Intro

**What is epigenetics?**

Epigenetics is the study of heritable and reversible changes of the genome that do not change the underlying sequence. These changes are fundamental for controlling gene expression in all organisms, and can occur through intrinsic methods (e.g. X-chromosome inactivation in females) or external causes (e.g. exposure to ultra-violet radiation). The molecular marks of epigenetics span several levels from overall chromatin structure, through histone modifications, down to changes of individual bases, with methylation.

**Describe epigenetic marks**

Methylation is the most studied epigenetic modification. Adenine and cytosine both readily accept methylation on the N6 and 4,5 C4,5, respectively. However, while methylated adenine has shown presence in mammalian DNA, little evidence has shown it has any effect. Cytosine modifications have been long studied in archaea, bacteria, and eukaryotes. 5-methyl-cytosine (5mC) is the most common epigenetic mark in humans, with about 28 million sites present in the genome. These sites, known as CpG sites for the cytosine-phosphate-guanine bridge, are not random but instead highly localized in regulatory motifs (e.g. 70% of promoters show high density of sites, also known as a CpG island). Methylated CpG sites provide a physical block to RNA polymerase attempting to bind to and transcribe a strand of DNA, so they have great capacity to affect gene expression. As such, mechanisms for controlling methylation must be strictly controlled.

These epigenetic marks can be *de novo* added by DNA Methyltransferase proteins (DNMT3a/b), or removed by ten-eleven translocation methylcytosine dioxygenases (TET family), so these modifications are considered reversible. These are necessary for normal development of tissues, as Example of adding and removing. Furthermore, these modifications can also be maintained through cell division by DNMT1, hence they are considered heritable. Example of heritable

These mechanisms must be strictly controlled, as dysregulation can significantly affect health and disease. Example of sickness by epigenetics

Implications to cancer

Malignant cell growth is also strongly linked to epigenetic abnormalities. Abnormal overexpression of DNMT proteins has shown significant effects on multiple cancers (e.g. DNMT3A in 25% of acute myeloid leukemia cases and DNMT1 in 12% of uterine cancer cases <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465608/#:~:text=found%20that%20the%20coding%20exons,characteristic%20change%20of%20tumor%20cells>.). Conversely, underexpression of TET proteins has been shown in some cancer types (e.g. up to 58% of chronic myelomonocytic leukemia cases). Thus, epimutations can both increase and decrease gene expression, which is problematic with pro-tumor and tumor-suppressing genes, respectively. Example related to later work. Alterations of the epigenomic landscape can cause widespread tumor and disease, and understanding this process is critical for well being.

Whole genome bisulfite sequencing (WGBS) is the *de facto* technique for studying the epigenomic landscape. In this technique, sodium bisulfite converts unmethylated cytosines to uracil (and later thymine) which allows differentiation by comparing pre- and post-treatment sequencing and alignment to the reference methylome. The resulting single nucleotide polymorphisms (SNPs) allow generation of the epigenomic landscape. However, coverage of at least 30x is recommended with this technique, as during treatment, single strand nicks are randomly introduced, so up to 95% of the CpG sites lost during sequencing cite. This causes problems with low population samples. Recently, new techniques focusing on single cell data have emerged to address these problems. Single cell bisulfite sequencing (sc-BS) and XXX

Existing methods

Importance of single cell methods

However, WGBS suffers some drawbacks due to its bulk processing. Low population samples cannot give suitable coverage and is difficult to differentiate between heterogenous cells. Single cell bisulfite sequencing (sc-BS) fills this gap.

Furthermore, the data obtained from sc-BS presents its own problems. While low coverage can be accommodated by many tools, sparsity of the data is a challenge. There are inherent losses with bisulfite treatment as described above. This imparts many NA values in the output data if a dense matrix is used, as is necessary for most existing tools. Conversely, high proportion of NA values are also not accepted by these tools, in addition to the high memory requirement for such structures. Hence, we have developed a new tool to address these problems and allow a pipeline more similar to traditional WGBS tools.

To address these problems, we introduce scMethrix, a methylation-specific single cell data storage and manipulation tool. This R package is well-suited for the specific needs of single cell data and for integration into the Bioconductor ecosystem.