Introduction to Reduced-Representation Methylation Sequencing (RRMS)

CpG dinucleotides frequently occur in high-density clusters called CpG islands (CGI) and most vertebrate genes have their promoters embedded within CGIs. Determining the methylation status of cytosines within CpGs is of substantial biological interest: alterations in methylation patterns within promoters is associated with changes in gene expression and disease states such as cancer. Exploring methylation differences between tumour samples and normal samples can help to elucidate mechanisms associated with tumour formation and development.

There are two primary methods which are commonly used to identify and quantify genome-wide DNA methylation:

- Affinity capture of methylated DNA – locus-specific or genome-wide, e.g. MeDIP, EPIC arrays
- Sodium bisulfite conversion and sequencing

Being antibody-based, MeDIP, is only really effective where there are high densities of CpGs. There is a high degree of off-target capture because of antibody cross-reactivity, but the main drawback is that MeDIP does not give single-base resolution. Although arrays can give single-base resolution, the number of CpG sites that an array can target is limited to around 935,000. As a consequence, sodium bisulfite sequencing has risen in popularity over the past few years. This method does provide single-base resolution and targets millions of CpGs, but it has several drawbacks:

- The conversion process make the reads hard to map, meaning that 1/3 of the data is unmappable
- The library prep is laborious and takes a long time
- It’s computationally intensive to do the analysis, meaning that this also takes a long time
- Regions with high densities of CpGs tend to be GC-rich, meaning that they don’t amplify well. As a result, regions are missed and the results are biased
- Consequently, only about 75% of the CpGs in a genome are accessed with 50x bisulfite sequencing
- Whole-genome bisulfite sequencing is also expensive – around $3,000 per sample

A more cost-effective alternative is reduced representation bisulfite sequencing (RRBS). This can be thought of as the methylation equivalent of exome sequencing. RRBS looks at around 1% of the genome, but because CpGs are not evenly distributed throughout the genome, RRBS captures 10-15% of CpGs in a mammalian genome, making it a more cost-efficient approach than WGS.

Adaptive sampling

Adaptive sampling (AS) offers a fast, flexible and precise method to enrich for regions of interest (e.g. CGIs) by depleting off-target regions during the sequencing run itself with no requirement for upfront sample manipulation: due to the real-time nature of nanopore sequencing it is possible to identify whether or not the strand that is being sequenced is within the region of interest (ROI): if the read does not map to the ROI the strand is ejected from the pore so it is able to accept a new strand. Off-target strands are continually rejected until a strand from the ROI is detected, and sequencing is allowed to proceed (Fig. 1).

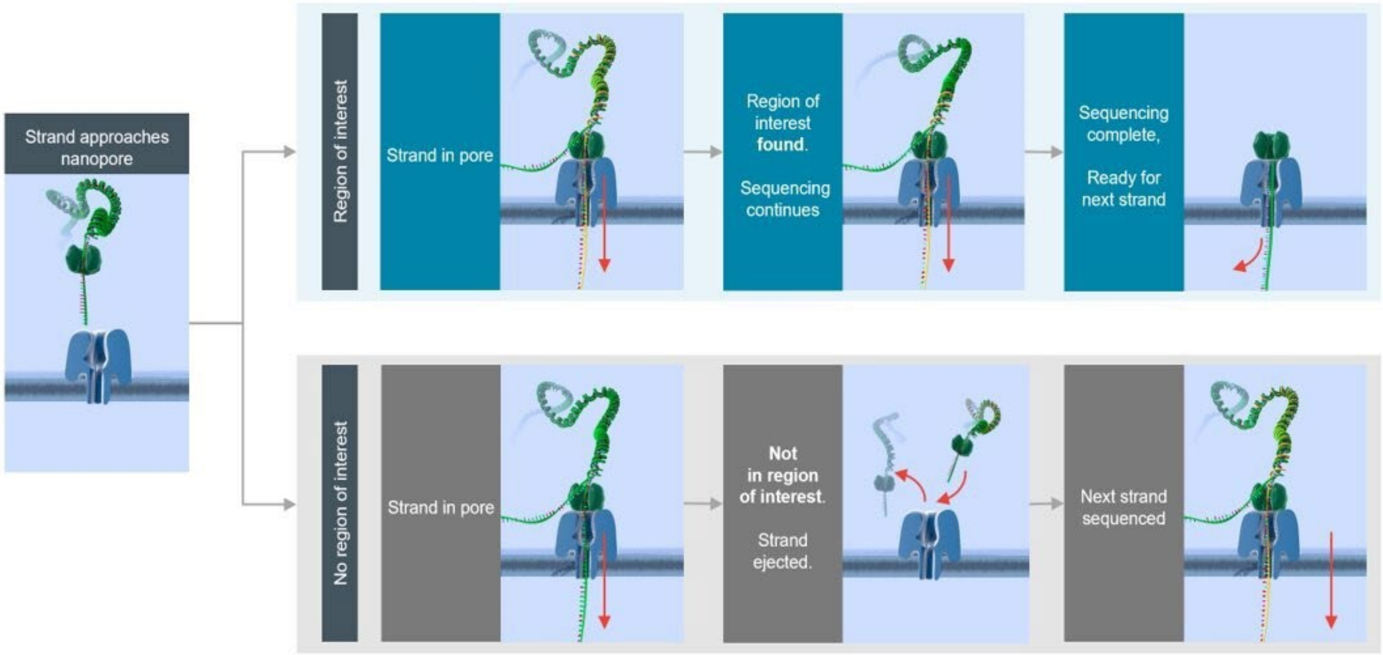

Figure 1. Schematic showing the principle behind adaptive sampling (AS)

What is RRMS?

Nanopore sequencing enables direct detection of methylated cytosines (e.g. at CpG sites), without the need for bisulphite conversion. Reduced Representation Methylation Sequencing (RRMS) uses Adaptive Sampling to target CpG islands, shores, shelves and annotated promoters within a reference genome. Remora, Oxford Nanopore's tool for direct methylation detection, is then used to generate high-confidence methylation calls. Currently, RRMS target files are available for human and mouse genomes. For human 310 Mb of the genome is targeted including all annotated CpG islands, shores and shelves as well as >90% of promoter regions (100% of promoter with more than 4 CpGs). In total, around 7.1 million CpGs are targeted in a single MinION run representing 24% of total CpGs in the human genome. For mouse 308 Mb of the genome is targeted including all annotated CpG islands, shores and shelves as well as all promoter regions. In total, around 4.7 million CpGs are targeted (22% of total CpGs in the mouse genome). This method enables the detection of differentially methylated regions between samples.

In addition, reads rejected as part of the adaptive sampling process can be used to provide whole genome copy number variation calling. RRMS therefore provides key information for tumour characterisation as well as methylation status, providing deeper insights into the mechanisms behind diseases such as cancer and monitoring tumour progression.

How is RRMS done?

Genomic DNA is extracted from human or mouse cells grown in culture using the QIAGEN Gentra Puregene Cell Kit. For gDNA extraction from blood or tissues, please see the Nanopore Community for more details. Extracted DNA is then fragmented using a g-TUBE (Covaris) and 2 µg of sheared DNA is prepared for sequencing using the Ligation Sequencing Kit.

There are currently two recommended sequencing configurations:

- MinION Flow Cell on MinION Mk1b, GridION Mk1 (single sample per flow cell)
- PromethION Flow Cell run on P2Solo or P24/P48 (up to at least 4 samples per flow cell)

To optimise output, flow cells are washed (and 150 ng of library re-loaded) twice.

Basecalling and calling of 5mC modifications is performed using *dorado basecaller* together with alignment against the desired reference using the option “--reference”. When barcoding multiple samples on a PromethION Flow Cell, demultiplexing can be performed while sequencing using the option “--kit <kit-name>” followed by “dorado demux” using “—no-classify” option to obtain an individual BAM per barcode.

Modkit pileup should be used to compute 5mC frequencies for all genomic CpG positions setting "--cpg --combine-strands”, this will combine methylation frequencies from forward and reverse strands and will aggregate frequencies per CpG position. CpGs with > 10 overlapping reads will be considered as high-confidence CpGs. Modkit dmr modules can be used to explore differences across different samples. For more information on the different options see *modkit* github repository (https://github.com/nanoporetech/modkit) (Fig. 2).

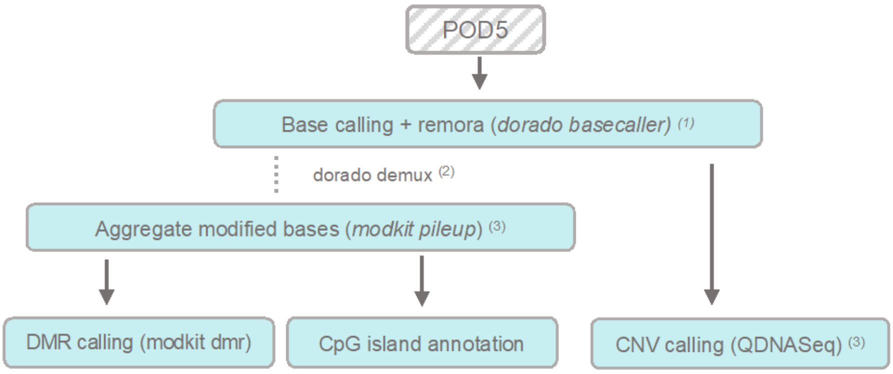
POD5
Base calling + remora (*dorado basecaller*) [1]
dorado demux [2]
Aggregate modified bases (*modkit pileup*) [3]
DMR calling (modkit dmr)
CpG island annotation
CNV calling (QDNASeq) [3]

Figure 2. Analysis pipeline for RRMS. [1] Alignment can be performed while basecalling using --reference option [2] Only required if working with barcoded samples – note demux step can be performed while basecalling by providing --kit option, if this is the case ---no-classify needs to be set when running dorado demux. [3] Implemented in wf-human-variation (https://github.com/epi2me-labs/wf-human-variation)

RRMS typically generates >20x average coverage per sample across target regions and recovers >22% of total CpGs in human or mouse genomes covering >90% of promoters, CGIs, shores and shelves. A comparison of number of high-confidence CpGs recovered and relevant features covered between RRMS and RRBS in human is shown below (Fig. 3).

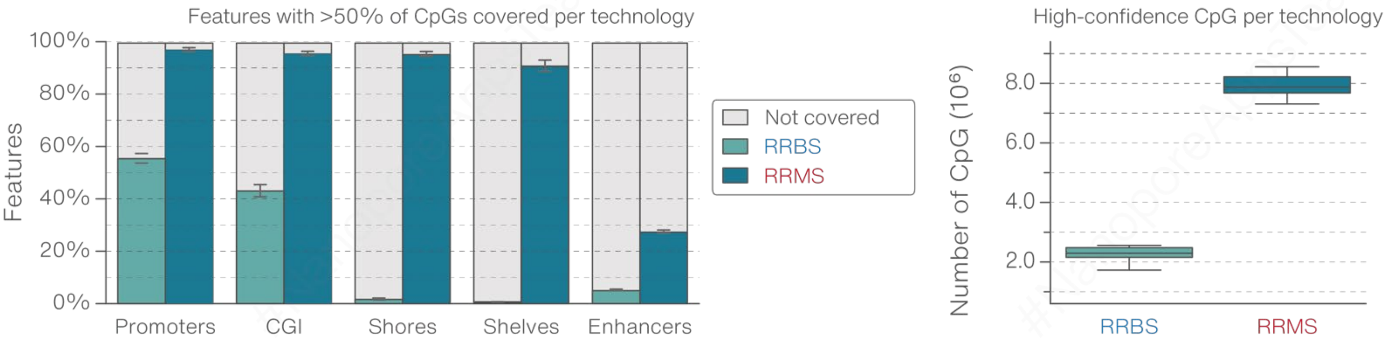

Figure 3. Typical performance of an RRMS run.

As an additional benefit of RRMS, it is possible to call copy-number variants (CNV) across the whole genome without any additional sample preparation (i.e. you get CNV calling "for free"), by using the reads which are ejected by AS during the course of an RRMS run (Fig. 4).

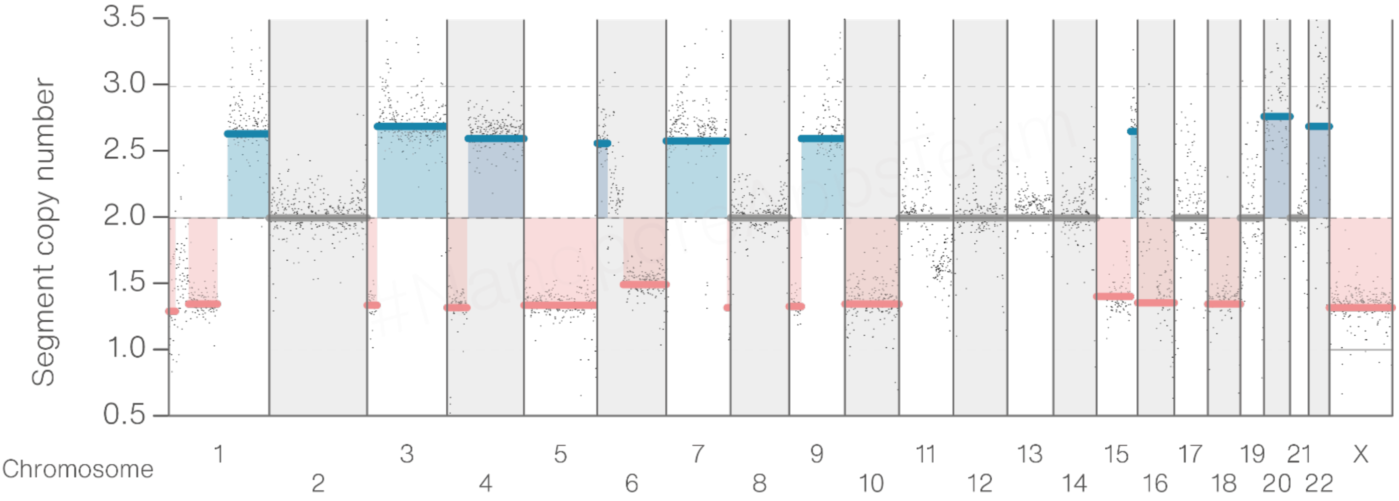

Figure 4. CNV calling using RRMS. The reads which are ejected during the course of an RRMS run can be used to call copy-number variation across the whole genome.

RRMS therefore provides key information for tumour characterisation as well as methylation status. Combined with its ease of use and ability to scale to a high number of samples, RRMS is well suited to investigating methylation differences in large cohorts, as well as providing deeper insights into the mechanisms behind diseases such as cancer and monitoring tumour progression.