires detailed documentation of data, with subject and sample details, as well as a description of the data capture platforms and important experimental and analytical parameters. We will discuss this process in our lecture on data archival, but we will not cover it in our practicals.

- The journal you are submitting requires supplementary material to be made available openly. In our fourth practical you will create an archive of your analysis data and code, excluding the sensitive human data. You will then publish this archive on FAIRDOMHub, and provide essential bibliographic information such as contributors, funders etc, you will also link your supplementary material with your paper.