**Adaptive sampling**

**Targeted sequencing**

In some sequencing applications, the focus of study — a single gene, or a selection of genomic regions — comprises a small fraction of the genome/sample. In these cases, characterisation through whole-genome sequencing can be inefficient and costly. Targeted sequencing is a term used to describe strategies that reduce the time spent sequencing regions that are not of interest thereby significantly reducing the amount of data required to achieve the desired depth of the regions of interest. As well as reducing sequencing cost, this reduces the data analysis burden and enables a quicker workflow. Targeted sequencing using nanopore technology can be achieved in several ways: amplicon sequencing, pull-down, cas9-based enrichment and adaptive sampling (AS).

**Introduction to adaptive sampling**

Adaptive sampling offers a fast and flexible method to enrich regions of interest by depleting off-target regions: target selection takes place during sequencing itself, with no requirement for upfront sample manipulation – the library is prepared and loaded as normal and “adaptive sampling” is selected in MinKNOW (typically a .bed file detailing target regions will need to be uploaded). Once the run commences, sequencing will begin and 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, MinKNOW will reverse the polarity of the applied potential, ejecting the strand from the pore so it is able to accept a new, different template strand. Off target strands are continually rejected until a strand from the ROI is detected and sequencing is allowed to proceed. AS can run in two different modes: “enrichment” and “depletion”. In “enrichment”, ROIs are uploaded to MinKNOW and strands that fall outside of this are rejected. In “depletion” mode, targets that are not of interest (e.g. host DNA in a host:microbiome metagenomic analysis) are uploaded to MinKNOW and strands that fall within these regions are rejected. In this document we discuss “enrichment” strategies only.  We generally, observe an enrichment for ROI of ~5-10-fold, when targeting using AS and the sections below outline our advice on how this can be achieved. For targeting regions within human genomes, we find this level of enrichment is robust as long as the total fraction that is being targeted is <10% of the total genome, and can enable users to obtain a mean depth >20-40x of ROI on a MinION flowcell.

**Experimental considerations**

Library input

For conventional (non-AS sequencing runs) sequencing runs, we recommend adding ~5-50 fmol of sequencing library to obtain optimum sequencing yield. During an adaptive sampling run, most of the template strands that enter the pores are NOT from the region of interest – these strands are thus quickly rejected/ejected from the pore and the pore returns to the open pore state: this can lead to an overall increase in open pore time. To mitigate, we recommend adding 50 fmol of sequencing library. However, we find that adding more than this does not provide additional benefit.



**Figure 1: Impact of library loading amount on flow cell output in adaptive sampling.** To demonstrate the impact of library input amount on flow cell output in AS runs, we generated a library from human genomic DNA. 10, 50 or 100 fmol of the library were run on GridION under AS conditions (targeting ~1% of the genome, split into 800 40,000 bp targets) and the output recorded. Panel A) displays the health of each flow cell (as represented in the GUI on MinKNOW) and shows that at a library load of 10 fmol, there is a significant number of pores in the open pore state. At 50 fmol, the pores are almost fully occupied. Panel B) displays the rate of data acquisition (Gb) over time and shows as the library load increases from 10 fmol to 50 fmol the output increases. However, little increase in output is observed by exceeding this.

Sequencing Kit

To date, our work with AS has focussed primarily on libraries generated using the Ligation Sequencing Kit. We have found that due to the improvements in sequencing kit components, the performance/output of AS is improved as you progress through the kit iterations: LSK109 < LSK110. Therefore, we recommend using the most up-to-date version of the kit that is available.

 

**Figure 2: Impact of sequencing kit chemistry on flow cell output in adaptive sampling.**To demonstrate the impact of sequencing kit chemistry on flow cell output in AS runs, we generated libraries from unfragmented and size selected human genomic DNA using various iterations of the Ligation Sequencing Kit: SQK-LSK109 and SQK-LSK110. 50 fmol of each library was run on GridION under AS conditions (targeting ~1% of the genome, split into 800 40,000 bp targets) and the output recorded.

Fragment length

The reversal of the potential difference applied across the membrane to eject off-target reads is the same mechanism by which MinKNOW “unblocks” pores. On rare occasions, pores are not successfully unblocked by this unblock mechanism: this is documented [elsewhere](https://community.nanoporetech.com/posts/blocking-unblocking-and-f#comment_26266). In this situation, the pore becomes “unavailable” for sequencing (terminally blocked) and over time, as more pores become unavailable, the rate of data acquisition begins to slow. Due to the persistent application of the “unblock” mechanism in AS runs (to reject off-target reads), the rate of pores becoming terminally blocked is often higher when compared with conventional runs. This effect can be exacerbated when input fragment lengths are longer. In order to maintain high data outputs with longer read libraries in AS runs, we recommend performing nuclease flushes (using SQK-WSH004) to clear terminally blocked pores, and then re-loading library.



**Figure 3: Impact of fragment length on flow cell output in adaptive sampling.**To demonstrate the impact on flow cell output in AS runs of different length libraries, we generated two libraries from human genomic DNA; one was fragmented using a Covaris g-TUBE to produce a read-N50 ~8 kb and the other was unfragmented and size selected to generate a read-N50 of 25 kb. The libraries were run on GridION under AS conditions (targeting ~1% of the genome, split into 800 40,000 bp targets). The rate of data acquisition (Gb) slows over time as an increased number of pores become unavailable: this rate of decay is faster with the longer libraries. Output can be increased by performing nuclease washes every ~20 hours (denoted by vertical arrows)

“Buffer” size

“Buffer” regions are flanking regions added to the side of every single target described in the .bed file. These regions allow reads which begin with a sequence that may not initially map to our target region, but may extend into our target region, to be accepted. By accepting reads which map into these flanking regions we increase the number of accepted reads that hit our target: while increasing the size of this buffer does lead to an increase in off target bases sequenced, it can also increase the depth of the regions of interest also. The size of the buffer chosen relates to the read length of the underlying library – we recommend setting the buffer size to the read-N10 of the library (this is point at which 10% of the bases sequenced are from reads that are this length or longer). Having a buffer size that exceeds the length of all reads present in the library will lead to reads being accepted and sequenced that never enter the target region.



**Figure 4: Impact of buffer size on sequencing depth in adaptive sampling.**To demonstrate the impact of buffer size on sequencing depth in AS runs, we generated libraries from unsheared and size selected human genomic DNA (to produce a read-N50 ~25 kb; the read length of the library was determined by running on GridION under non-AS conditions). We then performed the AS runs: we targeted (at random) 800 40,000 bp targets (this equates to 1% of the genome). The buffer size upstream and downstream of each target was set to ~20 kb, ~47 kb, ~62 kb and ~88 kb to represent the read-N75, -N25, -N10 and -N01 of the library. We observed that maximum target depth was obtained when the buffer size was set to the -N10 of the library.

Note, as the number of targets increases, the percentage of the genome that is being targeted will increase. Once the total fraction being targeted breaches 10-15% (ROI + buffer), we find that the level of enrichment can drop – see Figure 5. Therefore, it may be necessary to either fragment the library (this has the impact of reducing the buffer size around each target), or reduce the buffer size from read-N10 (e.g. to read-N75).

Target region

In preliminary AS runs we typically observed 5-10 fold enrichment of the ROI: this equates to ~20-40x sequencing depth of the targets in human genome tests. To test how robust this level of enrichment was (for human genome studies), we titrated the number of targets and the size of the targets in various scenarios and recorded the sequencing depth obtained in control runs (no AS) and in AS runs. We found that this level of enrichment/depth is observed when the fraction of the genome targeted is <10% regardless of the number of targets or the size of the targets.



**Figure 5: Impact of configuration of target region on sequencing depth in adaptive sampling.**To demonstrate the effect of different target configurations on sequencing depth in AS runs we generated libraries from unsheared and size selected human genomic DNA (to produce a read-N50 ~25 kb; the read length of the library was determined by running on GridION under non-AS conditions). We then performed the AS runs with various configurations of target regions (different numbers of targets, different sizes of targets, with different fractions of the genome targeted) and recorded the sequencing depth of our ROIs. The upstream and downstream buffer size was fixed as the read-N10 of the library (~50 kb) and this value added to both side of each individual ROI in all conditions. Panel A) In this experiment, we enriched for different numbers of 30,000 bp targets to cover ~0.06% (60 targets), ~0.11% (120 targets), ~0.56% (600 targets), ~1.1% (1,200 targets) and ~5.6% (6,000x targets) of the genome. Panel B) In this experiment, the number of targets is set at 600, but the size of each target is altered to cover ~0.06% (3,000 bp targets), ~0.11% (6,000 bp targets), ~0.56% (30,000 bp targets), ~1.1% (60,000 bp targets) and ~5.6% (300,000 bp targets) of the genome. Panel C) In this experiment, the fraction of the genome targeted was set at ~0.56% (1.8 Mb), but the number and size of targets was altered. Panel D) The data from panels A, B and C is amalgamated to show how the size of the AS.bed file (which comprises ROI + buffer regions) relates to depth of target.

As the % of the genome targeted increases (particularly above 20%), the depth will drop to 10-20x. However, it may be possible to recover performance by either fragmenting the library (this has the impact of reducing the buffer size around each target by reducing the length of the read-N90), or reducing the buffer size from the read-N90 value (e.g. to read-N25): these strategies can reduce the % of the genome targeted and may increase sequencing depth of target region.

Chromosome enrichment

AS experiments were also performed to target whole chromosomes: no buffer region is required in such a scenario. We targeted chromosome 7 (159 Mb) and chromosome 21 (47 Mb) by AS and recorded the depth obtained compared with control runs: the results obtained are in line with previous observations – sequencing depth increases ~5-10-fold to around 30x.



**Figure 6: Targeting whole chromosomes by AS.**We prepared sequencing libraries from unsheared and size selected human genomic DNA (to produce a read-N50 ~25 kb; the read length of the library was determined by running on GridION under non-AS conditions). We then performed the AS runs to target either chromosome 7 or chromosome 21 and recorded the depth of the targeted chromosome.