Screw: tools for building reproducible single-cell epigenomics workflows

Kieran O'Neill[1,2], B Decato[3], A Goncearenco[4], A Khandekar[4], B Busby[4], and A Karsan[1,2]

[1]Pathology Department, University of British Columbia, Vancouver, Canada
[2]Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada
[3]Molecular & Computational Biology Department, University of Southern California, Los Angeles, California, USA
[4]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA

DNA methylation is a heritable epigenetic mark that shows a strong correlation with transcriptional activity. The gold standard for detecting DNA methylation is whole genome bisulfite sequencing (WGBS). Recently, WGBS has been performed successfully on single cells (SC-WGBS) [1]. The resulting data represents a fundamental shift in the capacity to measure and interpret DNA methylation, especially in rare cell types and contexts where subtle cell-to-cell heterogeneity is crucial, such as in stem cells or cancer. However, SC-WGBS comes with unique technical challenges which require new analysis techniques to address. Furthermore, although some software tools have been published, and several existing studies have tended to use similar methods, no standardized pipeline for the analysis of SC-WGBS yet exists.

Simultaneously, there has been a drive within bioinformatics towards improved reproducibility. Textual descriptions of bioinformatic analyses are deeply inadequate, and often require "forensic bioinformatics" to reproduce [2]. Recreating the exact results of a study requires not only the exact code, but also the exact software versions compiled in the same way. Common Workflow Language (CWL) provides a framework for specifying complete workflows, while Docker allows for bundling of the exact software and auxiliary data used in an analysis within a container that can be executed anywhere. Together, these have the potential, via repositories such as Dockstore [3], to enable completely reproducible bioinformatics research.

Here we present Screw (Single Cell Reproducible Epigenomics Workfow). Screw is a collection of standard tools and workflows for analysing SC-WGBS data, implemented in CWL, with an accompanying Docker image. Screw provides the parts for software carpentry of fully-reproducible SC-WGBS analyses. Tools provided include quality control visualization, clustering and visualisation of cells by pairwise dissimilarity measures, construction of recapitulated-bulk methylomes from single cells of the same lineage, generation of bigWig methylation tracks for downstream visualization, and wrappers around published tools such as DeepCpG [4] and LOLA [5]. Screw has the added benefit that CWL's compatibility with interactive GUI-based workflow tools such as Galaxy can lower the barriers to use for less-technical wet lab biologist users.

CWL sources for Screw are available under the MIT license at https://github.com/Epigenomics-Screw/Screw. Tools and workflows are available from Dockstore under Epigenomics-Screw namespace, for example https://dockstore.org/workflows/Epigenomics-Screw/Screw/screw-preprocess

1. Schwartzman, Tanay (2015) Nature Reviews Genetics 16:716–26.

2. Gentleman (2005) Statistical applications in genetics and molecular biology. doi: 10.2202/1544-6115.1034

3. O'Connor et al. (2017) F1000Research 6:52.

4. Angermueller, Lee, Reik, Stegle (2016) bioRxiv 055715.

5. Sheffield, Bock (2015) Bioinformatics 32:587–589.