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

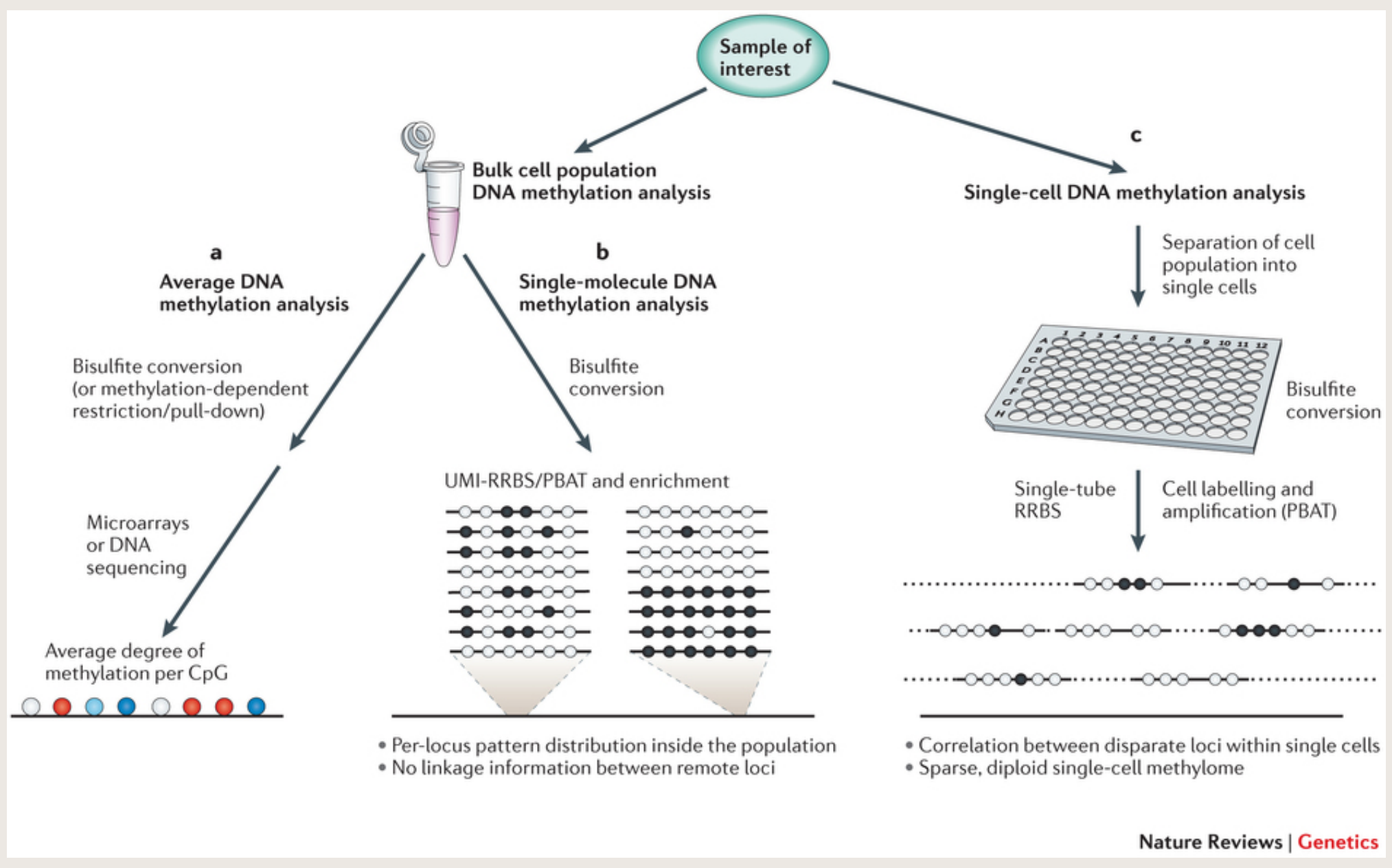
Screw: tools for building reproducible single-cell epigenomics workflows

Kieran O'Neill[1], Chelsey Fang[1], Benjamin Decato[2], Azhar Khandekar[3], Alexander Goncearenco[3], Ben Busby[3] , Aly Karsan[1]

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

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. However, 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.


Sample of interest
Bulk cell population DNA methylation analysis
a
Average DNA methylation analysis
b
Single-molecule DNA methylation analysis
c
Single-cell DNA methylation analysis
Bisulfite conversion (or methylation-dependent restriction/pull-down)
Bisulfite conversion
Separation of cell population into single cells
Bisulfite conversion
Microarrays or DNA sequencing
UMI-RRBS/PBAT and enrichment
Single-tube RRBS
Cell labelling and amplification (PBAT)
Average degree of methylation per CpG
• Per-locus pattern distribution inside the population
• No linkage information between remote loci
• Correlation between disparate loci within single cells
• Sparse, diploid single-cell methylome
Nature Reviews | Genetics

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

Screw: Single Cell Reproducible Epigenomics Workflow

Screw aims to provide a series of CWL+Dockerised tools, mini-workflows and complete template workflows for creating fully reproducible SC-WGBS analyses.

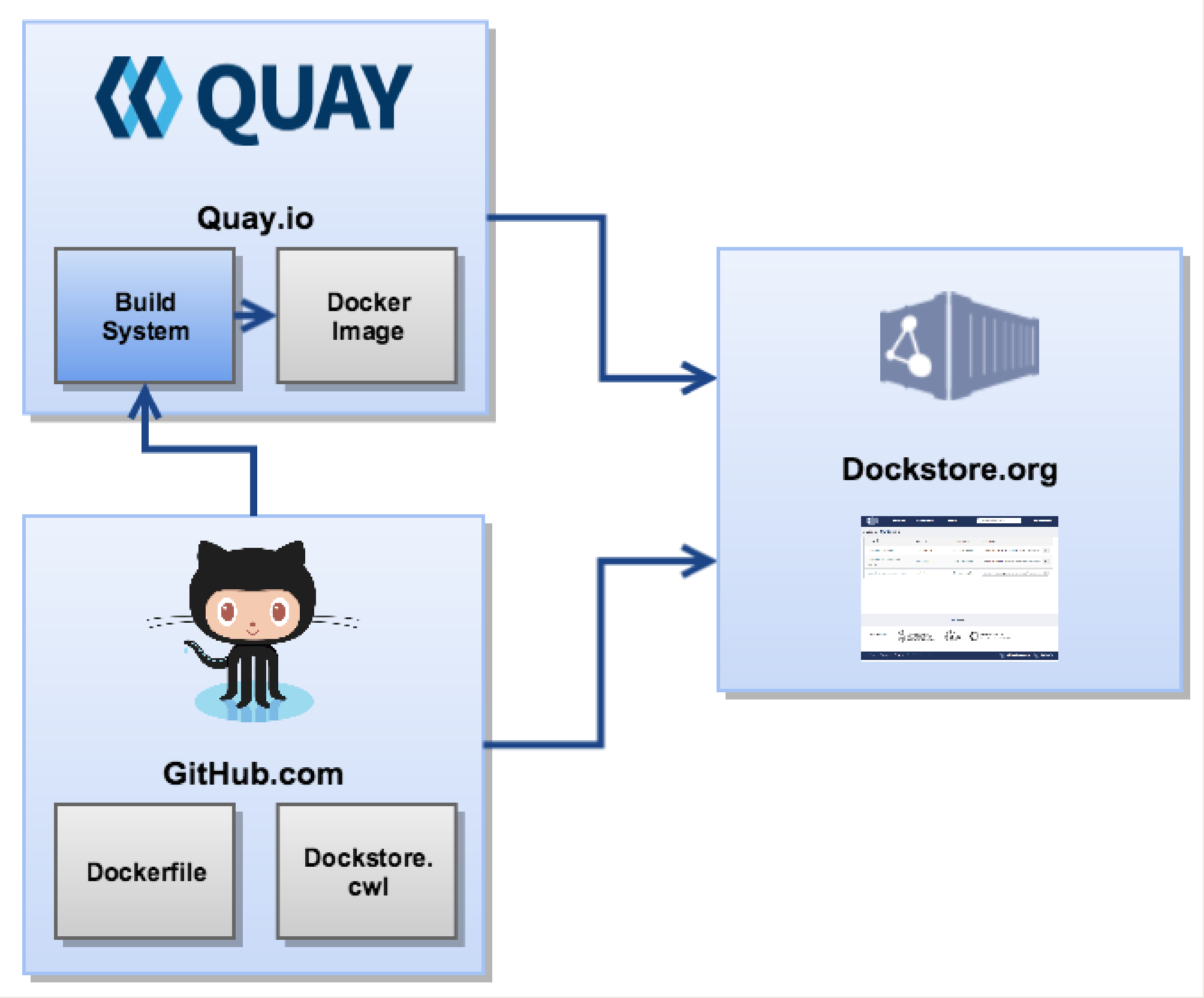
Reproducible Research with CWL and Docker

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 solve the code problem by providing a uniform and fully reproducible way of representing bioinformatics workflows. **Docker** aims to solve the software version problem by providing the exact environment in which an analysis was run. Together, they promise to help bioinformaticians to publish fully reproducible research.

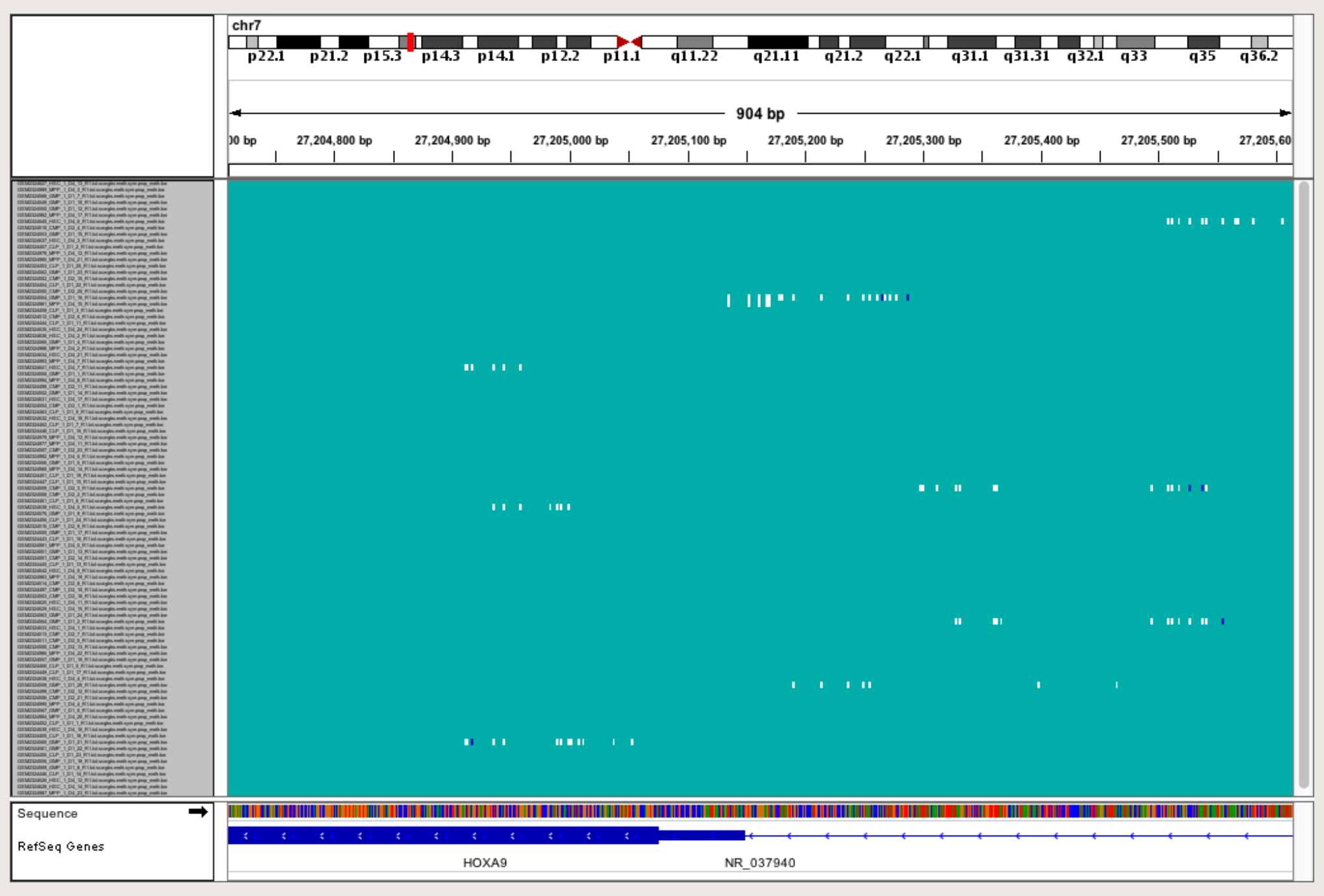
Dockstore enables easy sharing of workflow components to help build new, reproducible, workflows.



Docker, CWL and Dockstore. CWL workflows (and Docker build files) are kept on GitHub. Quay.io automatically builds Docker images from these. Dockstore enables sharing of CWL-specified tools and workflows, along with a Docker image containing exact software. Image reused from dockstore.org under the Apache 2.0 license.

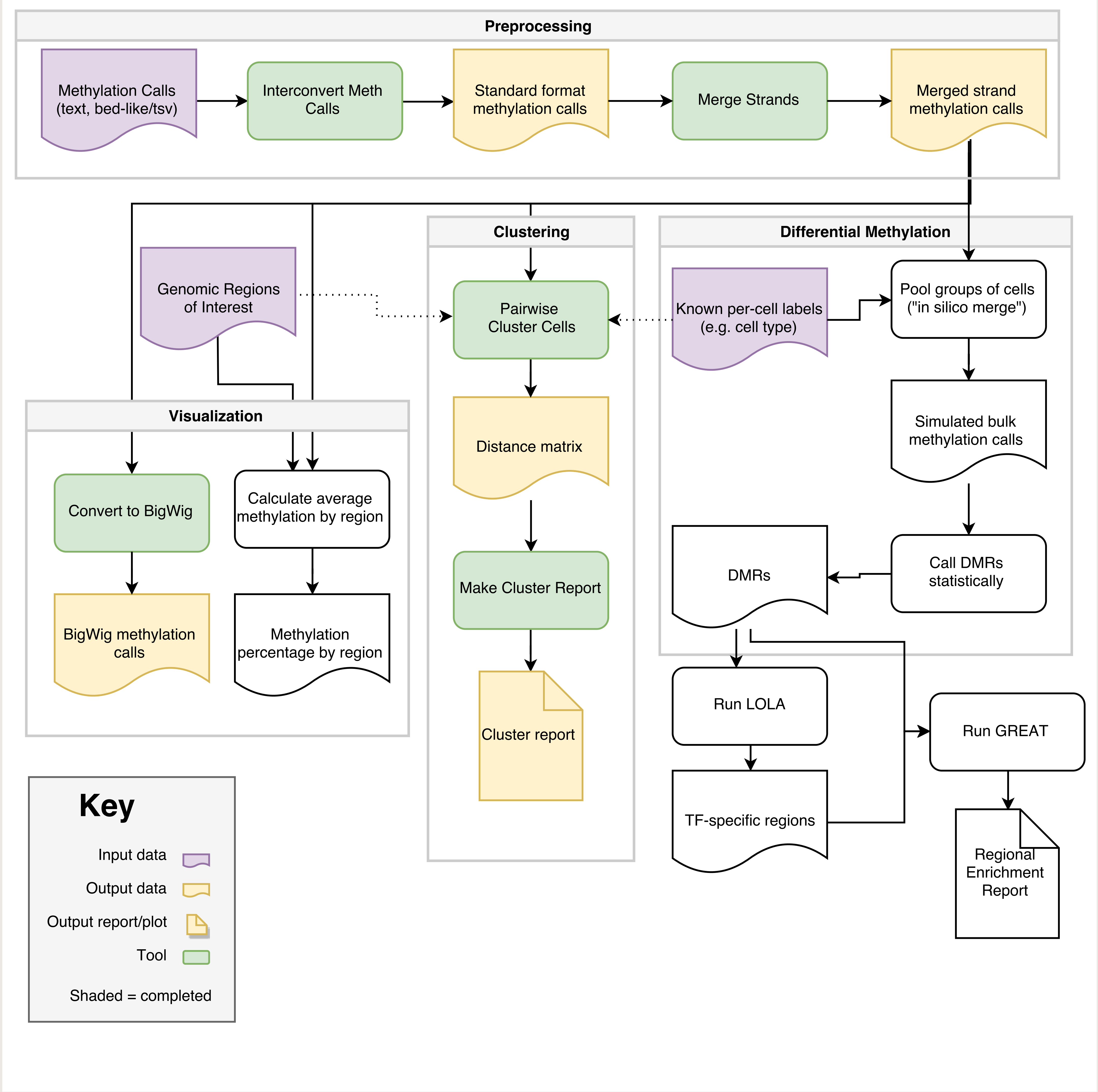
What SC-WGBS Data Looks Like

SC-WGBS data tends to have a reasonable number of cells (100 in currently published data, up to thousands in data being generated now). It also tends to have a lot of dropout – lateral coverage in each cell ranges from 1% to 20%.



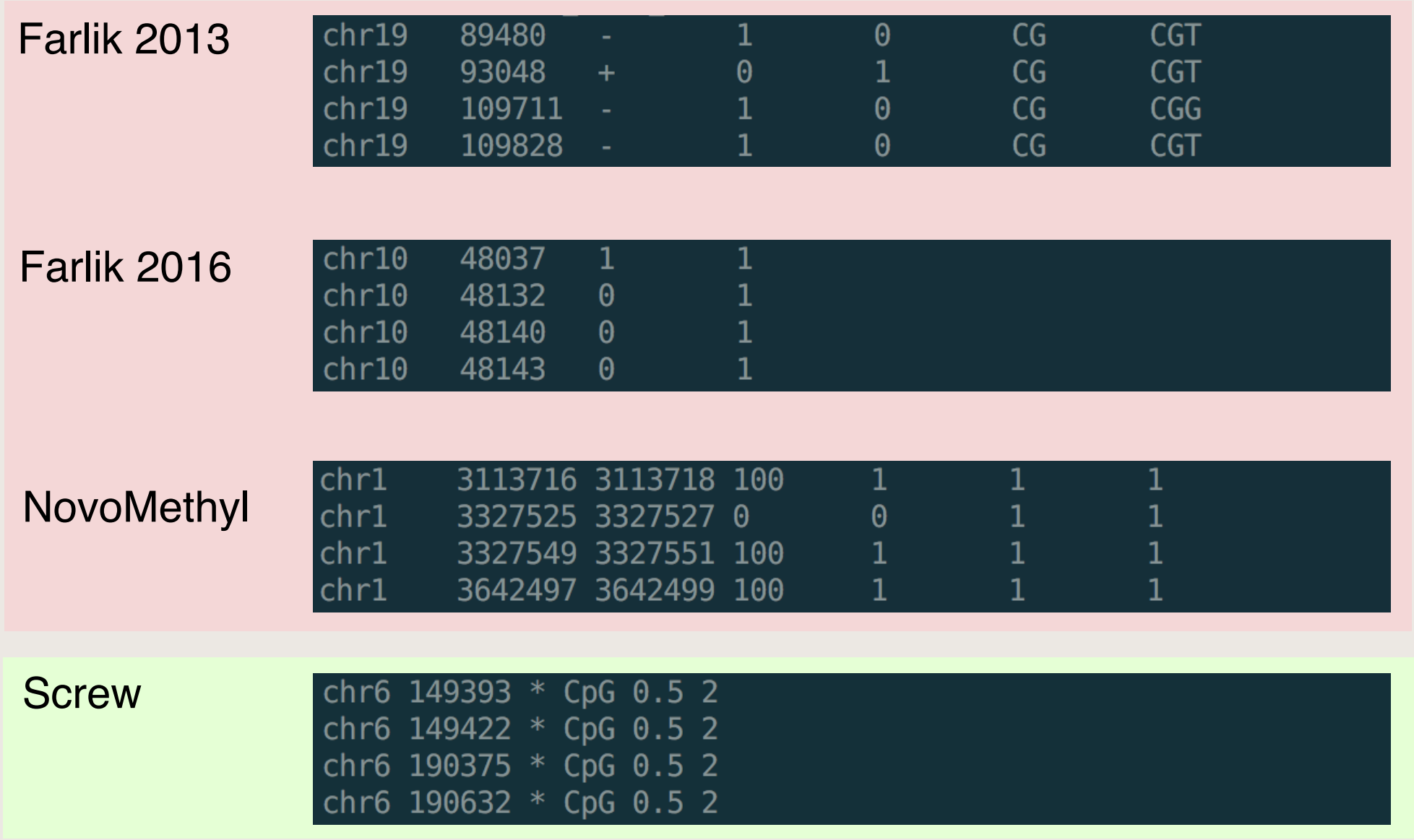
DNA methylation around the *HOXA9* TSS, from 100 single cells. Blue blocks indicate methylated CpGs, white unmethylated. Teal indicates missing data. DNA methylation data is very sparse – only a handful of libraries have any coverage at all within a given window. Data from [1]; BigWIGS were generated by Screw and visualised in IGV.

Screw Workflow


Preprocessing
Methylation Calls (text, bed-like/tsv)
Interconvert Meth Calls
Standard format methylation calls
Merge Strands
Merged strand methylation calls
Clustering
Differential Methylation
Genomic Regions of Interest
Pairwise Cluster Cells
Known per-cell labels (e.g. cell type)
Pool groups of cells ("in silico merge")
Visualization
Distance matrix
Simulated bulk methylation calls
Convert to BigWig
Calculate average methylation by region
Make Cluster Report
DMRs
Call DMRs statistically
BigWig methylation calls
Methylation percentage by region
Cluster report
Run LOLA
Run GREAT
TF-specific regions
Regional Enrichment Report
Key
Input data
Output data
Output report/plot
Tool
Shaded = completed

Preprocessing: Format Interconversion

One of the challenges in DNA methylation is the lack of a standard data format. Neither BED nor WIG/BigWIG quite suit the requirements, so every data set uses its own custom BED-like format. Screw provides a script to convert from different formats into a format used internally by the rest of the Screw tools. We will add formats as we analyse existing public data sets.



Farlik 2013

```
chr19   89480   -       1       0       CG      CGT
chr19   93048   +       0       1       CG      CGT
chr19   109711  -       1       0       CG      CGG
chr19   109828  -       1       0       CG      CGT
```

Farlik 2016

```
chr10   48037   1       1
chr10   48132   0       1
chr10   48140   0       1
chr10   48143   0       1
```

NovoMethyl

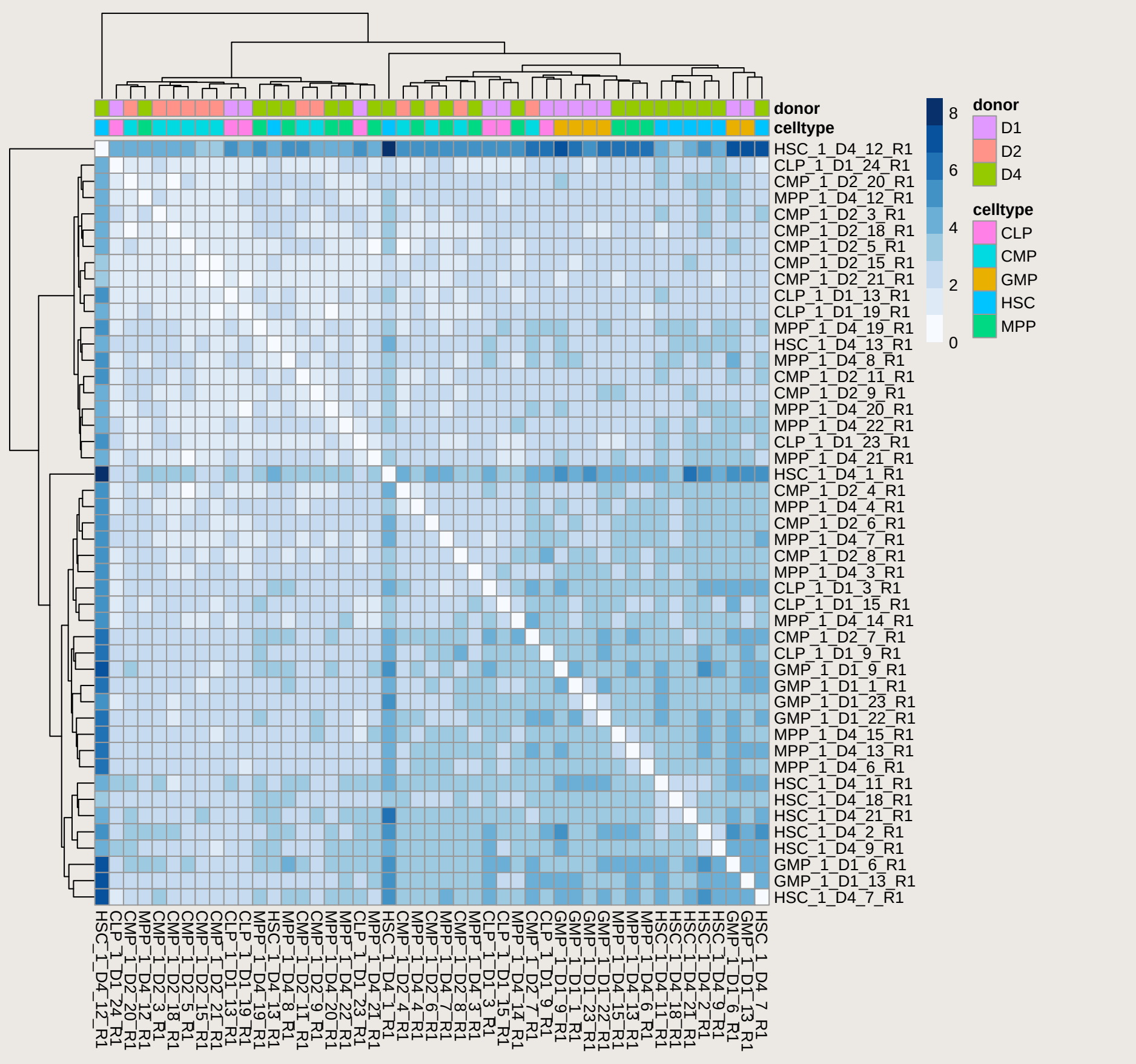
```
chr1    3113716 3113718 100     1       1       1
chr1    3327525 3327527 0       0       1       1
chr1    3327549 3327551 100     1       1       1
chr1    3642497 3642499 100     1       1       1
```

Screw

```
chr6 149393 * CpG 0.5 2
chr6 149422 * CpG 0.5 2
chr6 190375 * CpG 0.5 2
chr6 190632 * CpG 0.5 2
```

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

Clustering



Clustering of data from [1] using Screw. A BED file containing locations of active enhancer sites in CD34+ bone marrow was used to filter the CpG sites. (Data from Enhancer Atlas.) Only 47/100 cells had coverage in those enhancer regions.

CWL+Docker Technical Stumbling Blocks

- CWL does not have a good way of processing a whole directory full of files, then passing them on to a single tool.
- Quay.io maps tags to GitHub branches, but Docker treats tags like Git graph nodes, leading to problems refreshing branches while developing.
- Our hyper-secure clinical IT environment won't let us use Docker, but will install Udocker, which CWL runner does not support (yet?)

Future Work

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
- Epiphylogenomics
- DeepCpG
- Other machine learning pipelines (e.g. from [1])

There are also around half a dozen SC-WGBS (or SC-RRBS) data sets published. We intend to use Screw to perform the first single cell DNA methylation meta-analysis. All code and data will be published in a fully-reproducible form.

Powered by hackathons and interns

Development of Screw began at the NCBI Genomics Hackathon in March 2017, organised by Ben Busby. Since then, Chelsey Fang, a summer intern, has been working on the framework.

The Screw team. From left to right, Azhar Khandekar, Benjamin Decato, Paul Cantalupo (who worked in parallel on a similar project), 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