1. Data Summary

The purpose of data collection is to analyse the role of the X transcription factor in the control of gene expression in embryonic stem cells and during their differentiation to primordial germ cells.

Types of data that will be generated are: CUT&RUN (epi)genomic profiling data / Analysis of protein expression in stem cell and primordial germ cell cultures / 3D chromatin interaction data / Co-immunoprecipitations / Excel/CSV/html files of gene expression and genomic occupancy data.

Data will be generated from cell lines and will be stored in the appropriate format for the type of data: FASTQ and BigWig format files for sequence data files (CUT&RUN, Hi-C, Capture-C), tiff for image files and FlowJo for FACS data. We estimate generating ~500Mb data per year.

Data publicly available and deposited in the European Nucleotide Archive (ENA) will be re-analysed and compared to CUT&RUN data generated in the project. The data will be accessible to the scientific community and they will be used for further comparison and genomic/stem cells study perform in the host and other labs.

2. FAIR data

**2. 1. Making data findable, including provisions for metadata**

Generated data will be initially stored using University of Edinburgh DataStore file system. This is a high quality, enterprise-class storage with guaranteed backup and resilience. After publication, data will be submitted to the appropriate public databases (SRA/ENA or GEO) with sample data and annotations as recommended by the public repositories. CUT&RUN, Hi-C and Capture-C data will be stored in the GEO database and SRA/ENA sequence archives. Image files will be uploaded into the Mendeley data repository and FACS data will be stored in the FlowRepository database. Each set of data stored in the cited repositories will be identifiable via a Digital Object Identifier (DOI).

Will search keywords be provided that optimize possibilities for re-use?

**2.2. Making data openly accessible**

I aim to publish the data in a timely manner and so make it available from within public repositories. Prior to publication, data will be made available to a restricted set of collaborators from within UoE DataStore. explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Data will be accessible from public depository:

European Nucleotide Archive <http://www.ebi.ac.uk/ena/>

Gene Expresion Omnibus <http://www.ncbi.nlm.nih.gov/geo/>

Mendeley Data: <https://data.mendeley.com/>

FlowRepository: [http://flowrepository.org/](http://flowrepository.org/\)

The identifiers from public repositories will be added to our manuscripts, which we will submit to journals that support open-access publication.

**2.3. Making data interoperable**

Data produced within the project will be highly interoperable, since they will be generated using current standards (tiff files for images, FASTQ and BigWig for sequencing data, FlowJo files for FACS data). Epigenomic data will be generated with at least three biological replicates for CUT&RUN for each factor and cell lines analysed, and two replicates for Hi-C and Capture C.

**2.4. Increase data re-use (through clarifying licences)**

Data will be licenced under CC-BY licence term, and they will be available for re-use and sharing, after their publication. Output epigenomic data (BigWig files) will be accessible on the GEO database, while input data (raw and unanalysed – FASTQ files) will be accessible for re-use on the ENA archive. Quality of next-generation sequencing raw data will be tested with FastQC software (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Excel/CSV/html files of gene expression and genomic occupancy data will be stored on the Zenodo database and made publicly available. Data will be re-usable for an indefinite timeframe.

4. Data security

Generated data will be stored using UoE DataStore file system. This is a high quality, enterprise-class storage with guaranteed backup and resilience. Once published, data will be made available from within public repositories. Long-term data preservation for 10 years after the end of the fellowship will be achieved by the cost-effective UoE DataVault service.

5. Ethical aspects

Data are generated from murine cell lines. No human subjects or mice will be used during the project.