Material Book

Marcel Ferreira

2025-07-23

Table of contents

About this book

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

Part I

Sequencing

This page is under construction

2 Nanopore sequencing

2.1 Intro

Within the panorama of third-generation sequencing, ONT stands as a vanguard, contributing distinctive innovations to the DNA/RNA sequencing landscape. Pioneering the concept of nanopore sequencing, ONT employs biological nanopores embedded in synthetic membranes to analyze nucleic acids in real time [37], [39]. The portable MinION device, emblematic of Oxford Nanopore's approach, enables on-demand, long-read sequencing with unparalleled flexibility, given its small size (105 mm×23 mm x 33 mm) and weight (87 g) [47], [48], [49]. The nanopore-based sequencing mechanism allows for the direct, electronic detection of nucleotide sequences as they pass through the nanopore, offering advantages regarding read length and adaptability to various sample. For example, this allows the RNA strand to be sequenced directly, i.e., without synthesizing complementary DNA, making it easier to identify isoforms and determine the length of the poly-A tail [39], [50], [51], [52]. Bypassing bias-inducing steps like polymerase chain reaction (PCR), direct sequencing provides accurate, native data on the target molecule [53], [54], [55]. Nanopore sequencing has two modes: simplex and duplex. In the simplex mode (also known as 1D), a nanopore sequences only one DNA strand. In the duplex mode, both strands of DNA are sequenced (one immediately after the other), resulting in a twofold sequencing procedure that facilitates base correction [39], [56], [57].

Nanopore sequencing employs specialized flow cells tailored to distinct devices, each with unique capacities. The Flongle Flow Cell (R9.4.1) has 126 channels, interfaces with Flongle, MinION, and GridION devices, and achieves 2.8 Gb theoretical maximum output (TMO) for 1D experiments. The MinION Flow Cell (R9.4.1) supports MinION and GridION with 512 channels and a 50 Gb TMO. The PromethION Flow Cell (R10.4.1) caters to PromethION 2 Solo, P24, and P48, offering 2675 channels and a 290 Gb TMO. It features R10 nanopores with a double reader-head configuration, ideal for high-accuracy experiments, achieving over 99 % accuracy with Kit 14 chemistry. The MinION is a portable sequencing device with a 50 Gb TMO, providing immediate access to long-read data in various settings. The GridION, accommodating 1–5 flow cells, is a benchtop device with integrated computing, delivering a 250 Gb TMO for versatile applications. PromethION 2 devices offer high-yield sequencing (580 Gb) for small to medium-sized labs, supporting up to 200 flow cells per year. The PromethION series includes P24 and P48, with P48 providing twice the capacity (14 Tb) of P24, suitable for large-scale projects.

In 2023, Nature declared long-read sequencing as the Method of the Year, marking a significant milestone in genomic research [58]. These innovations have driven modern genomics forward, promoting substantial advances in personalized medicine, genetic variability studies, and understanding genomic complexities.

2.2 DNA/Extraction

2.3 Sequencing

2.4 Bioinformatics analysis

> The generated files are very large!
>
> Pay attention to the resources available on your system. POD5 files generated from whole genome sequencing with PromethION can easily exceed 1 TB. Aligned BAMs can exceed 200 GB.
> If you do not have sufficient storage, we recommend that you cut out a region (a gene, for example) to continue.

2.4.1 Basecalling

The first step is to convert the electrical signal stored in the POD5 files into the sequence bases. In nanopore this is performed by Dorado (Github| Documentation).

> Download the pre-compiled release!
>
> https://github.com/nanoporetech/dorado/releases

2.4.1.1 Dorado basecalling

The Dorado basecalling command is:

```
dorado basecaller hac pod5s/ > calls.bam
```

Where `hac` is the argument for the ***High Accuracy model***, `pod5s/` the path for your POD5 files, and `calls.bam` is the unaligned BAM file. This command will save the output in your current directory. Although this command is sufficient to run the basecall, there are some additional arguments that you may find useful.

2.4.1.2 Dorado models

Dorado can automatically select a basecalling model based on a chosen model speed (`fast`, `hac`, `sup`) and the POD5 data. If the model is not available locally, Dorado will automatically download it for use.

Additionally, Dorado supports the use of model paths.

To list the available models, run:

```
dorado download --list "all"
```

If you wish and have the space, you can download all available templates. Otherwise, please specify the desired model. I recommend that you use the `--models-directory` argument so that you can control where the models will be stored.

```
#All models
dorado download --model all --models-directory {PATH}

#Specific model
dorado download --model {MODEL} --models-directory {PATH}
```

The naming of Dorado models is systematically structured; each segment corresponds to a different aspect of the model, including both chemistry and run settings. Below are some examples of the simplex basecalling models:

`dna_r10.4.1_e8.2_400bps_sup@v5.2.0`

`rna004_130bps_hac@v5.2.0`

The structure of dorado models names is:

`{analyte}_{pore}_{chemistry}_{speed}@version`

2.4.2 Quality control (QC)

2.4.3 Genome Mapping

2.4.4 Variant calling

2.4.5 Base modification analysis

Part II

R packages

This page is under construction

Part III

Machine Learing

This page is under construction

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.