Data management plan

I agree and fully commit to the UKRI policies and principles of data preservation, management and sharing policy as outlined here: https://www.ukri.org/apply-for-funding/before-you-apply/your-responsibilities-if-you-get-funding/

Data areas and data types

We will produce whole genome sequencing data, ATAC-seq data and large-scale stochastic computer simulations. Whole genome sequencing will generate data in the form of FASTQ (for sequencing data). Simulation output is stored in .csv format. Following analysis, results will be in the form of BAM alignment files for NGS data. All the post-analysis data and results will be csv files. For all data analysis and stochastic simulations, R, Python/C++ programming codes will be produced and made publicly available on https://github.com/BenWernerScripts.

Standards and metadata

Sequencing and stochastic simulation analysis and storage will be conducted using the High-Performance Computing (HPC) services provided by the Research IT group at QMUL. This guarantees fast and sufficient computing power for our data analyses with secure access and storage, and also a constant data back-up system. All data will be stored with unique identifiers, experimental conditions, crucial relational link to other data files and quality matrices. Quality will be checked immediately after data collection and searched using data specific quality matrices including read length, percentage of mapped reads and coverage. We will upload all codes and derived software/database into the development version control system (https://github.com/BenWernerScripts). Raw data will be kept private until after publication. Simulation results and code will be available with the publication of corresponding preprints. All preprints will be published on the free print server https://www.biorxiv.org.

Methods for data sharing

Data will be made available at the point of publication of the associated paper or publication or latest by the end of the fellowship. We will publicize our results and data via peer-reviewed scientific journals Gold level open access. PDF reproductions or author-formatted copies of accepted publications will be available at my page: http://benjaminwerner.org/publications/ at the day of manuscript publication. I will work to ensure adherence to relevant copyright legislation. In addition, project results and data will be disseminated to the wider scientific community at national and international conferences. Contributions to conferences will be made by the PDRA. My group will use funds requested in the grant application. Conference abstracts will be published in the official meeting proceedings and via the Barts Research Database. The raw data reported in journal articles and conference presentation will available on appropriate archives, e.g. EGA for sequencing and ArrayExpress for single cell data. All scripts, programming code for stochastic simulations, data and figure generation and potentially other computational applications will be made publicly available under https://github.com/BenWernerScripts at the day of preprint publication. We commit to add all manuscripts on preprint server at the moment of journal submission. Projects between labs will share and manage data/code via Apocrita and Github access. I will mandate the use of Anaconda notebooks for code transparency across different programming languages to maximize user friendliness, transparency and communications between collaborators.

Timeframes

Setting up the wet lab expertise, beginning experiments and producing sequencing data will take 24 months. Theoretical modelling and applied analysis in collaboration with other groups will be faster and will occur throughout the entire grant duration. A publication will become available within the first 12 month of the grant. We hope to generate publications of grant specific generated data within the first 36 month and publish the major body of the work in multiple publications by the end of the first 48 month.