biomodal

Bioinformatics Pipeline

Data Interpretation Guide

For analysis of libraries sequenced on an Illumina platform and prepared using:

- **biomodal duet multiomics solution +modC**
- **biomodal duet multiomics solution evoC**

This documentation is compatible with duet pipeline version 1.3

For Research Use Only. Not for use in diagnostic procedures.

Table of Contents

This documentation is compatible with duet pipeline version 1.3.*

Introduction

The **biomodal duet multiomics solution bioinformatics pipeline** is a bioinformatics tool used for analysing genetic and epigenetic information present in a sample. It can be used with both double-stranded genomic DNA and cell-free DNA libraries prepared using the **biomodal duet multiomics solution +modC** or **evoC** and sequenced on an NGS (Next Generation Sequencing) sequencer. The +modC assay is capable of detecting modified cytosine bases (modC), which can mean either 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC) without distinguishing between them. The evoC assay is capable of detecting 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) in CpG contexts and modified cytosine bases (modC) in CpH contexts. This is

achieved using a two-base code, which translates into 16 unambiguous states and enables suppression of errors that may have been introduced during sample preparation or sequencing.

To use the pipeline, standard per-sample per-lane FASTQ files generated from postsequencing demultiplexing are required with a specific file naming format, as well as a metadata file. The pipeline can be deployed on either a local High Performance Compute (HPC) cluster or all major cloud providers, and is orchestrated by Nextflow, which leverages the executor of choice. The duet software installation and running guide provides instructions for the setup and configuration of your environment. This guide helps you understand the pipeline outputs and explains the format of the files generated.

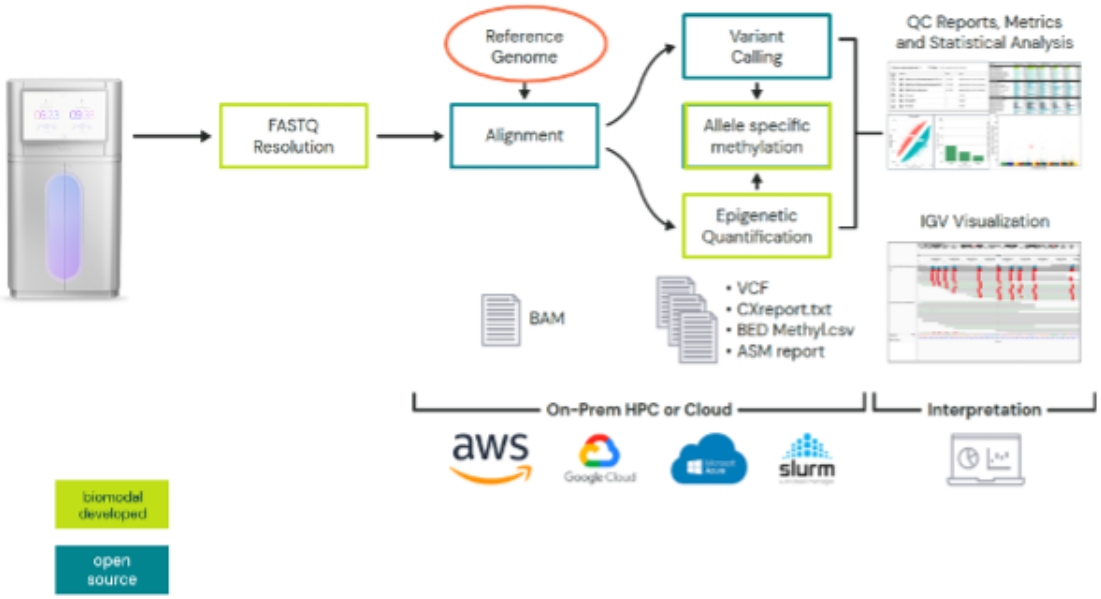
Resources

Access the following guides and additional resources on the biomodal documentation portal

Guide Name	Description
duet software installation and running guide	Provides instructions for setting up your environment and configuring the Command-Line Interface (CLI) tool that is used for running the pipeline.
biomodal Data Interpretation Guide (this document)	Provides analysis workflow and sequencing data analysis information for the biomodal pipeline.
Laboratory user guide: duet +modC	Provides instructions for preparing DNA libraries for sequencing using the biomodal duet multiomics solution +modC.
Laboratory user guide: duet evoC	Provides instructions for preparing DNA libraries for sequencing using the biomodal duet multiomics solution evoC.

Overview

Figure 1 presents a high-level overview of the data transformations that take place within the **biomodal duet multiomics solution bioinformatics pipeline**. Input FASTQ files are trimmed, resolved and aligned; duplicates are removed; variant calling and epigenetic quantification are performed. Optionally, variant call information can be combined with epigenetic quantification to call allele-specific methylation. Summary reports are generated at the conclusion of the pipeline.



4. Analysis Workflow

4.1. Input

The biomodal analysis pipeline requires 2 types of input files:

- FASTQ Read 1 and Read 2 files in pairs that are per-sample and per-lane.
- A sequencing run metadata file.

For FASTQ file and metadata file naming conventions, please see the CLI installation guide in Section 2.2 Input file requirements. Initial primary processing stages, including trimming, resolution, and alignment, are performed per-sample per-lane, and then after alignment, the lane-wise BAM files for each sample are merged.

4.2. Trimming, filtering and resolution

The trimming, filtering and resolution of reads takes place in a bespoke module called Prelude which converts the raw deaminated R1/R2 FASTQ file-pairs for a sample into a FASTQ file containing trimmed resolved single-end reads with epigenomic tags encoding the modifications present.

The trimming step identifies and removes the hairpin, which is expected to be present on the ends of any reads derived from constructs where the original DNA fragment had a length that was shorter than the read length. The trimming step also performs removal of other potential artefacts, such as poly-G tails that are longer than 8nt, reads with more than five Ns, and reads that are shorter than 15nt after the removal of all potential artefacts.

Resolution refers to the process of converting a pair of paired-end reads obtained from sequencing a deaminated, unfolded hairpin construct into a single-end read in an unambiguous 4-letter genomic alphabet annotated with epigenetic calls. This process pairs bases from each paired-end read, and for each pair of bases in the read-pair, a resolution rule is applied to determine whether the pair of bases should be resolved to:

1. An unmodified genetic base call (A, C, G, or T).
2. A genetic base call of C that is epigenetically modified:

- In duet +modC, this is referred to as modC and could be either methylcytosine (mC) or hydroxymethylcytosine (hmC).
- In duet evoC, methylcytosine (mC) and hydroxymethylcytosine (hmC) are differentiated in CpG contexts; in CpH contexts, modC is reported, which could be either methylcytosine (mC) or hydroxymethylcytosine (hmC). Note that the ability to differentiate mC and hmC in CpG contexts depends upon reading the succeeding G base in a CpG. There may be some cases where it is not possible to differentiate methylcytosine (mC) and hydroxymethylcytosine (hmC) because the succeeding G base is absent (e.g. if the C is the last base on a fragment, the last base on a read, or if the succeeding base is an N). In this case the modification would be reported as modC.

3. A suppressed error, represented as N in the resolved genomic sequence.

With 4 possible base calls from a position on read 1, and 4 possible base calls from the corresponding position on read 2, there are 16 possible pairings.

- In duet +modC, 5 are expected to be observed and 11 are not expected to be observed.
- In duet evoC, 6 are expected to be observed and 10 are not expected to be observed.

Pairings expected to be observed are referred to as plausible pairings; those not expected to be observed are referred to as implausible pairings. Suppressed errors derive from implausible pairings.

The resolution process applied to a pair of paired-end reads begins by aligning the reads naively such that the n^{th} base of read 1 is paired with the n^{th} base of read 2. The proportion of plausible pairings is assessed to determine whether the resolution can proceed. If the proportion of plausible pairings is low, it is likely that either a shift is needed to correctly align the two reads, or the sequenced construct is an unexpected artefact that should be discarded. For read pairs that do not align naively, a modified pairwise alignment is applied to determine whether the reads require a shift relative to one another in order to be resolved, or whether they need to be discarded.

For read pairs aligned naively or corrected via a pairwise alignment, resolution proceeds, and a resolved single-end genomic read with epigenetic annotations and error-suppressed bases is generated. Additional trimming removes potential artefacts from the ends of resolved reads.

Finally, any Cs in the last three bases of the resolved read are masked by being converted to Ns. This is to limit the potential impact that the end repair step of the duet assay can have on methylation calling at the 3′ end of DNA fragments. When a fragment of double-stranded input DNA to the assay features a 5′ overhang (sometimes referred to as a ‘jagged end’), the end repair step uses a polymerase to extend the 3′ end, but when this occurs the methylation pattern is lost from the repaired stretch.

4.3. Alignment

BWA-MEM2 is used for a standard alignment against a four-letter reference genome combined with the sequences of the spiked-in controls. Epigenetic calls are carried forward from the resolved FASTQ files into the aligned BAM files and feature in an ‘MM’ tag compliant with the definition of the ‘MM’ tag in the SAM file specification.

After alignment, the BAM files from each lane for a given sample are merged.

4.4. Duplicate marking

The marking of duplicates is performed using samtools markdup.

4.5. BAM file filtering and collection of stats

After the marking of duplicates, genome-aligned reads are separated from unaligned reads and from the reads aligning to each category of the control sequence. Controls are grouped into two categories:

- The ‘long controls’ are the methylated lambda and unmethylated pUC19 spike-ins.
- The ‘short controls’ are a set of 80bp oligonucleotide spike-ins.

The long controls are down-sampled to a maximum of 200x prior to the removal of duplicates and subsequent downstream processing. This is to ensure that the coverage on the controls does not differ significantly from the maximum coverage on the genome.

During this filtering step, secondary alignments and supplementary alignments and reads with a mapping quality of zero are also removed. Note that reads with a mapping quality of zero are reads that align equally-well to more than

one location in the reference genome.

A range of metrics associated with the aligned reads are calculated, such as coverage and bias metrics.

Duplicate removal is performed on the genome BAM file and on the long control BAM file but not on the short control BAM file.

The duplication rate in the genome BAM is reported in downstream reports.

Note that if the 'targeted' mode of the pipeline is invoked, then duplicates are marked but not removed. This is necessary to facilitate the accurate calculation of target-related metrics.

4.6. Epigenetic quantification

Epigenetic quantification commences after the filtering of the aligned lane-merged BAM files. By default, epigenetic status is quantified at CpGs and is performed on the genome-aligned reads as well as on the controls. Quantification is performed at sites that are identified as CpGs from the reference, so does not include CpG sites that are unique to the individual sample but absent from the reference genome.

Epigenetic quantification counts each type of epigenetic call at each CpG site in order to report per-CpG modification calling rates. It is possible to additionally quantify epigenetic status at CHG and CHH sites on the genome via an additional parameter.

Note that the possible base calls aligning to the C position of a reference CpG on a single read are:

- Cytosine with an associated methylation status (e.g.unmethylated, methylated, or hydroxymethylated)
- One of the other genetic bases, i.e. G, A, or T (which may represent a genuine single nucleotide variant or could be a miscalled base)
- N, i.e. a masked, erroneous, or suppressed call (which could arise from the sequencer or from the resolution algorithm)

The N, G, A, and T calls that align to a reference CpG account for the difference sometimes observed between the sum of coverage at the different methylated cytosine states and the total coverage at that CpG.

The output of the quantification step is:

- A methylation cytosine report, per sample, containing one row per CpG.
- A zarr file storing modification quantification information associated with each CpG and each sample in a compressed multi-dimensional array format, suitable for using with the biomodal modality Python library.

Additionally:

- In duet +modC, a bedMethyl file is generated per sample containing one row per CpG.
- In evoC, no bedMethyl file is generated by default. However, via an additional parameter in evoC, separate per-sample bedMethyl files can be generated for mC, hmC and undifferentiated modC (where a methylation call is made, but it cannot be determined whether it is mC or hmC)

These file formats are described in further detail below.

4.7. Spike-in quality control metrics

To assess the accuracy of detection of modified cytosines (modC) and unmodified cytosines (C) in duet +modC, and the accuracy of detection of methylcytosines (mC), hydroxymethycytosines (hmC), and unmodified cytosines (C) in duet evoC, spike-in controls are added to each reaction, analysed by the pipeline, and metrics are reported back to the user. These are also used to evaluate the genetic accuracy of the workflow. The spike-in controls

include the pUC19 unmethylated plasmid; a preparation of the lambda phage genome that has been artificially methylated at each CpG; and methylated, hydroxymethylated, and demethylated short oligonucleotides.

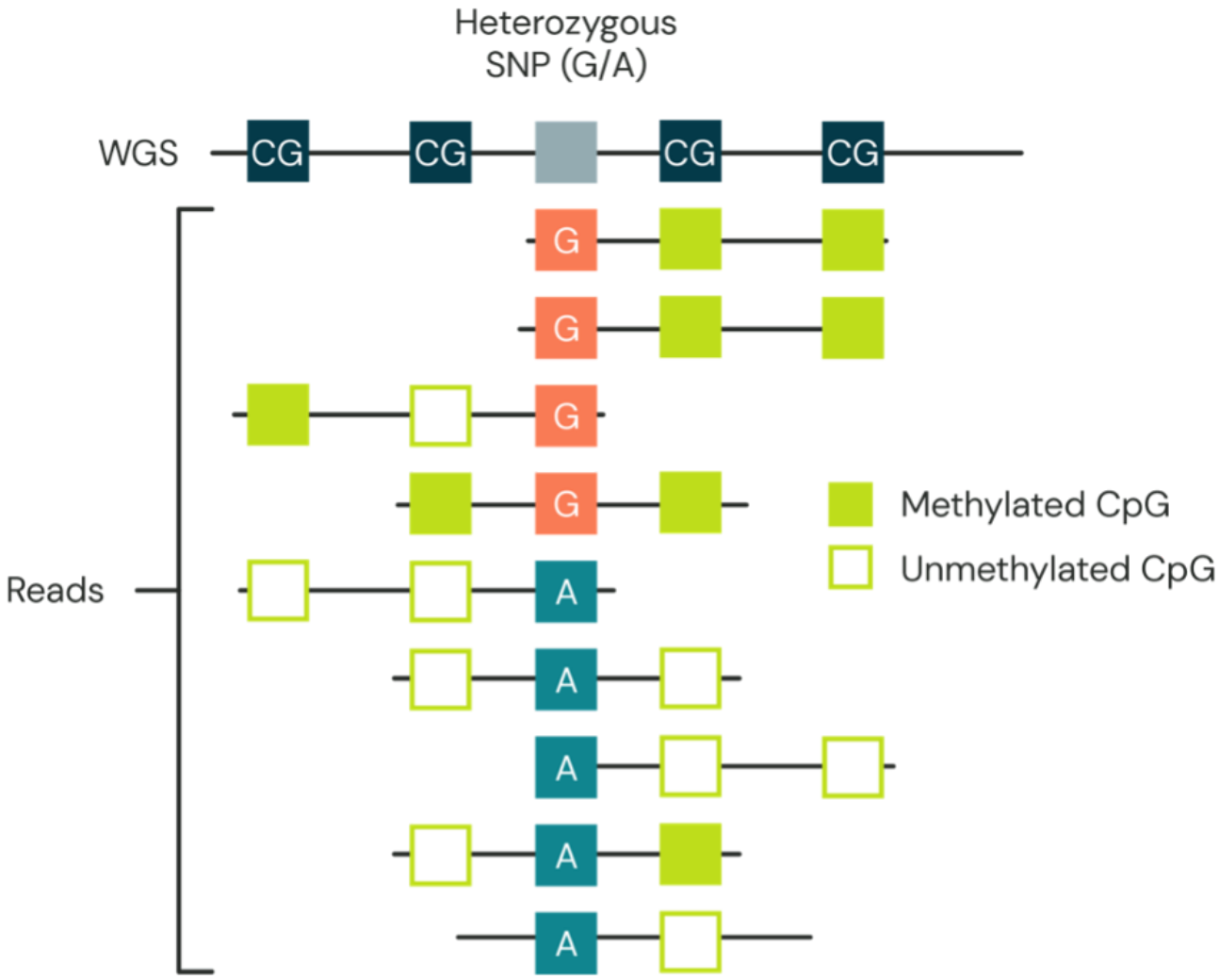
- In duet +mocC, the sensitivity of modC detection is calculated as the fraction of total versus expected modC calls in the lambda genome; specificity is calculated as the fraction of total versus expected C calls in pUC19.
- In duet evoC, the sensitivity of mC is calculated as the fraction of total versus expected mC calls in the lambda genome; the sensitivity of hmC is calculated as the fraction of total versus expected hmC calls on one of the short oligo controls; specificity is calculated as the fraction of total versus expected C calls in pUC19.

4.8. Variant calling

By default, the pipeline runs the Genome Analysis Toolkit (GATK) HaplotypeCaller for germline variant calling. By setting an additional parameter, somatic variant calling using Mutect2 can additionally be performed. When somatic variant calling is performed this way, Mutect2 is run in 'tumour-only' mode (i.e. with no paired normal sample) and the GATK-recommended FilterMutectCalls module is run afterwards to remove potential false positive somatic variant calls.

4.8.1. Variant associated methylation

The combination of accurate genomic and epigenomic data makes it possible to evaluate associations between variants and methylation. One such example is allele-specific methylation (ASM), which can be quantified by activating the ASM module in the pipeline. ASM is a biological event in which distinct differences in methylation patterns are observed across homologous chromosomes. Heterozygous single nucleotide variants (SNVs), which allow for separating reads by parental alleles, are critical for the identification of ASM. An example of such a heterozygous single nucleotide polymorphism is shown in Figure 2.



The ASM module takes a bam file and a standard VCF file as input, from which it extracts SNVs with a heterozygous genotype. For each heterozygous SNV, we quantify methylation that occurs at CpGs on reads associated with each allele. We then apply filtering; by default, only including loci that have 6 or more reads spanning the allele and at least one CpG, and where we observe more than a 30% difference in methylation level between the two alleles. We apply a statistical test to produce a p-value that quantifies how confident we are in the allele-specific methylation. The output of the ASM module is a bespoke csv file, referred to as the ASM file, with one row for every heterozygous SNV. The VCF and ASM file formats are described in section 5.

4.9. Report creation

The pipeline outputs data files and reports containing extensive sample information — see Output Files and Analysis Metrics sections.

4.10. Targeted

The biomodal duet pipeline includes a mode for processing data generated using targeted panels from pipeline version v1.1.2 onwards. The analysis pathways for processing targeted libraries are slightly different because the pipeline calculates targeted metrics and makes use of additional reference files describing the bait and target regions associated with the target panel being used.

Relevant reference files for the following target panels are available by default from pipeline version v1.1.2 onwards:

- Twist Alliance Pan-cancer Methylation Panel.
- Human Methylome Panel.

5. Output Files

Output from the biomodal pipeline organises data into the following top-level directory structure:

Directory	Contents
`reports`	Sample-level and multi-sample reports summarising information about the samples and controls.
`sample_outputs`	Primary data files generated by the pipeline (described in more detail below).
`controls`	BAM files and quantification files associated with the methylated lambda and unmethylated pUC19 controls. These small files are analogous to the BAM files and quantification files generated for your samples, and may be useful for familiarising yourself with the file formats. Note that there is an accompanying FASTA file for the controls in the reference file directory with the following name/location: `ss_ctrls/v24/ss-ctrls-long-v24.fa.gz`.
`diagnostics`	Secondary outputs from the pipeline including: • A parameters log recording the parameters that were used to execute the pipeline • More extensive metrics to support more detailed data investigations • The interim resolved FASTQ files that were passed into the aligner .

The biomodal pipeline produces the following **data files**:

File	File Name	Subdirectory
BAM	`{A}.genome.{B}.dedup.bam`	`sample_outputs/bams/`
	`{A}.genome.{B}.dedup.bam.bai`	
Germline VCF	`{A}.genome.{B}.dedup.vcf.gz`	`sample_outputs/variant_call_files/germline/`
	`{A}.genome.{B}.dedup.vcf.gz.tbi`	
Somatic VCF	`{A}.genome.{B}.dedup.somatic.vcf.gz`	`sample_outputs/variant_call_files/somatic/`
	`{A}.genome.{B}.dedup.somatic.vcf.gz.tbi`	

File	File Name	Subdirectory
Cytosine Report	{A}.genome.{B}.dedup.duet-{C}.CG_quant.CXreport.txt.gz	sample_outputs/modc_quantification
	{A}.genome.{B}.dedup.duet-{C}.CG_quant.CXreport.txt.gz.tbi	
ASM	{A}.asm.csv	sample_putputs/allele_specific_methylation/
Zarr Store	{ run_name }_{D}.zarrz	sample_outputs/zarr_store/

Where: {A} is the sample ID {B} is the genome tag, such as GRCh38Decoy. This is followed by _primary_assembly if you are working with a reference genome that features a subset of contigs that constitute the 'primary assembly'. For example, the GRCh38Decoy reference features some decoy contigs that are excluded from the 'primary assembly'. {C} is either modC or evoC depending on whether the data analysed is from the duet +modC kit or the duet evoC kit. {D} is CG, CHG, or CHH depending on the context and whether CHG/CHH calling has been requested at the time of launch. {run_name} is the parameter provided to the pipeline at the time of launch.

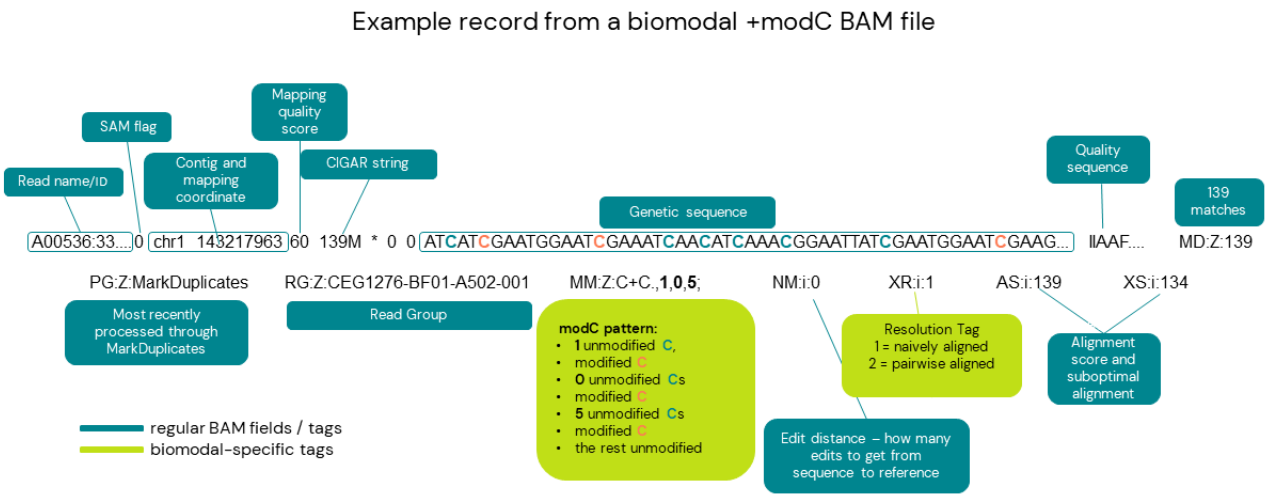
Note that:

- BAM, VCF, and Cytosine Report files have an accompanying index file. This is used by software that parses the files, such as IGV. Index files have the same name as the file they accompany, but with an additional extension.
- Somatic variant call files are only generated if requested at the time of launch.

Each file type is further described below.

5.1. BAM and CRAM files

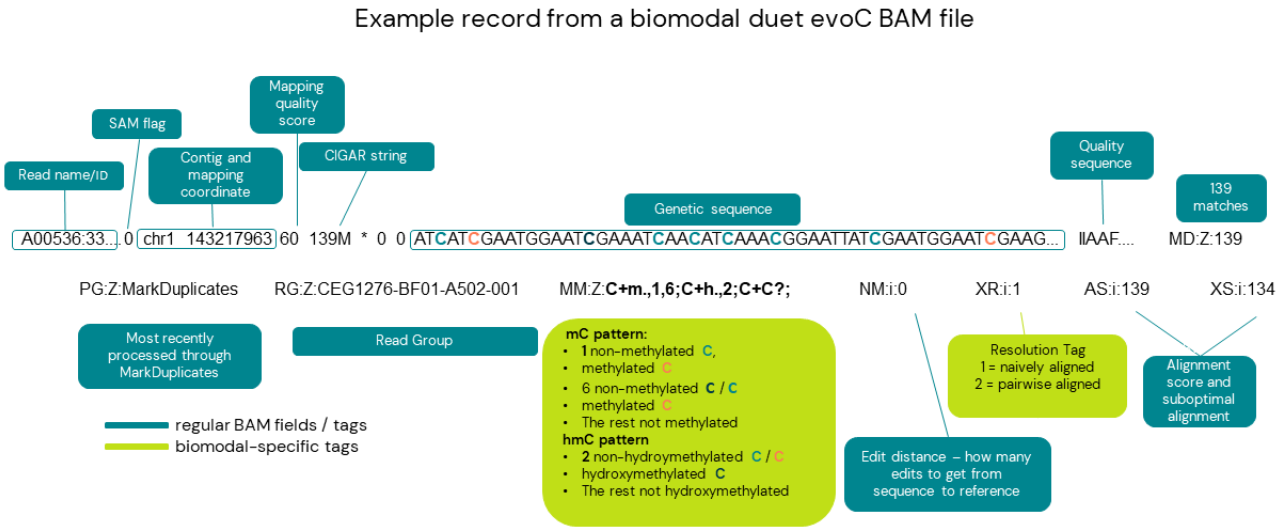
BAM files are in a binary format (and therefore compressed), but they can be converted into a SAM format to make them human readable. The diagram below in Figure 3 shows the SAM conversion of a BAM file:



Each read, as depicted above, is on a single line of the SAM file. The colour key indicates which fields are common features found in all SAM files, and which are unique to the biomodal methodology around recording information on methylation patterns.

Here, biomodal-unique features are represented by:

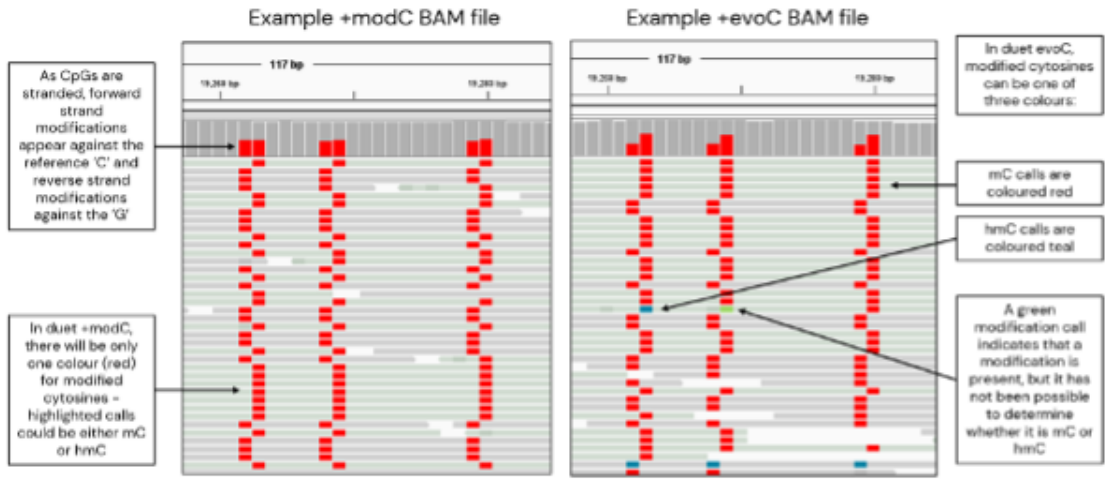
- MM tags – these record information about methylation and conform to a specification described in the SAM file. It is interpreted as follows: starting at the beginning, jump over the first number of Cs to arrive at a modC, and then jump over the second number of Cs to arrive at the next modC, and so on until the last number in the MM tag is reached.
- XR tag: this tag records whether each resolved read was resolved naively (XR:i:1) or whether the original R1 and R2 needed to be pairwise aligned before resolution (XR:i:2).



A duet evoC BAM file differs from a +modC BAM file only in the format of the MM tag. In a duet evoC BAM file, instead of a single list numbers there are three lists:

C+m.	This pattern describes which Cs have been called as methylated.
C+h.	This pattern describes which Cs have been called as hydroxymethylated.
C+C?	This pattern describes any Cs that have been called as modified, but where it has not been possible to determine whether the modification is mC or hmC. This occurs when a modification is called in a non CpG context, or when the dinucleotide context is unknown (for instance when the C is the last base on a read or the succeeding base is an N)

BAM files are a well-established format for storing sequence alignments. They can be loaded into IGV (the Integrated Genomics Viewer) for visualization. Methylation status can be visualized in IGV by using the right-click menu and toggling the 'color alignments by --> base modification' setting.



5.1.1 CRAM files

Compressed Reference-oriented Alignment Map (CRAM, defined here in detail) is a file format that offers greater compression compared to BAM for storing sequence alignments. This reduces disk space usage and storage costs. By default, the biomodal pipeline outputs sequence alignments as BAM files. However, if analysed with an additional parameter, the pipeline will output genome alignment files as CRAM files.

For example, storing a single alignment file with 30X mean coverage has a disk space of approximatley 60 GB when stored as a BAM, but this is reduced to 35 GB when stored as a CRAM, representing more than a 42% reduction in size (15% reduction in disk space of the pipeline)

	BAM	CRAM
Sequence Alignment	60 GB	35 GB
Whole Pipeline	160 GB	135 GB

CRAM files use reference-based compression, meaning that sequence data is only stored when it differs from the reference genome. Consequently, the reference FASTA file is required to read the CRAM file. When the pipeline is run with the parameter set to generate CRAM files, the output subdirectory `sample_outputs/crams` will contain both the CRAM files and the reference FASTA file required to interpret them.

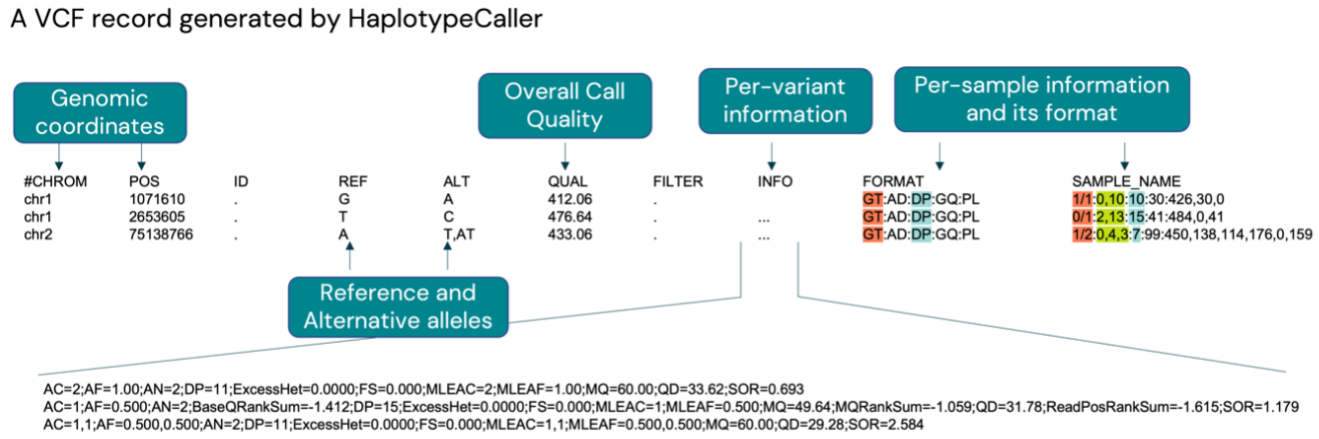
Most popular downstream tools for analysing alignment files, such as samtools, pysam, GATK, and IGV, support CRAM files when provided with their accompanying reference FASTA.

If needed, CRAM files can be easily converted back to BAM files using `samtools view`:

```
samtools faidx reference.fa
samtools view -@ 6 -T reference.fa -b -o out.bam  in.cram
```

5.2. VCF files

These are well-established bioinformatics file formats (defined here in detail) and contain a list of SNVs (single nucleotide variants) and INDELs (insertions/deletions) that have been found. For each variant, it lists the genomic coordinates (chromosome/contig and position), reference and alternative base observed, overall quality score, and both per-variant and per-sample information, as illustrated in Figure 6.



Per-variant information is contained in the “INFO” field, where the key and the data can be found in the format `KEY=data` (e.g., `AF=1.00`), and the definition of keys can be found in the header of the VCF file (e.g. `##INFO=`

`<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">`)

Per-sample information contains information specific to the sample in question (a VCF file can contain more than one sample, for example, in the case of joint variant calling). The keys are identified by the FORMAT field and also defined in the header. In this example from HaplotypeCaller, there are 5 pieces of information reported per sample: `GT:AD:DP:GQ:PL` (which correspond to Genotype, Allele Depth, Depth, Genotype Quality, and Phred Scaled Likelihood). Other variant callers, such as Mutect2 will output different information. The data is recorded in the column corresponding to each sample under their `SAMPLE_NAME` and in the same order:

- The Genotype field (GT) reports the predicted genotype in the format n/n, where n refers to the allele number, with 0 being the REF allele, 1 the first ALT allele, 2 the second ALT allele, and so on. In the example above, the first variant is homozygous ALT (1/1 corresponding to A/A); the second variant is heterozygous with one allele corresponding to REF and one to ALT (0/1 or T/C); the third variant is a heterozygous with neither allele corresponding to the REF allele (1/2 or T/AT).
- The Allele Depth field (AD) reports the number of reads supporting the REF and ALT allele (in this example, 0 reads support the REF allele and 4 reads support the first ALT allele, and 3 reads support the second ALT allele for the third variant).
- The Depth field (DP) reports the total sequencing depth at the indicated position.
- The Genotype Quality (GQ) field reports the overall quality of the genotype call in that particular sample.
- The Phred Scaled Likelihood (PL) reports how much less likely each genotype is compared to the genotype that has been called for all the possible genotypes that can be called at that position. In the case of a biallelic site (e.g., the first row in the example), these are REF/REF, REF/ALT, and ALT/ALT, with scores of 426, 30, and 0, respectively. These indicate that the most likely genotype is ALT/ALT, followed by REF/ALT 1000 times less likely (Phred=30) and REF/REF 2.5*1043 times less likely, (Phred=426).

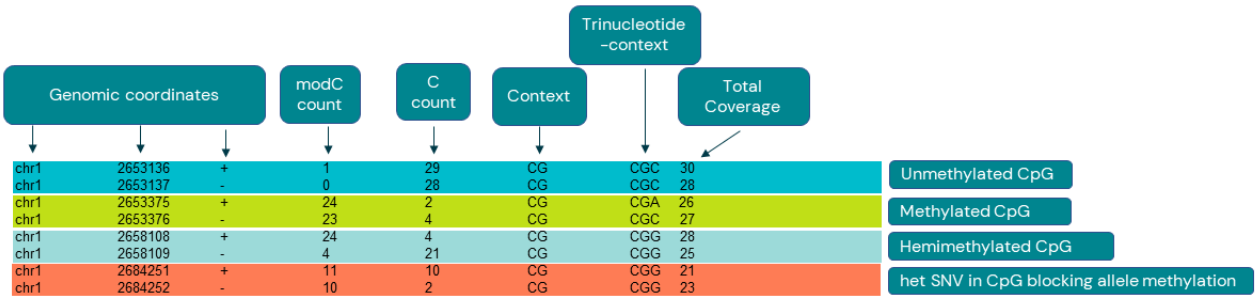
VCF files are a well-established format for storing variant calls. They can be loaded into IGV (the Integrated Genomics Viewer) for visualization.

5.3. Duet cytosine report

5.3.1 Duet cytosine report +modC quantification file

This file summarises the state of each cytosine by reporting the number of reads supporting the modified and unmodified state of each cytosine in the genome covered by sequencing reads, as seen in Figure 7.

Example records from the Cytosine Report +modC quantification file



By default, only cytosines in a reference CpG context are reported, although reporting cytosines in reference CHH/CHG contexts is also possible. The file contains the following information:

- Genomic coordinates: including information about the contig, position, and strand of the cytosine on which information is provided. Note that in most cases, this means that two records are provided for each CpG reporting separate information for the C in the forward and reverse strand.
- modC/C counts: the number of reads supporting a modified or unmodified cytosine at each position.
- Context and Trinucleotide context: the sequence context around the cytosine. By default, only cytosines in CG context are reported. Trinucleotide context refers to C plus the two following bases in 5′-3′ direction relative to the strand on which the cytosine is located. This context comes from the reference file, not from the aligned reads.
- Total coverage: the number of reads covering the position including non-C base calls.

The figure above reports examples of four CpGs with different states. The first CpG is fully unmodified (modified fraction: 1/58 ≈ 1.7%); the second is fully modified (modified fraction: 47/53 ≈ 88.6%); the third CpG is hemi-modified (modified fraction: 28/53 ≈ 52.8%, with the modification concentrated on the forward strand (24/28 ≈ 85.7%) rather than the reverse strand (4/25 = 16%); finally, the fourth CpG is also partially modified, but in this case, one strand is ~50% modified (11/21), while the other strand is also 50% modified (10/23), but the majority of unmodified reads are unmodified because they do not have a cytosine. This is consistent with a biallelic site in which one allele (CG/GC) is fully modified while the other allele is fully unmodified because it is not in a CpG context - i.e. it is CH/GD (where H and D are IUPAC disambiguation codes).

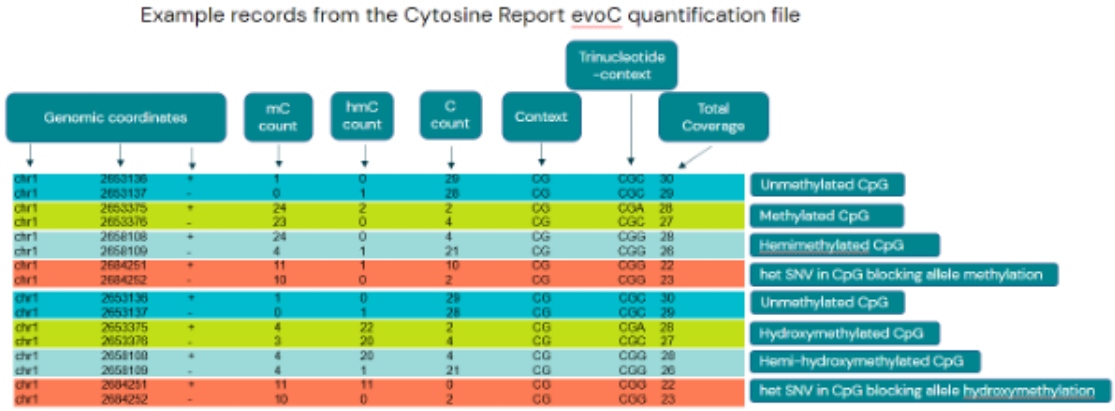
Note that the total coverage column includes N bases and any non-C genetic bases that align to the CpG; therefore the value in the coverage column may be greater than the sum of coverage in the modC and C columns.

The +modC Cytosine Report is a plain text format data file that can be loaded into a data frame in R or Python, or another programming language. It is also compatible with a number of existing libraries used for methylation analysis and visualisation, such as:

- The methylKit R library.
- The RnBeads R library.

5.3.2 Duet cytosine report duet evoC quantification file

In duet evoC, an additional column for hmC is added to the Cytosine Report, as shown in Figure 8.



Note that in the duet evoC cytosine report, undifferentiated modC calls (where a modification is present but it cannot be determined whether it is mC or hmC) feature in the total coverage column, but not in either the mC or hmC columns.

The addition of the hmC column is not immediately compatible with the methylKit R library; the dataframe can be loaded and adjusted as necessary. For example, the following code converts the duet evoC cytosine report into separate methylKit-compatible mC and hmC cytosine reports and then plots the mC and hmC distributions:

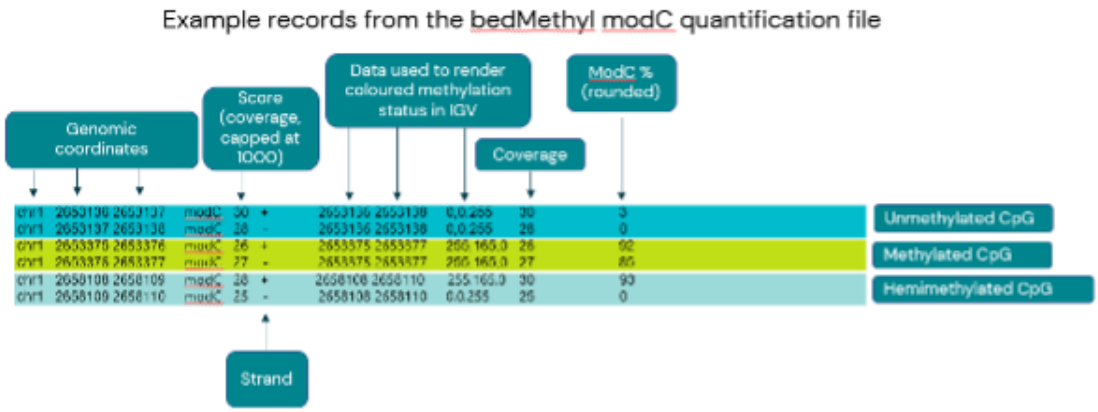
```
library(data.table)
library(tidyverse)
library(methylKit)

#  5mC
fread("~/CEG1485-EL01-D1115-005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport.txt.gz",
           col.names =c('contig', 'coordinate', 'strand', 'mC', 'hmC', 'C',
'context', 'trinucleotide', 'coverage') )%>%
##       Chromosome Position    Strand   Count methylated. Count unmethylated.
C-context  Trinucleotide context
  select(contig,    coordinate, strand,  mC,               C,
context,    trinucleotide)%>%
  fwrite("~/CEG1485-EL01-D1115-
005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport_reformatted_mC.txt.gz",quote = FALSE,col.names =
FALSE,row.names = FALSE,sep="\t")

#.  5hmC
fread("~/CEG1485-EL01-D1115-005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport.txt.gz",
      col.names =c('contig', 'coordinate', 'strand', 'mC', 'hmC', 'C',
'context', 'trinucleotide', 'coverage') )%>%
  ##       Chromosome Position    Strand   Count methylated. Count unmethylated.
C-context  Trinucleotide context
  select(contig,    coordinate, strand,  hmC,               C,
context,    trinucleotide)%>%
  fwrite("~/CEG1485-EL01-D1115-
005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport_reformatted_hmC.txt.gz",quote = FALSE,col.names =
FALSE,row.names = FALSE,sep="\t")
```

```
##Read the reformatted files in methylkit
myobj<-methRead(list('~/CEG1485-EL01-D1115-
005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport_reformatted_mC.txt.gz',
                     '~/CEG1485-EL01-D1115-
005.genome.GRCh38Decoy_primary_assembly.dedup.duet-
evoC.CG_quant.CXreport_reformatted_hmC.txt.gz'),
         sample.id=list("mC","hmC"),
         assembly="hg38",
         treatment=c(0,0),
         context="CpG",
         mincov = 10,
         pipeline='bismarkCytosineReport'
)
#Make the plots
getMethylationStats(myobj[[1]],plot=TRUE,both.strands=FALSE)
getMethylationStats(myobj[[2]],plot=TRUE,both.strands=FALSE)
```

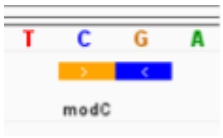
5.4. Duet BedMethyl modC quantification file



By default, only cytosines in a CpG context are reported, although reporting cytosines in CHH/CHG contexts is also possible. The file contains the following information:

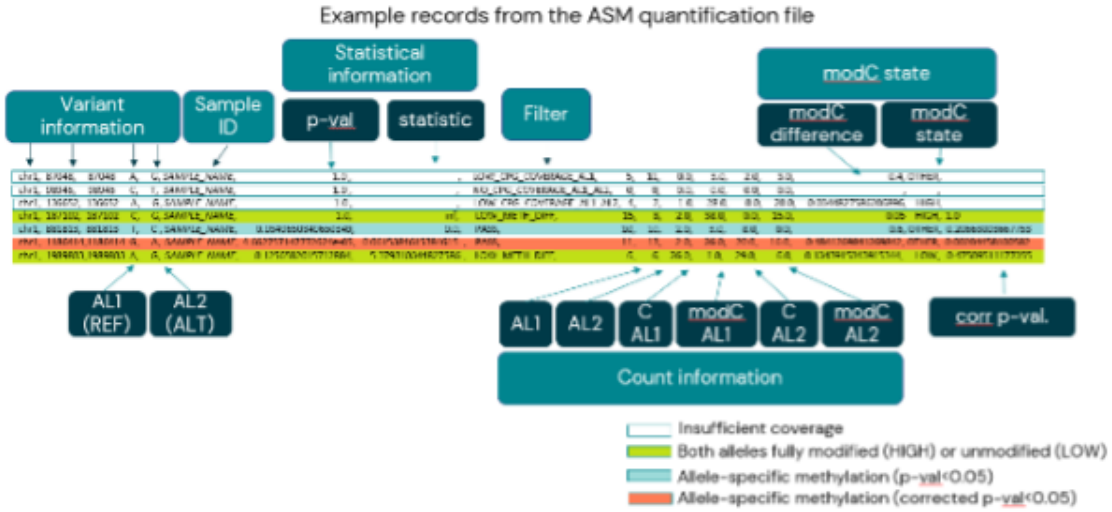
- Genomic coordinates: including information about the contig, position, and strand of the cytosine on which information is provided. Note that in most cases, this means that two records are provided for each CpG reporting separate information for the C in the forward and reverse strand.
- Coverage and coverage score: the number of reads covering the CpG, including a score that is the coverage, but capped at 1000.
- modC %: The percentage of cytosines at the given location that were reported as modC. This percentage is rounded to a whole number.
- modC status colour encoding: Data used to render coloured modC status markers when viewed in IGV, as shown in the following table:

	Methylated CpG.

	Unmethylated CpG.
	Hemimethylated CpG Forward Strand.
	Hemimethylated CpG Reverse Strand.

5.5. Allele-specific methylation (ASM) file format

The ASM file format created by biomodal is depicted in Figure 10. It reports for each heterozygous variant present in the sample the following information:



- **Variant information**, including genomic coordinates (chromosome and position), as well as allele1 and allele2, which correspond to REF and ALT in a VCF file (see the relevant section).
- **Sample name**
- **Filter**: this field is used to exclude sites with insufficient coverage (less than 6 reads covering each allele) from subsequent analyses. It can take any of the values in the table below - normally only sites with sufficient coverage are analysed ("LOW_METH_DIFF" or "PASS"):

Category	Description
PASS	This site meets all requirements and is available for ASM evaluation
LOW_METH_DIFF	This site meets the read depth requirement and is available for ASM evaluation, BUT the methylation difference between the two alleles is <0.3, which is often a prerequisite for ASM calling.
{NO,LOW}_CPG_COVERAGE_{AL1,AL2}	This site lacks any reads (NO) or sufficient reads (LOW) covering one or more CpGs (default requirement is min 6 reads at at least 1 CpG).

Category	Description
`NO_READS`	This site lacks any reads with a sufficiently high mapping quality (default requirement is Q30)

- **Count information**: these fields report the number of reads covering each allele, as well as the number of modified and unmodified CpGs associated with each allele. The latter is obtained by summing the number of modC and C in the CpG context in all the reads supporting each allele. For example, in the first record in the figure, a total of 5 reads are associated with the REF allele (A), and 11 reads are associated with the ALT allele (G; therefore, the site is filtered as AL1_LOW_READ_COUNT). In the 5 reads supporting allele A, a total of 3 CpGs contain unmodified Cs, and no CpG contains a modified C; in the 11 reads supporting allele G, a total of 2 CpGs contain unmodified Cs and 3 CpGs contain modified Cs. Note that the total number of observed CpGs will differ from the total number of reads because reads may have different numbers of CpGs on them.
- **modC state information** includes an estimate of the absolute difference in the modification levels of the two alleles, calculated as the difference between the mean methylation across all observed CpGs on reads associated with each allele. For example, in the fourth SNV in Figure 6, 38/40 = 95% of CpGs on allele 1 are modified, while 15/15 = 100% of CpGs on allele 2 are modified, giving a methylation difference of 0.05 or 5% (which is less than 30%, and for this reason this site is reported as LOW_METH_DIFF, in the filter field). The second field uses this information to categorise each SNV in a modC state: if both alleles have more than 80% modification, this will be 'HIGH', and if both of them have less than 20%, this will be 'LOW' (green-highlighted samples in the figure) ; if neither condition applies the site will be categorized as 'OTHER' (or it will be empty if there are insufficient reads to generate any data). A site having 'OTHER' modC state is a pre-requisite for it being considered a possible ASM site.
- **Statistical information** is reported on the last three fields: A statistic is calculated and reported in the statistic field (normally, this is the odds ratio of the Fisher's Exact Test), and a corresponding p-value calculated for that statistic, measuring the confidence of the ASM call. This value is then corrected for multiple hypothesis testing (Benjamini/Hochberg correction, considering for each sample only the PASS sites).

Users have different strategies to identify sites with allele specific methylation with a certain degree of confidence from this dataset.

- Filtering the file for p-val < 0.01 will result in a list of all possible single sites of allele-specific methylation with a non-negligible false discovery rate. We have observed that this strategy works well if you want to track large-scale movements of het variants to/from the ASM state between samples sharing the same set of variants (e.g., the same cell line treated with different drugs/conditions/etc).
- Filtering the file for genomic regions carrying a certain number of ASM sites (identified as p-val < 0.01) will result in higher specificity, greatly improving the ability to discriminate between real ASM loci and false discoveries at the cost of reducing sensitivity and resolution at the genomic level.
- Filtering the file for genomic regions carrying a certain number of ASM sites, identified as log10(corrected_p-val) < -10, will only identify regions with very strong ASM (e.g., imprinted regions) and reduce false discoveries the most.

In all cases, it is important to remember that higher genome-wide coverage will generally result in lower p-values and, therefore, more ASM calls, so when comparing two samples, it is best to aim for comparable coverage.

The ASM file format is a bespoke plain-text format. It can be loaded as a dataframe into R or Python for bespoke analysis. It can be filtered using standard Unix commands, such as `grep`.

5.6. Zarr store

The Zarr store is a multi-sample, multi-dimensional, compressed, indexed data store. This data store contains equivalent information to the quantification files, but wraps them into an extremely efficient data format unlocking accelerated downstream analysis. It can be used in conjunction with the biomodal modality Python package to efficiently load, query, analyse, and present modification data.

5.7. Resolved reads FASTQ file

The resolved FASTQ files are single-end reads in a regular FASTQ file format including the following features:

- An XR tag, as described above in the BAM Files section, indicating whether the original R1/R2 read pair were naively aligned or pairwise aligned.
- An MM tag, as described above in the BAM Files section, indicating the methylation calls made.
- The resolved genetic sequence determined from the original R1/R2 sequences.
- The resolved quality sequence determined using sequencer-specific and read-length-specific empirical q-tables.

6. Analysis Metrics

Aggregate summary metrics report

An aggregate summary report of important metrics for each stage of the pipeline workflow is available in an Excel format in the `reports/summary_reports` subdirectory. Each column of the report presents the metrics for one sample, with the sample noted in the column heading. Additionally, there is a csv file containing this data. The csv file contains one *row* per sample. There is also a csv file with the suffix `Metrics_Definitions.csv` that provides a mapping of the metric names in the Excel file to their corresponding field names in the csv file and to their descriptions.

The metric names in the left-column of the Excel file feature a ‘Note’ that can be shown by right-clicking and selecting ‘Show/Hide Note’. The ‘Note’ provides a description of the metric.

Some sections of the Excel report differ between the +modC product and the duet evoC product. Where this is the case, the sections are described separately.

The Excel report features the following metrics per sample grouped into sections. All values provided are guidelines developed on the Illumina NovaSeq6000 sequencing platform.

6.1. Modified cytosine accuracy: control DNA (+modC)

In the +modC product, the following metrics are reported to characterise the evaluation of epigenetic accuracy using the methylated lambda and unmethylated pUC19 controls.

Field Name	Description

Field Name	Description
modC sensitivity on fully methylated lambda control	Sensitivity for measuring methylated CpGs calculated from a fully methylated lambda control. A value >95% would be considered satisfactory, and a value >98% would be considered good, >98.5% would be considered excellent.
modC specificity on fully unmethylated pUC19 control	Specificity for measuring unmodified CpGs calculated from a fully non-methylated pUC19 control. A value > 99% would be considered satisfactory, >99.8% would be considered good, > 99.9% would be considered excellent.
Non-C calls at CpG sites on fully methylated lambda control	Percentage of positions aligning to a CpG on the lambda reference where the called base is not a C. Such bases could be A, G, T, or N. The N bases will include cases where a sequencing error has been suppressed during resolution and cases where C's in the last three bases of a resolved read have been masked.
Non-C calls at CpG sites on fully unmethylated pUC19 control	Percentage of positions aligning to a CpG on the pUC19 reference where the called base is not a C. Such bases could be A, G, T, or N. The N bases will include cases where a sequencing error has been suppressed during resolution and cases where C's in the last three bases of a resolved read have been masked.

6.2. Modified cytosine accuracy: control DNA (duet evoC)

In the duet evoC product, the following metrics are reported to characterise the evaluation of epigenetic accuracy using the methylated lambda and unmethylated pUC19 controls.

Field Name	Description
Methylated lambda modC sensitivity	Sensitivity for measuring modC at mC sites calculated from a fully methylated lambda control. modC refers to a call of mC, hmC or undifferentiated modC.
Non-methylated pUC19 control modC specificity	Specificity for measuring modC at C sites calculated from an unmethylated pUC19 control. modC refers to a call of mC, hmC or undifferentiated modC.
Methylated lambda control mC sensitivity	Sensitivity for measuring mC calculated from a fully methylated lambda control.
Methylated lambda control hmC specificity	Specificity for measuring hmC calculated from a fully methylated lambda control.
Methylated lambda control mC precision	Precision for measuring mC calculated from a fully methylated lambda control.

Field Name	Description
Non-methylated pUC19 mC specificity	Specificity for measuring mC calculated from an unmethylated pUC19 control.
Non-methylated pUC19 hmC specificity	Specificity for measuring hmC calculated from an unmethylated pUC19 control.
Non-C calls at CpG sites on fully methylated lambda control	Percentage of positions aligning to a CpG on the lambda reference where the called base is not a C. Such bases could be A, G, T, or N. The N bases will include cases where a sequencing error has been suppressed during resolution and cases where Cs in the last three bases of a resolved read have been masked.
Non-C calls at CpG sites on fully unmethylated pUC19 control	Percentage of positions aligning to a CpG on the pUC19 reference where the called base is not a C. Such bases could be A, G, T, or N. The N bases will include cases where a sequencing error has been suppressed during resolution and cases where Cs in the last three bases of a resolved read have been masked.

6.2.1 SQ2hmC (Hydroxymethylated Oligo) Control Accuracy (duet evoC)

Additionally, in duet evoC, the following metric is presented as an indication of hmC sensitivity:

Field Name	Description
Percent hmC called as hmC on 80bp SQ2hmC mixed C/hmC short control	The percentage of hmC sites correctly called as hmC on an 80bp synthetic oligo with a variety of different C and hmC states at CpGs

6.3. Genetic accuracy: control DNA

Metrics associated with the evaluation of genetic accuracy using the methylated lambda control.

Field Name	Description
Genetic accuracy lambda control	Overall genetic accuracy from the lambda-aligned reads as a percentage calculated relative to a lambda truth set. A value > 99.90% would be considered satisfactory, > 99.97% would be considered excellent.
Genetic accuracy lambda control Q-score	Overall genetic accuracy from the lambda-aligned reads as a Q-score calculated relative to a lambda truth set.
Genetic accuracy lambda control A	Overall genetic accuracy from the lambda-aligned reads of A bases as a percentage calculated relative to a lambda truth set.
Genetic accuracy lambda control C	Overall genetic accuracy from the lambda-aligned reads of C bases as a percentage calculated relative to a lambda truth set.
Genetic accuracy lambda control G	Overall genetic accuracy from the lambda-aligned reads of G bases as a percentage calculated relative to a lambda truth set.
Genetic accuracy lambda control T	Overall genetic accuracy from the lambda-aligned reads of T bases as a percentage calculated relative to a lambda truth set.

Field Name	Description
Genetic accuracy lambda control Q-score A	Overall genetic accuracy from the lambda-aligned reads of A bases as a Q-score calculated relative to a lambda truth set.
Genetic accuracy lambda control Q-score C	Overall genetic accuracy from the lambda-aligned reads of C bases as a Q-score calculated relative to a lambda truth set.
Genetic accuracy lambda control Q-score G	Overall genetic accuracy from the lambda-aligned reads of G bases as a Q-score calculated relative to a lambda truth set.
Genetic accuracy lambda control Q-score T	Overall genetic accuracy from the lambda-aligned reads of T bases as a Q-score calculated relative to a lambda truth set.

6.4. Quantification of modified cytosines

Metrics are reported to summarise genome-wide CpG methylation rates in the autosomes. The allosomes are excluded from this calculation to ensure comparability of this rate between samples of different sex. Additionally, the genome-wide rate of unmodified C in the mitochondria is reported. The acts as an additional control because methylation is expected to be extremely rare or entirely absent in the mitochondria. Base calls of G, A, T, and N at sites that align to a reference CpG are also excluded from these calculations.

6.4.1 Quantification of modified cytosines; duet +modC

In the +modC product, the following metrics are reported:

Field Name	Description
Mitochondrial genome rate of C at CpG	Rate of observing an unmethylated C at CpG sites on the mitochondrial genome.
Autosomal chromosomes rate of C at CpG	Rate of observing an unmethylated C at CpG sites on the autosomes.
Autosomal chromosomes rate of modC at CpG	Rate of observing an modified (mC or hmC) C at CpG sites on the autosomes.

The following metrics will be present only if the pipeline has been run with the CHG/CHH quantification mode enabled:

Field Name	Description
Mitochondrial chromosome rate of C at CHG	Rate of observing an unmethylated C at CHG sites on the mitochondrial chromosome.
Autosomal chromosomes rate of C at CHG	Rate of observing an unmethylated C at CHG sites on the autosomes.
Autosomal chromosomes rate of modC at CHG	Rate of observing an modified C (mC or hmC) at CHG sites on the autosomes.
Mitochondrial chromosome rate of C at CHH	Rate of observing an unmethylated C at CHH sites on the mitochondrial chromosome.
Autosomal chromosomes rate of C at CHH	The rate of observing an unmethylated C at CHH sites on the autosomes.

Field Name	Description
Autosomal chromosomes rate of modC at CHH	The rate of observing an modified C (mC or hmC) at CHH sites on the autosomes.

6.4.2 Quantification of modified cytosines; duet evoC

In the evoC product, the following metrics are reported:

Field Name	Description
Mitochondrial genome rate of C at CpG	Rate of observing an unmethylated C at CpG sites on the mitochondrial genome.
Autosomal chromosomes rate of C at CpG	Rate of observing an unmethylated C at CpG sites on the autosomes.
Autosomal chromosomes rate of modC at CpG	Rate of observing an modified (mC or hmC) C at CpG sites on the autosomes.
Autosomal chromosomes rate of mC at CpG	Rate of observing methylated C at CpG sites on the autosomes.
Autosomal chromosomes rate of hmC at CpG	Rate of observing hydroxymethylated C at CpG sites on the autosomes.

The following metrics will be present only if the pipeline has been run with the CHG/CHH quantification mode enabled:

Field Name	Description
Mitochondrial chromosome rate of C at CHG	Rate of observing an unmethylated C at CHG sites on the mitochondrial chromosome.
Autosomal chromosomes rate of C at CHG	Rate of observing an unmethylated C at CHG sites on the autosomes.
Autosomal chromosomes rate of modC at CHG	Rate of observing an modified C (mC or hmC) at CHG sites on the autosomes.
Autosomal chromosomes rate of mC at CHG	Rate of observing methylated C at CHG sites on the autosomes.
Autosomal chromosomes rate of hmC at CHG	Rate of observing hydroxymethylated C at CHG sites on the autosomes.
Mitochondrial chromosome rate of C at CHH	Rate of observing an unmethylated C at CHH sites on the mitochondrial chromosome.
Autosomal chromosomes rate of C at CHH	The rate of observing an unmethylated C at CHH sites on the autosomes.
Autosomal chromosomes rate of modC at CHH	The rate of observing an modified C (mC or hmC) at CHH sites on the autosomes.
Autosomal chromosomes rate of mC at CHH	The rate of observing methylated C at CHH sites on the autosomes.

Field Name	Description
Autosomal chromosomes rate of hmC at CHH	The rate of observing hydroxymethylated C at CHH sites on the autosomes.

6.5. Genome duplication coverage

Metrics associated with the alignment of reads to the genome and the identification and removal of duplicates.

Field Name	Description
Genome-mapped reads (including duplicates)	The total number of genome-aligned reads including duplicates.
Genome-mapped read duplicates	The number of resolved genome-aligned reads identified and removed as potential duplicates.
Genome-mapped duplication rate	The duplication rate in the genome-aligned reads.
Genome deduplicated reads	The total number of reads in the deduplicated genome primary assembly-aligned BAM file.
Genome deduplicated bases	The total number of bases in the deduplicated genome primary assembly-aligned BAM file.
Percent of input bases aligned to genome primary assembly	The percentage of deduplicated primary genome assembly aligned bases, not including soft-clipped bases.
Genome mean mapped bases per read	The average number of bases in trimmed, resolved, genome aligned and deduplicated reads, excluding soft-clipped bases.
Genome reads mean quality	The average quality of trimmed, resolved, genome-aligned, deduplicated reads (excluding soft-clipped bases).
Genome mean MAPQ	The average mapping quality of trimmed, resolved, genome-aligned, deduplicated reads.
GC bias global error	A measure of GC bias. Lower values indicate less bias towards or away from GC-rich regions. The metric is not directional, so a high value could be associated with bias towards or away from GC-rich regions.
Mean coverage	The mean genome-wide coverage.
Genome percentage no coverage	Percentage of the genome with no coverage.
Genome percentage 1x	Percentage of the genome covered at 1X or above.

Field Name	Description
Genome percentage 2x	Percentage of the genome covered at 2X or above.
Genome percentage 5x	Percentage of the genome covered at 5X or above.
Genome percentage 10x	Percentage of the genome covered at 10X or above.
Genome percentage 25x	Percentage of the genome covered at 25X or above.
Genome percentage 30x	Percentage of the genome covered at 30X or above.
CpG to genome-wide coverage ratio	Ratio of mean coverage at CpGs to mean coverage genome-wide. Because CpGs are stranded, a 'perfect' value for the metric would be 0.5. Values > 0.5 indicate bias towards CpG; values < 0.5 indicate bias away from CpGs
TSS to non-TSS coverage ratio	Ratio of mean coverage near transcription start sites (TSS regions) to mean coverage at regions not near transcription start sites. A 'perfect' value for the metric would be 1.0. Values > 1.0 indicate bias towards TSS regions; values < 1.0 indicate bias away from TSS regions.

6.6. Read pair resolution (Prelude)

Metrics associated with the transformation of read-pairs in a deaminated alphabet into resolved single-end 4-letter genomic reads with epigenetic annotations.

Field Name	Description
Reads after trimming and quality filtering	Total number of read-pairs (R1/R2 read-pairs) that remain after trimming and quality filtering.
Bases after trimming and quality filtering	Total number of bases that remain after trimming and quality filtering.
Reads that resolve naively	Total number of reads that can be resolved into genetic and epigenetic sequences without prior pairwise alignment.
Percentage of reads that resolve naively	Percentage of trimmed reads that can be resolved into genetic and epigenetic sequences without prior pairwise alignment.
Reads to rescue via pairwise alignment	Total number of reads for which a pairwise alignment will be attempted in order to resolve into genetic and epigenetic sequences.
Percentage of reads to rescue via pairwise alignment	Percentage of trimmed reads for which a pairwise alignment will be attempted in order to resolve into genetic and epigenetic sequences.
Reads rescued via pairwise alignment	Number of reads that were able to be resolved into genetic and epigenetic sequences after a pairwise alignment.

Field Name	Description
Percentage of reads rescued via pairwise alignment	Percentage of reads for which a pairwise alignment was attempted that were able to be resolved into genetic and epigenetic sequences after a pairwise alignment.
Total resolved reads	Total number of resolved reads usable for downstream analysis (made up of those that resolved naively and those that were rescued).
Percentage of trimmed reads that resolve	Percentage of trimmed, quality-filtered reads usable for downstream analysis (made up of those that resolved naively and those that were rescued)
Percentage of input reads that resolve	The fraction of the total input read pairs in the initial input FASTQ files that remains as resolved reads after processing through the resolution algorithm.
Total discarded reads	Total number of trimmed, quality-filtered reads discarded as not representing the expected construct / format to resolve into genetic / epigenetic sequences.
Percentage of trimmed reads that are discarded	Percentage of trimmed, quality-filtered reads discarded as not representing the expected construct / format to resolve into genetic / epigenetic sequences.
Percentage of trimmed bases that resolve	Percentage of trimmed, quality-filtered bases available for downstream analysis after resolution and further trimming.
Percentage of input bases that resolve	The fraction of the total input base-pairs in the initial input FASTQ files that remains as resolved bases after processing through the resolution algorithm.

6.7 Trimming (Prelude)

Metrics associated with the trimming step which removes the hairpin sequence from the reads and additionally performs some filtering, for instance removal of very short reads.

Field Name	Description
Total input read pairs	The total number of read-pairs (R1/R2 read-pairs) provided as input to the pipeline.
Filtered short reads	Number of read-pairs filtered out at the trimming stage due to being too short.
R1 processed bases	Total number of R1 bases provided as input to the pipeline.
R2 processed bases	Total number of R2 bases provided as input to the pipeline.
R1 hairpins trimmed	Total number of hairpins trimmed from R1.
R2 hairpins trimmed	Total number of hairpins trimmed from R2.
R1 poly-G tails trimmed	Total number of R1 poly-G tails trimmed. This occurs when there are at least 9 consecutive G's at the tail of a read.
R2 poly-G tails trimmed	Total number of R2 poly-G tails trimmed. This occurs when there are at least 9 consecutive G's at the tail of a read.
Percentage of input reads remaining after trimming and quality filtering	The fraction of the total input reads in the initial input FASTQ files that is remaining after initial trimming and quality filtering in the CUTADAPT module.

Field Name	Description
Percentage of input bases remaining after trimming and quality filtering	The fraction of the total input bases in the initial input FASTQ files that is remaining after initial trimming and quality filtering in the CUTADAPT module.

6.8 Targeted

This section of the summary report will only be visible if running the pipeline in targeted mode

Field Name	Description
On-target rate	The selected bases, fraction of aligned bases located on or near a baited region.
Fold-80 base penalty	The fold over-coverage necessary to raise 80% of bases in sequenced targets to the mean coverage level in those targets.
Target AT-dropout (%)	A measure of how under-covered AT-rich regions are relative to the mean. A 5% dropout implies that 5% of expected AT reads have mapped outside AT-rich regions.
Target GC-dropout (%)	A measure of how undercovered GC-rich regions are relative to the mean. A 5% dropout implies that 5% of expected GC reads have mapped outside GC-rich regions.
Mean target coverage	Mean coverage across targeted regions.
Median target coverage	Median coverage across targeted regions.
Max target coverage	Maximum coverage observed across targeted regions.
Zero coverage targets	The fraction of targets that had no coverage at any base.
Fold enrichment	The fold by which the baited region has been amplified above genomic background.
Target bases >=1x (%)	The percentage of all target bases achieving 1X or greater coverage.
Target bases >=30x (%)	The percentage of all target bases achieving 30X or greater coverage.
Target bases >=100x (%)	The percentage of all target bases achieving 100X or greater coverage.
Target bases >=1000x (%)	The percentage of all target bases achieving 1000X or greater coverage.

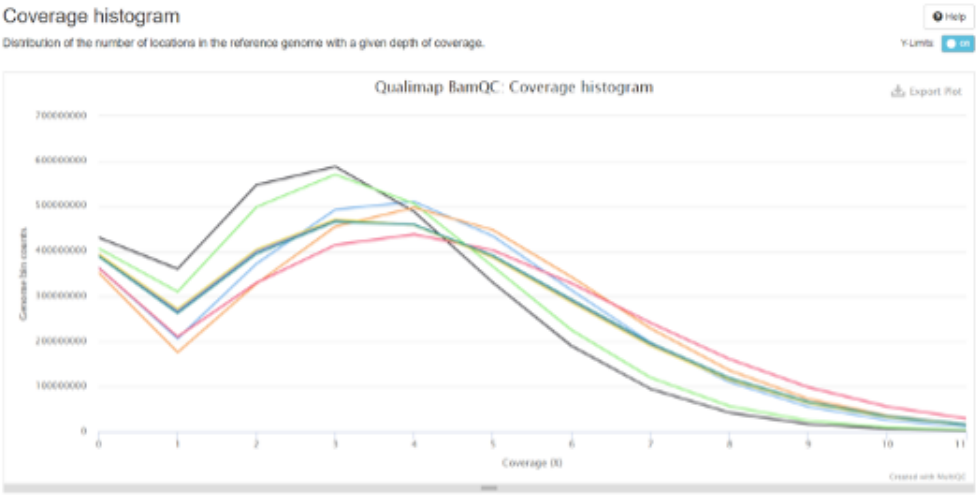
7. Summary Reports

7.1. MultiQC Summary Report

Several QC tests are run automatically on each sample, and the raw outputs are available to be inspected in each sample folder. Summaries of the most important metrics for each sample are collated in the MultiQC report. The MultiQC report presents data from all samples in a single report, enabling easy comparison between samples.

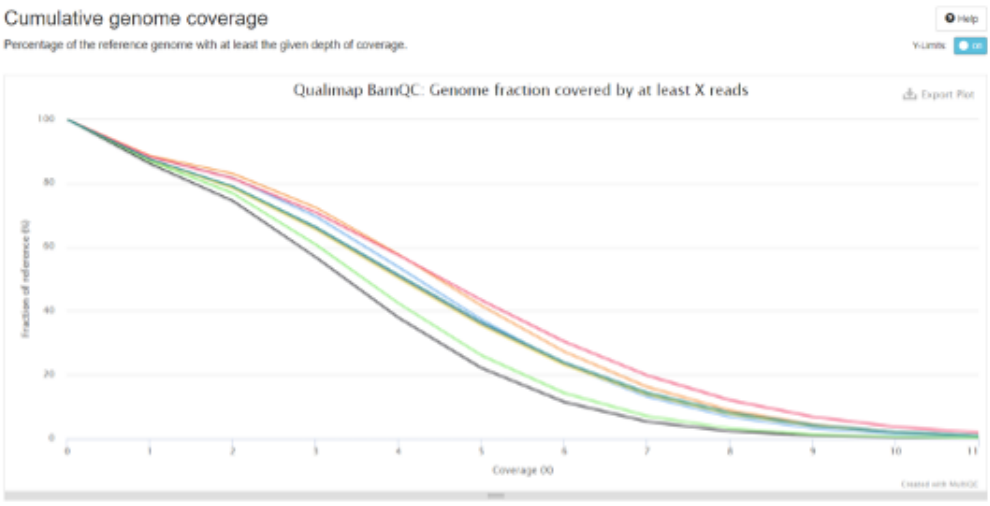
Data provided in the MultiQC report includes:

7.1.1. Qualimap coverage histogram



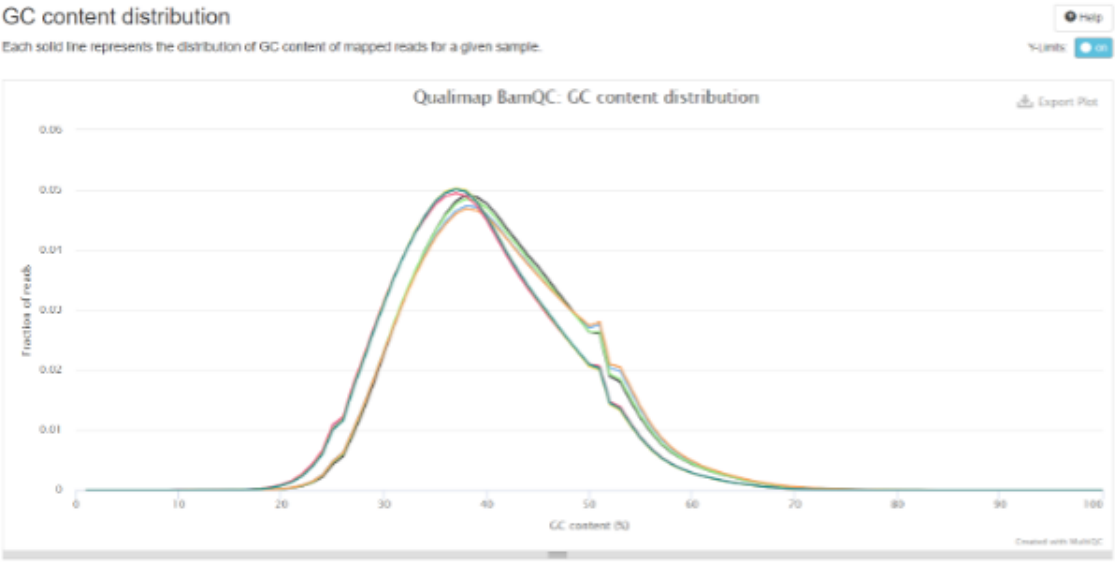
This histogram reports the number of genomic windows that have been sequenced at any given coverage.

7.1.2. Cumulative genome coverage



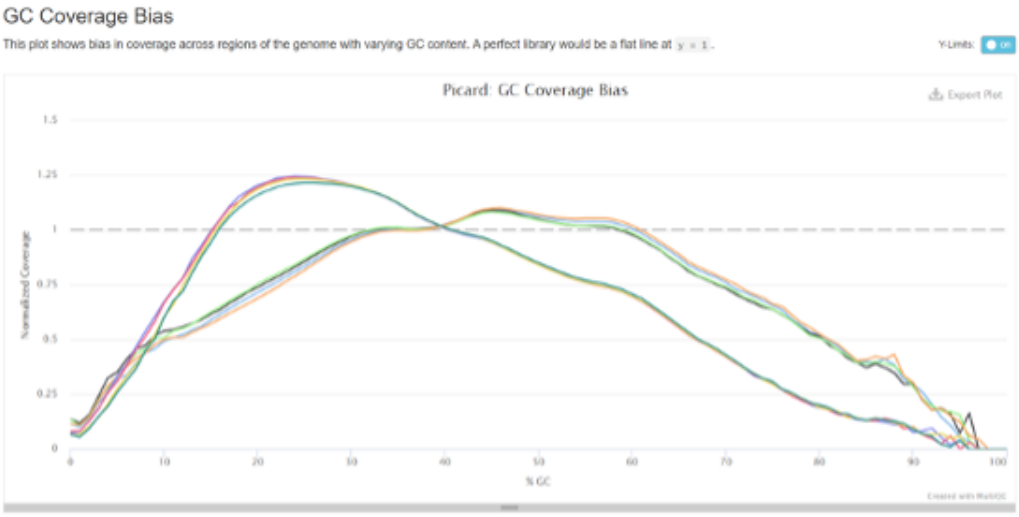
This graph reports the fraction of the genome that has been sequenced with at least nX coverage. It is expected that approximately 8.5% of the genome will be inaccessible to sequencing due to its repetitive nature (so the fraction of genome covered with at least 1X will be at most ~91.5%).

7.1.3. GC content distribution



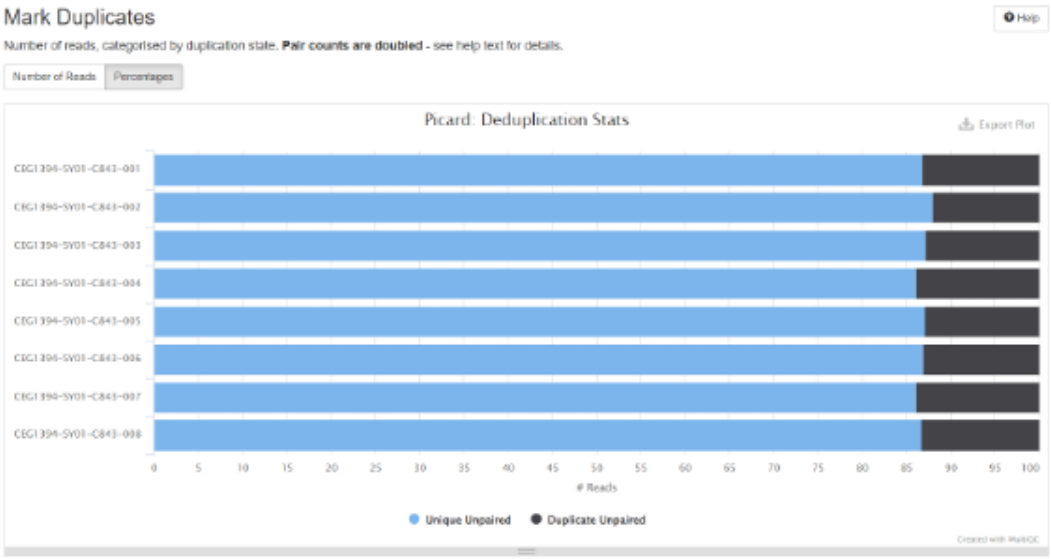
This graph reports the distribution of GC content for all mapped reads. In the case of the human genome, this is expected to be a relatively wide distribution centred on 35%~40%.

7.1.4. Picard GC coverage bias



This graph reports the normalized coverage (nc) across genomic windows with varying GC content. An ideal theoretical library would be a flat line with nc = 1 for all GC content bins. In practice good libraries will have `0.75<nc<1.25` for all genomic windows between ~25% and ~60% GC content. In the example below there is a set of gDNA libraries and a set of cfDNA libraries which show a differing pattern of GC coverage bias.

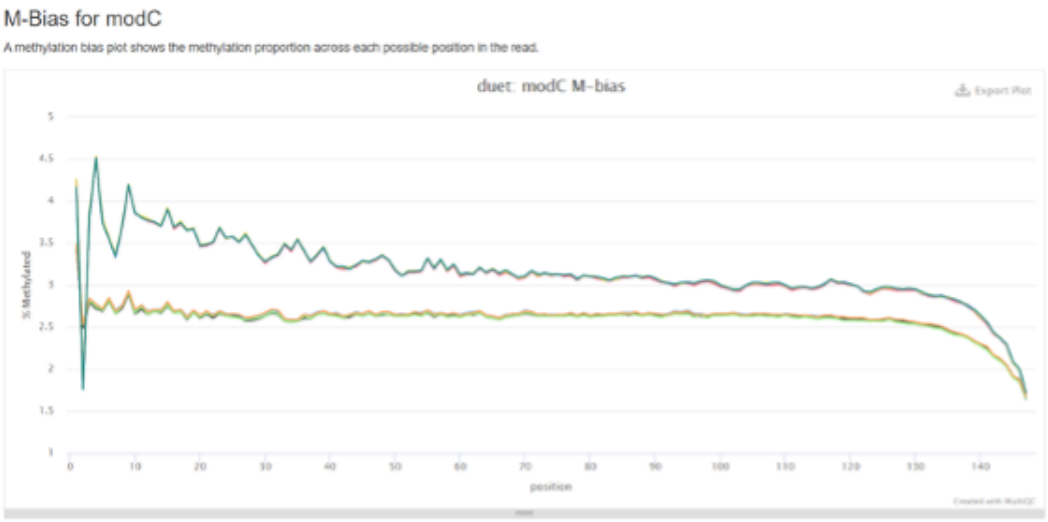
7.1.5. Picard MarkDuplicates



This graph reports the fraction (or number) of duplicate reads in each sample. Duplicate reads arise from PCR amplification or from incorrect segmentation of clusters during sequencing. With single-end reads, duplicates also arise from non-duplicate reads sharing the same start genomic position by chance, and thus overall duplication rate is expected to increase with coverage unless steps are taken to reduce this effect using UMIs. Without UMIs, duplication rates of ~14% for ~30X and ~20% for ~60X are considered normal.

7.1.6. M-bias

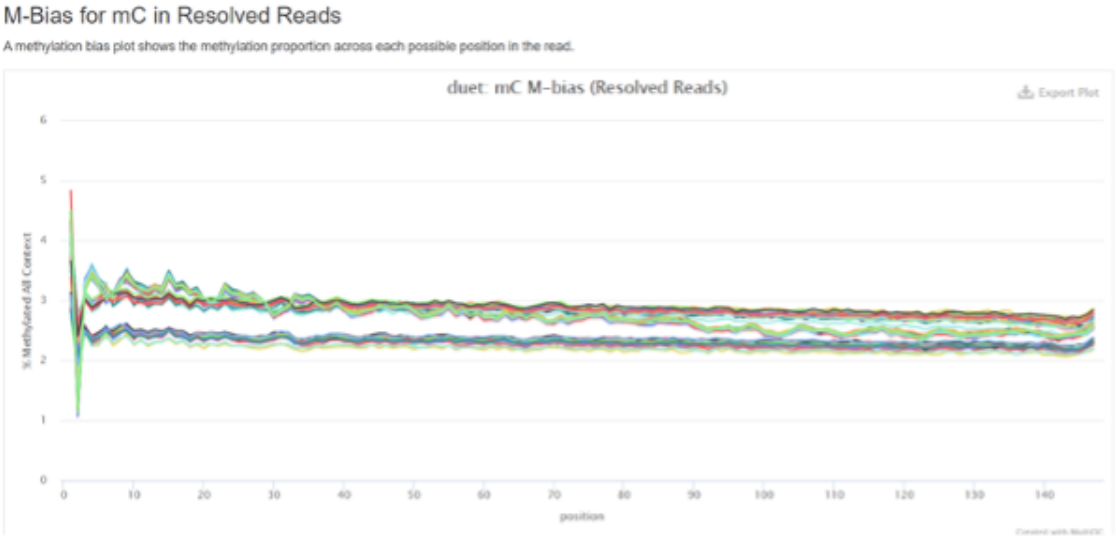
duet +modC



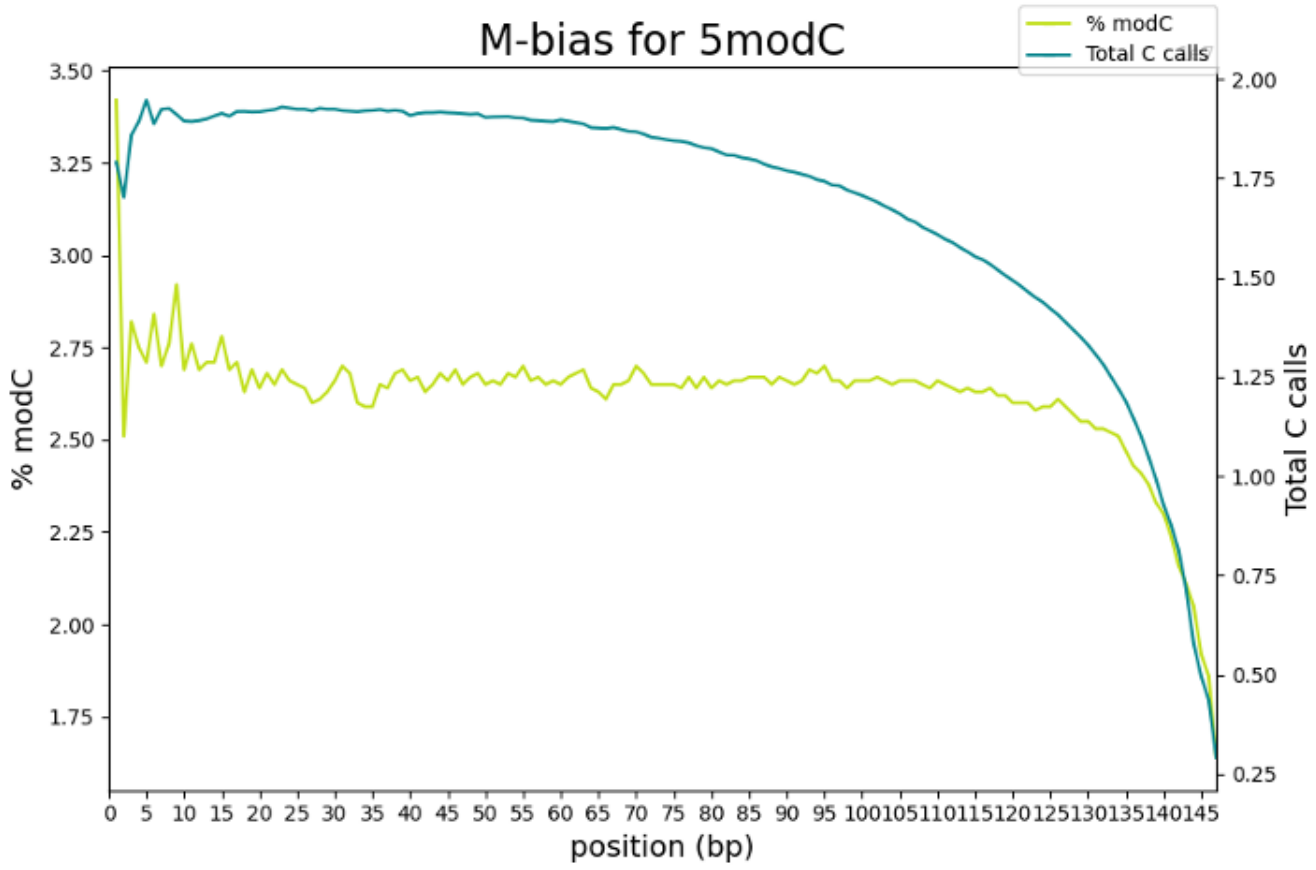
In duet +modC, this graph reports the fraction of modified cytosine sites as a function of the position in the resolved read. An ideal library from a human genome will have a flat line at around 2%~6% for most of the read, depending on the overall methylation of the particular sample. Increased variability is expected for positions approaching the maximum read length as the number of reads analysed decreases substantially towards the maximum read length due to hairpin trimming. A constant increase in the fraction of modified C for read positions towards the maximum read length could indicate incomplete hairpin trimming, as the hairpin is fully methylated. In the example above there is a set of gDNA libraries and a set of cfDNA libraries which show greater methylation on the cfDNA reads.

duet evoC

In duet evoC, separate plots are presented for mC, hmC, and undifferentiated modC. The following plots provide examples from a sequencing run featuring a set of gDNA libraries and a set of cfDNA libraries featuring greater methylation.



Note that in duet evoC, the modC category represents those cases where a modification has been detected, but it has not been possible to determine whether it was mC or hmC. This occurs when the modification occurs in a CH context, or where the succeeding base is an N, or where the modification occurs on the last base of the read or of the fragment. Therefore, this is expected to increase with read cycle and that effect is expected to be more prominent in gDNA than cfDNA due to increased occurrence of shorter fragments.



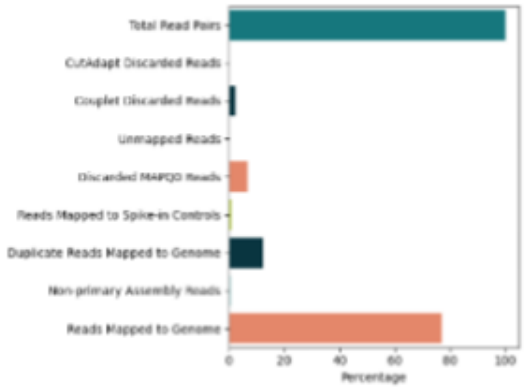
7.2. Sample-Level QC summary reports

Sample-level QC reports displayed graphically are available in the `dqsreport` subdirectory. There will be a separate report for each sample. Reports include sequencing, genomic, and epigenetic QC metrics.

7.2.1. Resolution and alignment stats

Resolution and Alignment Stats

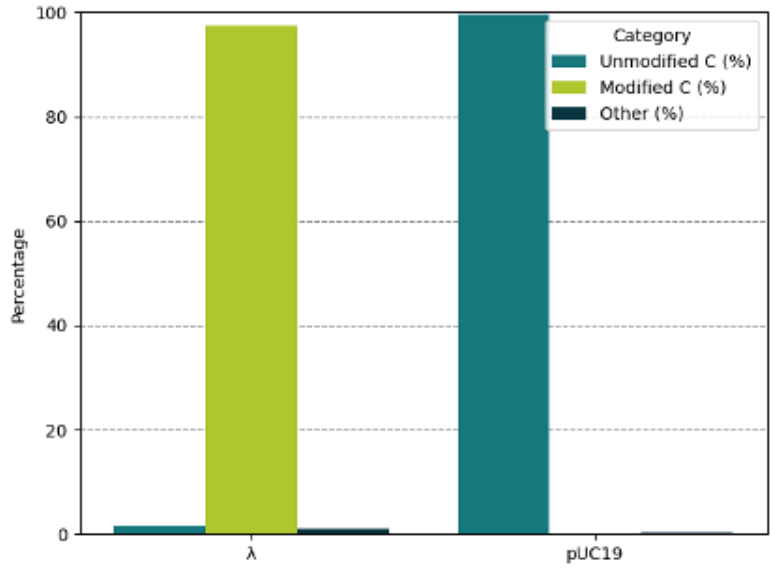
Category	Count	Percentage
Total Read Pairs	9366293	100.000
CutAdapt Discarded Reads	7224	0.077
Couplet Discarded Reads	218973	2.338
Unmapped Reads	5686	0.061
Discarded MAPQ0 Reads	626868	6.693
Reads Mapped to Spike-in Controls	75038	0.801
Duplicate Reads Mapped to Genome	1164641	12.434
Non-primary Assembly Reads	58997	0.630
Reads Mapped to Genome	7208866	76.966

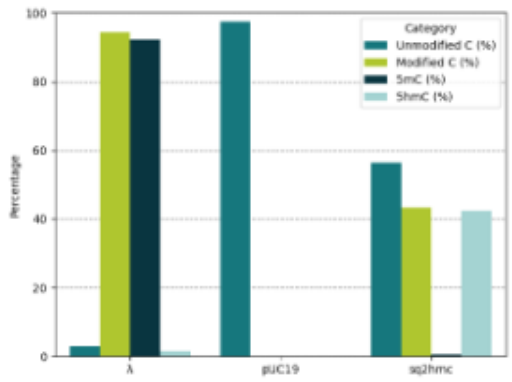


This plot provides insight into read retention rate at consecutive stages of processing:

- **Total Read Pairs:** This will always be presented as 100% and can be used as a reference to compare the other bars against. The other bars should sum to the Total Read Pairs.
- **Cutadapt Discarded Reads:** This shows the proportion of reads that were filtered out at the initial trimming and quality-filtering step.
- **Couplet Discarded Reads:** This shows the proportion of reads that were discarded because they failed to resolve via duet resolution rules. These are likely to be DNA sequences that were not in the expected format for a duet construct.
- **Unmapped Reads:** This shows the proportion of reads that did not map to either the genome or the controls. These might be, for example, contamination or low-complexity reads.
- **Discarded MAPQ0 Reads:** This shows the proportion of reads that were filtered out because they had a mapping quality of zero. This occurs when there are multiple possible alignment loci for a read that all have equal alignment scores; these are therefore likely to be reads associated with repetitive or low-complexity regions of the genome.
- **Reads Mapped to Spike-in Controls:** This shows the proportion of reads that mapped to the spike-in controls provided with the assayDuplicate Reads Mapped to Genome: This shows what proportion of all reads that both mapped to the genome and were identified and removed as duplicates (duplicates may be either PCR duplicates or sequencing duplicates). Non-Primary Assembly Reads: This shows what proportion of reads were filtered out because they mapped to decoy sequences in the reference genome, rather than to chromosomes/contigs that are part of the primary assembly.
- **Reads Mapped to Genome:** This shows what proportion of reads were mapped to the genome and were not identified as duplicates – these are the reads passed downstream for processes such as quantification and variant calling.

7.2.2. DNA modifications





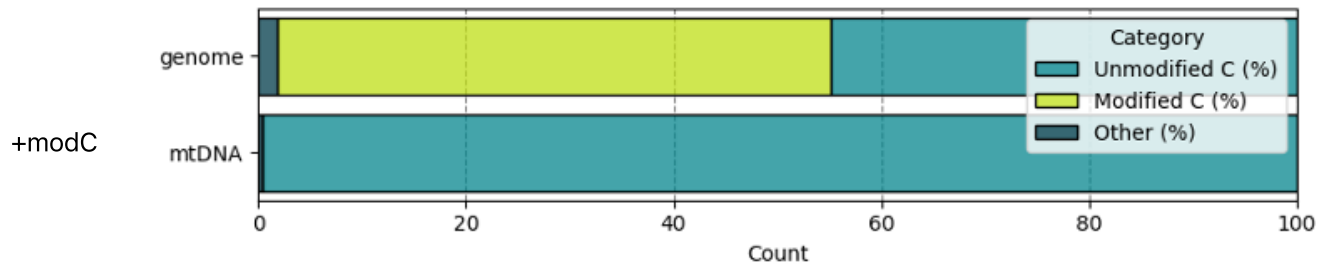
These plots present sensitivity and specificity information calculated from the fully methylated lambda control and the totally unmethylated pUC19 control. The plots show the percentage base calls at CpG sites that we reported as:

- Unmodified C
- Modified C
- Other (for example a non-C called in a CpG context)

These plots demonstrate the accuracy of the assay and would alert you to a failure of the library preparation process. We expect observe sensitivity on the lambda control > 95% and specificity on the pUC19 control > 99%.

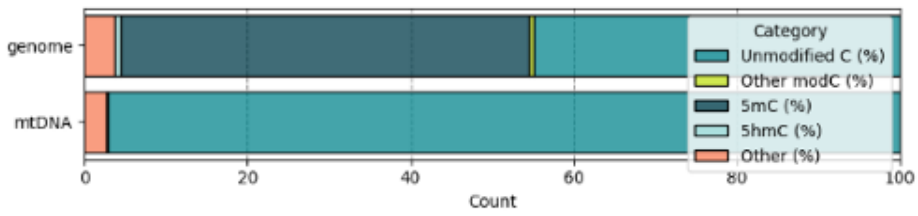
In duet evoC, there is an additional plot showing modification calling rates on a hmC short oligo control that features a mixtures of hmC and unmodified Cs.

7.2.3. Overall CG methylation levels



---

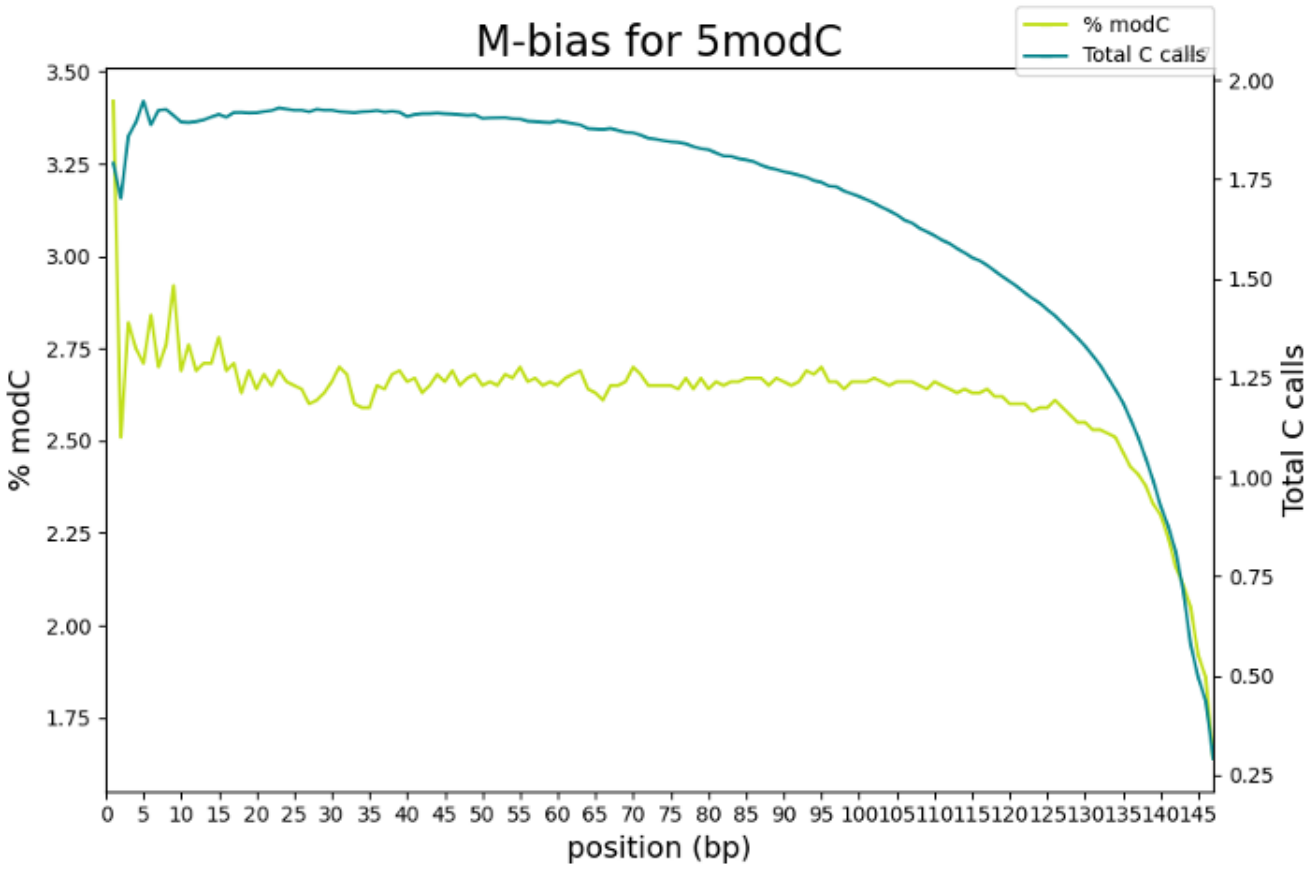
evoC



The genome bar on this plot shows the proportion of CpGs in the autosomes that are reported as modified or as unmodified. This gives an indication of the overall methylation rate across the genome. The allosomes are excluded to avoid any gender-bias.

The mtDNA bar shows the proportion of CpGs in the mitochondrial genome that are reported as modified or as unmodified. This can act as a form of control, because methylation in the mitochondrial genome is considered to be either extremely rare or non-existent.

7.2.4. M-bias



This plot (green line) presents the proportion of base calls that were modC at each position (read cycle) in the genome-aligned reads. This can be interpreted in the context of the total number of C calls (blue line). The number of C calls is expected to drop off towards the end of the sequencing cycles due to fewer fragments extending to the full read length. When the number of C calls becomes low, the signal for modC calls is expected to become noisier.

In evoC, there will be separate m-bias plots for mC, hmC, and undifferentiated modC. Note that an undifferentiated modC call is expected to occur when a modification is detected in CH context, or on the last base of a read, or on the last base of a fragment, or preceding an N. Other than the CH category, these cases are expected to occur more often further into the read cycles, so in evoC, the modC m-bias plot is expected to increase.

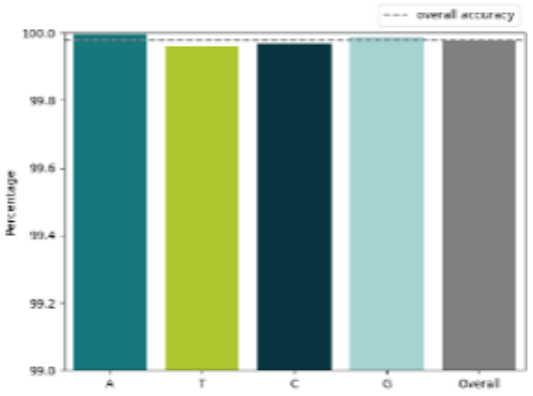
7.2.5. Genetic accuracy

The known genetic sequence of the methylated lambda control is used to calculate genetic accuracy and this is presented per base and overall:

Genetic Accuracy on Spike-in Controls

- Genetic accuracy is computed by counting the proportion of bases aligned to a position which agree with the reference base.* Ns are excluded from this calculation.

	Genetic Accuracy lambda
A	99.9952
T	99.9588
C	99.9669
G	99.9878
Overall	99.9771

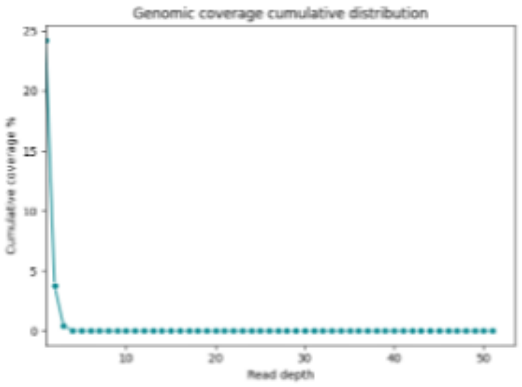


7.2.6. Coverage

Genomic coverage

Genomic coverage percentages are computed using qualimap outputs.

Metric	Value
Mean Coverage	0.29X
% Positions w/ >= Mean Coverage	78.143239999999999%
% Positions w/ >= Half Mean Coverage	89.07162%
% Positions w/ >= 1x Coverage	24.24%

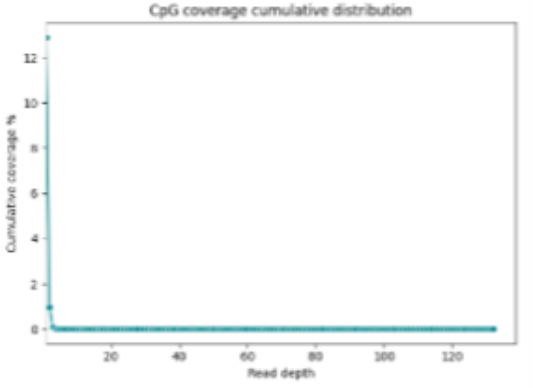


This plot presents the cumulative distribution of coverage on the genome.

CpG coverage

CpG coverage percentages are computed on the basis of human genome - 56,000,000 sites (both strands).

Metric	Value
Mean Coverage	0.14X
% Positions w/ >= Mean Coverage	12.9%
% Positions w/ >= Half Mean Coverage	12.9%
% Positions w/ >= 1x Coverage	12.9%



This plot presents the cumulative distribution of coverage on CpGs.

7.3. FASTQC report

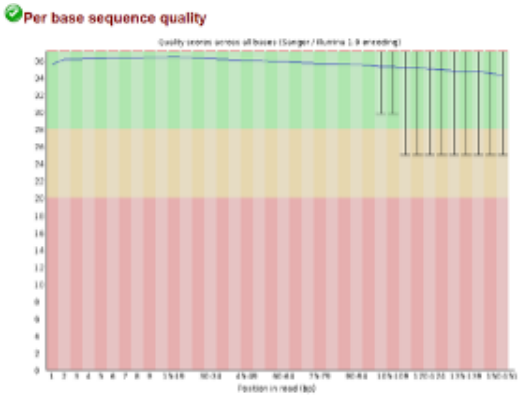
Optionally, the pipeline can generate FASTQC Reports for the raw input reads. If generated, these will be output into a `diagnostics/fastqc_reports/` subdirctory. FASTQC is a commonly used tool for characterising the quality of the raw reads generated from next-generation sequencing.

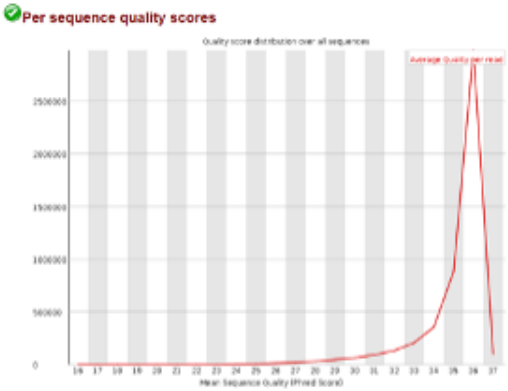
If generated, there will be one report for R1 and one for R2 for each sample for each lane.

The FASTQC Report includes plots characterising:

- Per base sequence quality.
- Per tile sequence quality.
- Per sequence quality scores.
- Per base sequence content.
- Per sequence GC content.
- Per base N content.
- Sequence length distribution.
- Sequence duplication levels.
- Overrepresented sequences.
- Adapter content.

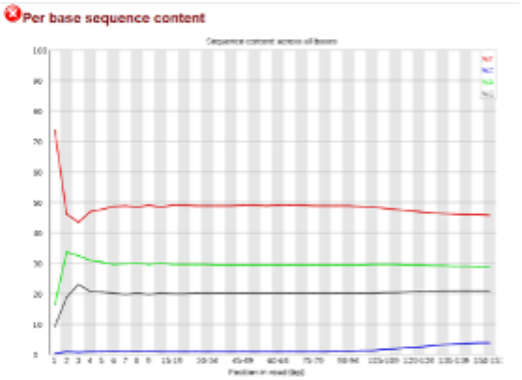
The following examples show the per-base sequence quality and per-sequence quality scores plots:

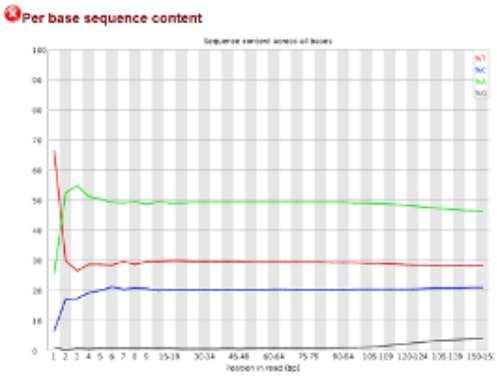




Note that due to the deaminated nature of duet libraries, the FASTQC report is expected to flag warnings associated with per-base sequence content and per-sequence GC content.

A typical per-base sequence content plot will look like this:





The first plot is from an R1 report, showing C-depleted R1 reads. The second plot is from an R2 report showing G-depleted R2 reads.

The bias at the first position is caused by an artefact left over from the duet construct and this gets trimmed off as part of the pipeline processing. The change in GC content towards the end of the reads is expected to coincide with reaching the end of some fragments and reflects the sequencing of the hairpin in the duet construct.