# Screw: tools for building reproducible single-cell epigenomics workflows

Kieran O'Neill, Chelsey Fang, Benjamin Decato, Azhar Khandekar, Alexander Goncearenco, Ben Busby, Aly Karsan

Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada

BC Cancer Agency CARE + RESEARCH
An agency of the Provincial Health Services Authority

Background: Single-cell DNA Methylation Sequencing

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) [2]. 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.
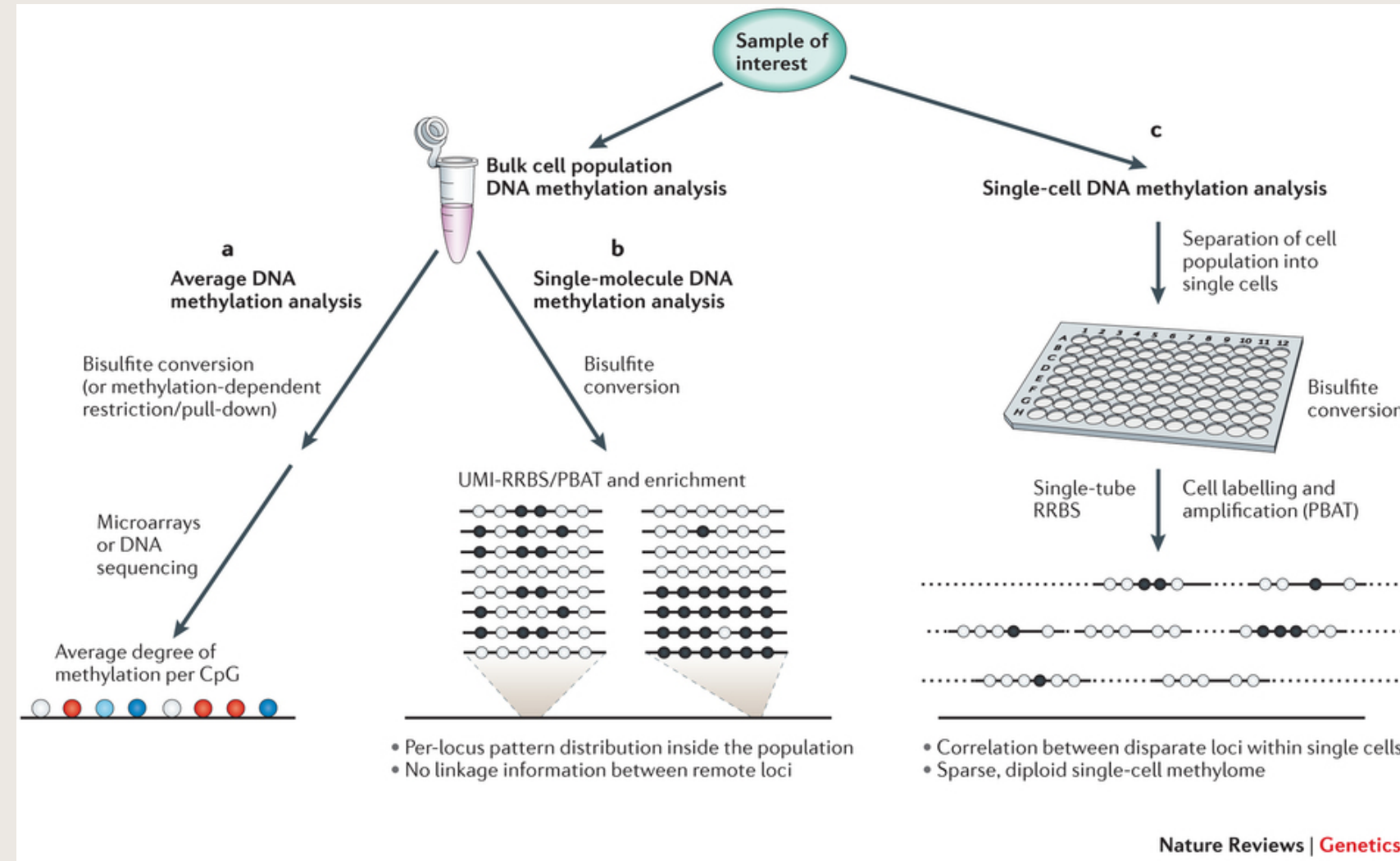
DNA Methylation, both bulk and single cell Schwartzman et al (2015) Nature Reviews in Genetics (used by permission)

Reproducible Research with CWL and Docker

Reproducible research means completely reproducing a given bioinformatic analysis This requires having the exact **data**, **code** and **software** that was used.

- Open data in bioinformatics is a fairly solved problem.
- Code is getting there with RMarkdown/Jupyter, but could be better.
- Software versions (and accompanying OS/ecosystem) are a big problem.

CWL aims to provide a uniform and fully reproducible way of representing bioinformatics workflows. Docker aims to provide the exact environment in which an analysis was run. Together, they promise to help bioinformaticians to publish fully reproducible research.

As a side effect, Dockstore enables easy sharing of workflow components to help build new workflows.

Enter Screw

Screw stands for Single Cell Reproducible Epigenomics Workflow

Screw aims to provide a series of CWL+Dockerised mini-workflows and workflow components for creating fully reproducible single-cell DNA methylation analysis.

What SC-WGBS Data Looks Like
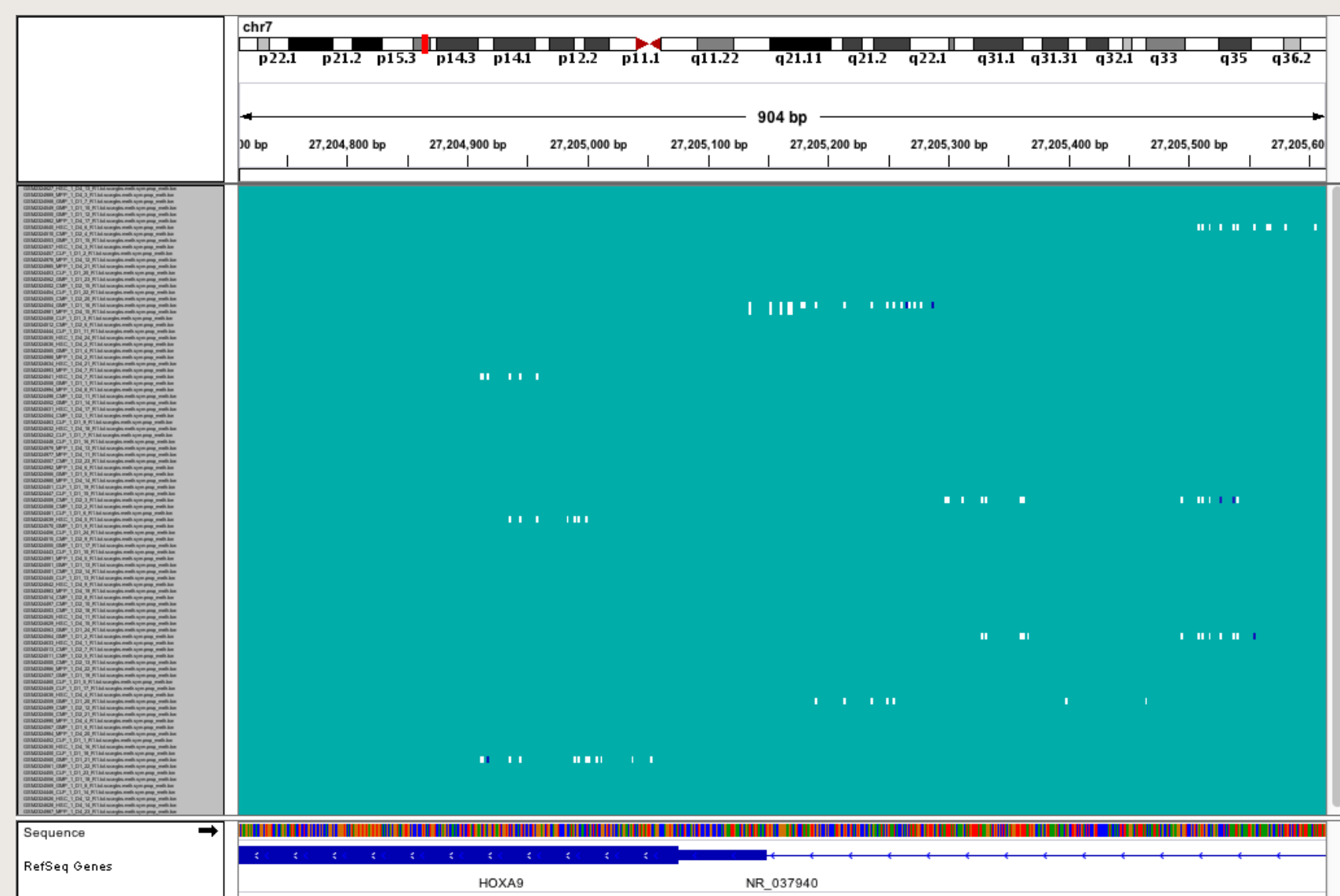
DNA methylation around the *HOXA9* TSS, from 100 single cells. Blue blocks indicate methylated CpGs, white unmethylated. Teal indicates missing data. DNA methylation data is very sparse – only a handful of libraries have any coverage at all within a given window. Data from [1].

Preprocessing: Format Interconversion

One of the challenges in DNA methylation is the lack of a standard data format. BED files are poorly suited, since multiple data columns are needed. Also, CpG sites are all point locations, so the end column is unncessary. WIG/BigWIG can help, but don't support enough data columns. The consequence is that every data set has its own custom BED-like format. We are working on providing
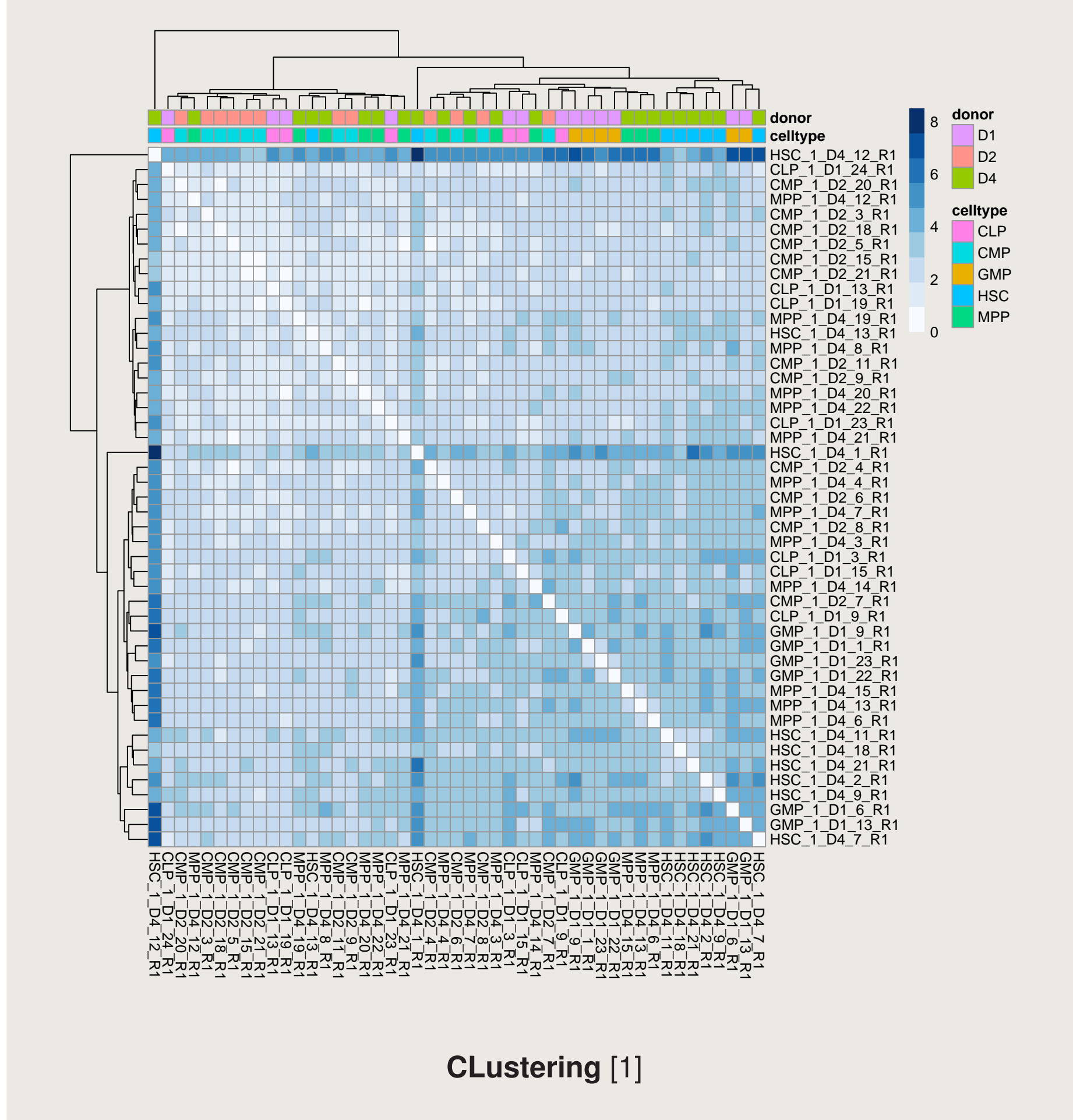

Diverse DNA methylation formats. Formats from two published papers ([**Farlik2013**] and [1]), NovoMethyl output format (unpublished), and Screw's standard internal format.

Screw Workflow
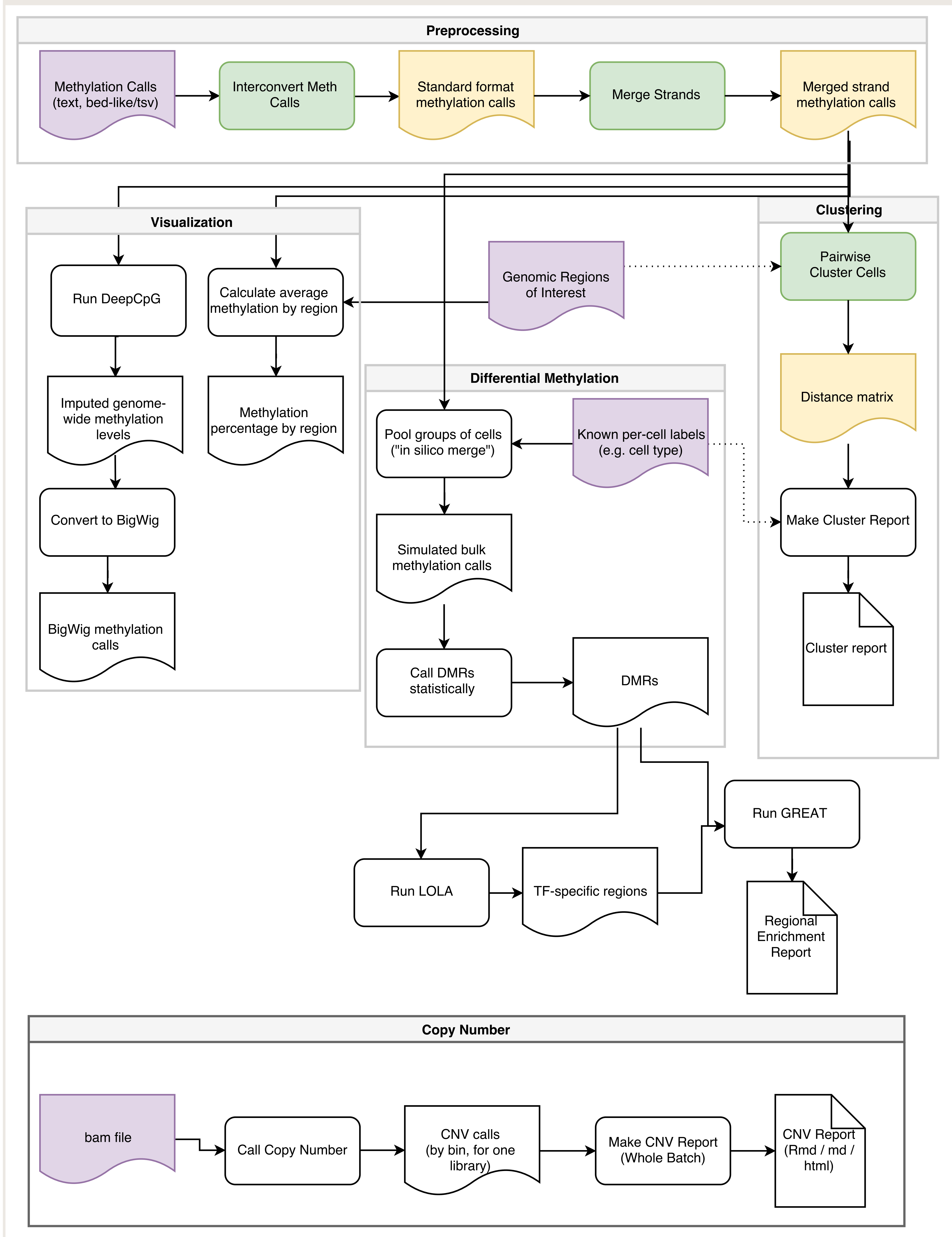
Clustering

CLustering [1]

CWL Stumbling Blocks

Future Plans

Many methods are currently being developed, and some have debuted in papers without (yet) an accompanying methods paper. In future we plan to include functionality for:

- tSNE plotting

Powered by hackathons and interns

Screw was initially created during the NCBI Genomics Hackathon in March 2017, organised by Ben Busby. Ben organises awesome hackathons that drive a lot of new open source tool development. Since then, Chelsey Fang, a summer intern, has been working on the framework.

The Screw team. From left to right, Azhar Khandekar, Benjamin Decato, , Kieran O'Neill, Alexander Goncearenco; separate photo: Chelsey Fang.

References

[1] M Farlik et al. "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation". In: *Cell Stem Cell* 19.6 (2016), pp. 808–822.

[2] O Schwartzman and A Tanay. "Single-cell epigenomics: techniques and emerging applications". In: *Nature Reviews Genetics* 16.12 (2015), pp. 716–26.

Contact/GitHub

Code/examples: https://github.com/Epigenomics-Screw

Kieran O'Neill: koneill@bcgsc.ca