duet Pipeline Documentation

Release 1.5.0

biomodal Ltd.

May 06, 2026

CONTENTS

This quick start guide gives a concise overview of how to use our Data Interpretation Guide. It outlines the key steps to assess your data quality and begin exploring your duet pipeline outputs, depending on the level of analysis you want to achieve, whether you are a lab scientist or a bioinformatician. The full Data Interpretation Guide is comprehensive, you don't need to read it all at once. Instead, use the sections that are most relevant to your role and the depth of analysis you're aiming for.

CHAPTER

ONE

WHO IS THIS FOR?

This guide is for anyone who has run the duet pipeline and wants to understand how to interpret the output data. It is particularly useful for:

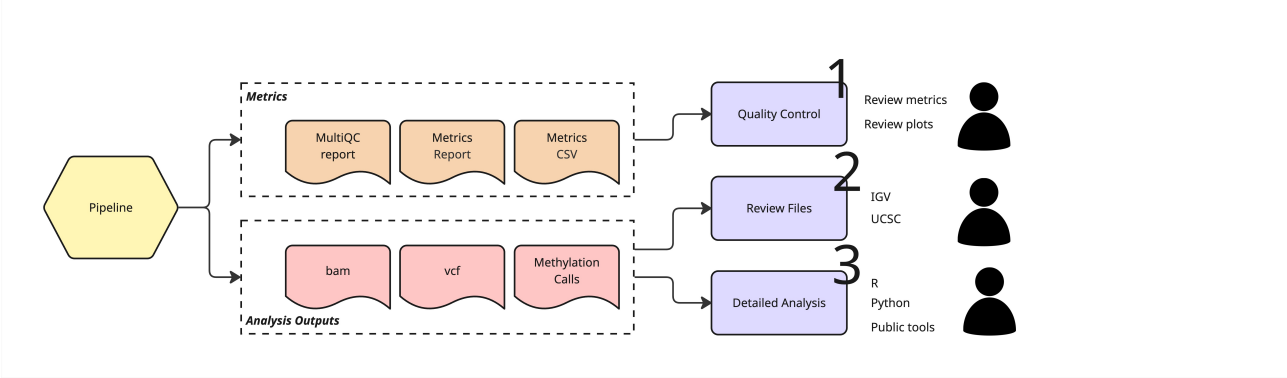
- **Lab Scientist**: You want to understand what the data says about your biological samples.
- **Bioinformatician**: You want to explore, process, or integrate the data into downstream analyses.

CHAPTER

TWO

OVERVIEW: FROM DATA TO INSIGHT

Note: The diagram below shows the main files the duet pipeline produces, and how they can be used to derive biological insights. The duet pipeline produces additional optional output files that are not listed in the diagram.



1. Review the metrics and plots generated by duet to make sure your data is of high quality.
2. Explore individual output files with standard tools to get a feel for your data.
3. Use the output files to derive biological insights, such as differential methylation or integration with other omics data.

CHAPTER

THREE

STEP-BY-STEP

3.1 For Lab Scientists

1. **Check Your Assay Type** - duet +modC or duet evoC? *See assay overview*
2. **Review technical QC measures** - Understand how duet assesses accuracy. *Summary metrics*
3. **Open the Summary Report** - Review key metrics. *Metrics summary* and *MultiQC report*
4. **Explore Methylation Calls** - Find the methylation output files that suit your needs. *Epigenetic Quantification*

3.2 For Bioinformaticians

1. **Locate Output Files** - BAM, BED, TSV, JSON. *File formats*
2. **Dive into the Metrics** - Use summary files or raw data. *Metrics analysis*
3. **Visualise your alignment files** - Using IGV, R, Python. *Alignment output files*
4. **Explore Epigenetic Quantification** - Using one of the output files. *Quantification*

CHAPTER

FOUR

HOW TO USE THIS GUIDE

This guide is structured to help you quickly find the information you need. You can navigate through the major sections using the tabs at the top of this page, and then by using the table of contents on the right side of the page.

Additionally, you can use the search function in the top right-hand corner of the page to find specific topics or keywords across all biomodal software pages.

To get back to this page, click the 'home' icon.

For any questions or issues, please contact your biomodal support team at **support@biomodal.com**.

CHAPTER

FIVE

USER DOCUMENTATION CONTENTS

5.1 duet data interpretation guide

This documentation is compatible with duet pipeline version 1.5.x

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 post-sequencing demultiplexing are required with a specific file naming format and, optionally, 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 your chosen executor. The duet software installation and running guide provides instructions for the setup and configuration of your environment. This guide helps you understand the pipeline workflow and outputs and explains the format of the files generated.

5.1.1 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.
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.

5.1.2 Pipeline 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 (ASM). Summary reports are generated at the conclusion of the pipeline.

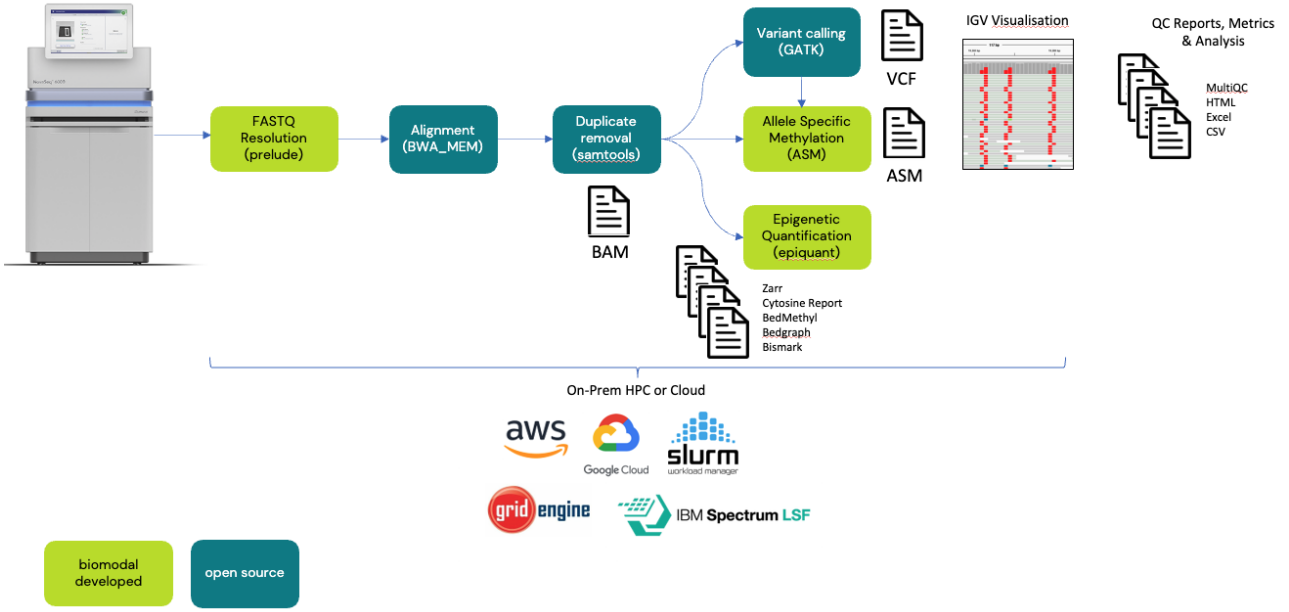


Fig. 1: Figure 1: biomodal analysis pipeline workflow

5.2 Pipeline workflow

5.2.1 Input files

The biomodal analysis pipeline requires two types of input files:

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

For FASTQ file and metadata file naming conventions, please see the CLI installation guide 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.

5.2.2 Trimming, filtering and resolution

The trimming, filtering and resolution of reads takes place in a bespoke module called Prelude which converts the lane-wise 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 first removes a single base from the 5' end of each read. This is because a single additional base is expected to be present left over from the excision of the hairpin in the event of a double hairpin ligation event, and this is expected to affect a high proportion of reads. Next, poly-G tails that are longer than 8nt are removed from the 3' end of each read. This is because poly-G tails are expected to indicate the absence of a signal from the sequencer. Additionally, any flanking N's are removed.

Next, the trimming algorithm identifies and removes the hairpin, which is expected to be present on the 3' ends of any reads derived from constructs where the original DNA fragment had a length that was shorter than the read length. Hairpin

identification requires an overlap of at least 3 bases with the beginning of the hairpin sequence, tolerates a 10% mismatch rate, and allows for variable deamination status at one unprotected site in the hairpin sequence. An *XL* tag is added to the read to record the original fragment length. This tag is set to *0* if no hairpin has been identified, indicating that the original fragment was longer than the read length.

Reads with more than five Ns present, and reads that are shorter than 15nt after the removal of all potential artefacts are filtered out.

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-base 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: 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, five are expected to be observed and 11 are not expected to be observed
- In duet evoC, six 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 nth base of read 1 is paired with the nth 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 prunes 3' tails until there are no N's in the last 15 bases of the read. Note that this quality trimming might result in a resolved read that is shorter than the original fragment length recorded in the XL tag.

Finally, any Cs in the last three bases of the resolved read are masked by being converted to Ns and any trailing N's after this masking are removed. 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.

5.2.3 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.

5.2.4 Duplicate marking

The marking of duplicates is performed using samtools markdup.

5.2.5 BAM file filtering and collection of alignment and coverage metrics

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, supplementary alignments and reads with a mapping quality of zero are also removed.

Note: Reads with a mapping quality of zero are reads that align equally-well to more than one location in the reference genome. In this instance, the aligner chooses one of the locations at random to become the primary alignment and the other locations become secondary alignments.

Note: Supplementary alignments are reads where the read has been split because different parts of the read align to different locations in the genome. The longer part of the read remains as the primary alignment and the shorter part becomes a supplementary alignment.

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: If the 'targeted' mode of the pipeline is invoked, then duplicates are first marked so that targeted metrics can be calculated, and then duplicates are removed for downstream analysis. The published BAM file will have had duplicates removed and will also have been filtered to include only reads that align to the target regions.

5.2.6 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 - i.e. unmethylated, methylated (mC), or hydroxymethylated (hmC) (or modC which means either mC or hmC without being able to differentiate between them)
- 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 primary output of the quantification step is:

- A multi-sample zarr store containing methylation information about all CpGs in all samples in a single compressed file format.

Additionally, a selection of per-sample plain-text quantification file formats can be generated. The default plain text quantification file generated is the Cytosine Report format. However, parameter changes can cause any combination of the following plain text quantification file formats to be generated:

- Cytosine Report
- BedMethyl
- Bedgraph
- Bismark

These file formats are described in further detail in *Epigenetic quantification output files*

With duet +modC, for each file type requested, each sample will have a single file generated, which will contain the modC calls

With duet evoC, each sample will have three files generated:

- A file containing mC calls
- A file containing hmC calls
- A file containing modC calls, which are the union of the mC calls, the hmC calls, and any undifferentiated modC calls where a modification was detected but it was not possible to determine whether the modification was an mC or an hmC.

Note: If CHG/CHH calling is requested, additional plain text quantification files will be generated for each sample for CHG calls and for CHH calls. These will only contain modC calls, even with the evoC assay, as duet evoC can only differentiate mC and hmC in CpG contexts, not at CH contexts.

5.2.7 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), hydroxymethylcytosines (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 +modC, 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; modC sensitivity is also calculated as the fraction of calls that are mC, hmC, or undifferentiated modC versus expected mC calls in the lambda genome.

5.2.8 Variant calling

Germline variant calling

By default, the pipeline runs the Genome Analysis Toolkit (GATK) HaplotypeCaller for germline variant calling. This will generate a single VCF file for each sample on the run. Note that germline variant calling can be disabled via a parameter override.

Joint variant calling

By default, germline variants are called on each sample independently. However, it is also possible to perform joint germline variant calling. In this scenario, a single gVCF file is generated for each sample on the run. These then get consolidated via GATK GenomicsDBImport followed by joint genotyping using GATK GenotypeGVCFs.

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. For example, a region on the maternal chromosome might have a different methylation pattern from the same region on the paternal chromosome. 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, or haplotypes, with a heterozygous genotype. For each heterozygous SNV or phased haplotype in the VCF file, the module quantifies methylation that occurs at CpGs on reads associated with each allele. It then applies filtering; by default, only loci that have six or more reads spanning the allele and at least one CpG are evaluated. Then, if more than a 30% difference in methylation level between the two alleles is observed, ASM is called. A statistical test is applied to produce a p-value that quantifies confidence 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, or phased haplotype. The VCF and ASM file formats are described in *Outputs*

Note: Note that if the joint variant calling mode is combined with the ASM mode, then the ASM module will be run on the variants identified from the joint genotyping step.

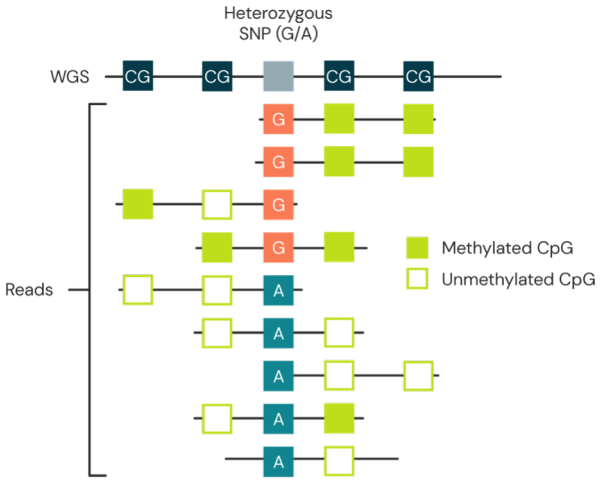


Fig. 2: Figure 2. Example of methylated and unmethylated CpGs on a heterozygous SNV

5.2.9 Somatic 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.

5.2.10 Report creation

The pipeline outputs data files and reports containing extensive sample information — see *Outputs* and *Metrics* sections.

5.2.11 Targeted sequencing

The biomodal duet pipeline includes a mode for processing data generated using target panels. 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:

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

In the targeted mode, a BAM with marked duplicates is passed into the GATK CollectHsMetrics module to calculate targeted sequencing metrics and into the QUALIMAP_BAMQC module to calculate genome-wide coverage metrics. The filename for this BAM has the pattern *{sample_id}.genome.{genome_tag}.markdup.bam*, for example *BM100.genome.GRCh38Decoy_primary_assembly.markdup.bam*

Separately, a BAM file with duplicates removed, and filtered to include only reads that align to the target regions, is published as the final output BAM file. This BAM file is used for variant calling and epigenetic quantification downstream. It is also used for the calculation of GC Bias and M-Bias metrics. The filename for this BAM has the pattern *{sample_id}.targeted.{genome_tag}.markdup.bam*, for example *BM100.targeted.GRCh38Decoy_primary_assembly.markdup.bam*.

5.3 Outputs

Outputs from the biomodal pipeline are organised into the following top-level directory structure:

Directory	Contents
`repor`	Multi-sample reports summarising information about the samples and controls
`sampl`	Primary data files generated by the pipeline (described in more detail below)
`contr`	BAM files and quantification files associated with the 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: `ss-ctrls-long-v24.fa.gz`
`diagn`	Secondary outputs from the pipeline including that can be useful for investigating and diagnosing issues

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

File	File Name	Subdirectory
BAM (default)	`{A}.genome.{B}.dedup.bam` and `{A}.genome.{B}.dedup.bam.bai`	`sample_outputs/bams/`
CRAM (optional - alternative to BAM)	`{A}.genome.{B}.dedup.cram` and `{A}.genome.{B}.dedup.cram.crai`	`sample_outputs/crams/`
Germline VCF (default)	`{A}.genome.{B}.dedup.vcf.gz` and `{A}.genome.{B}.dedup.vcf.gz.tbi`	`sample_outputs/variant_call_files/germline/`
Germline joint genotyping VCF (optional alternative to single-sample VCFs)	`{run_name}.joint_genotyping.vcf.gz` and `{run_name}.joint_genotyping.vcf.gz.tbi`	`sample_outputs/variant_call_files/germline/`
Somatic VCF (optional)	`{A}.genome.{B}.dedup.somatic.vcf.gz` and `{A}.genome.{B}.dedup.somatic.vcf.gz.tbi`	`sample_outputs/variant_call_files/somatic/`
Quantification Cytosine Report (default)	`{A}.genome.{B}.dedup.{C}.num_{D}_cxreport.txt.gz` and `{A}.genome.{B}.dedup.{C}.num_{D}_cxreport.txt.gz.tbi`	`sample_outputs/modc_quantification/`
Quantification BedMethyl File (optional)	`{A}.genome.{B}.dedup.{C}.{D}.bed.gz` and `{A}.genome.{B}.dedup.{C}.{D}.bed.gz.tbi`	`sample_outputs/modc_quantification/`
Quantification Bedgraph File (optional)	`{A}.genome.{B}.dedup.{C}.frac_{D}.bdg.gz` and `{A}.genome.{B}.dedup.{C}.frac_{D}.bdg.gz.tbi`	`sample_outputs/modc_quantification/`
Quantification Bismark File (optional)	`{A}.genome.{B}.dedup.{C}.num_{D}_bismark.txt.gz` and `{A}.genome.{B}.dedup.{C}.num_{D}_bismark.txt.gz.tbi`	`sample_outputs/modc_quantification/`
ASM File (optional)	`{A}.asm.csv`	`sample_outputs/allele_specific_methylation/`
Multi-sample zarr store (default)	`{run_name}.genome.{B}.dedup.{C}.zarrz`	`sample_outputs/zarr_store/{C}/`

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 `CG`, `CHG`, or `CHH` depending on the context and whether CHG/CHH calling has been requested at the time of launch.
- `{D}` is `modc`, `mc` or `hmc` to indicate the category of modification featured in the file
- `{run_name}` is the parameter provided to the pipeline at the time of launch.

Note that:

- In pipeline version 1.4.1, with CHG/CHH calling enabled, multi-sample zarr stores for CG, CHG, and CHH contexts are all output to the same subdirectory, `sample_outputs/zarr_store/`, but in pipeline version 1.4.2 they get separated into subdirectories
- In pipeline version 1.4.1, with joint variant calling enabled, the joint genotyping VCF is output to `variant_call_files/` and is not prefixed with the `{run_name}`. but in release 1.4.2 it is output to `sample_outputs/variant_call_files/germline/` and it is prefixed with the `{run_name}`
- BAM, VCF, and the quantification 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
- ASM files are only generated if requested at the time of launch
- By default, the only plain text quantification files generated are the Cytosine Reports. However, parameter overrides can cause any combination of the quantification file formats to be generated

Each file type is further described below.

5.3.1 Alignment output files

BAM output 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:

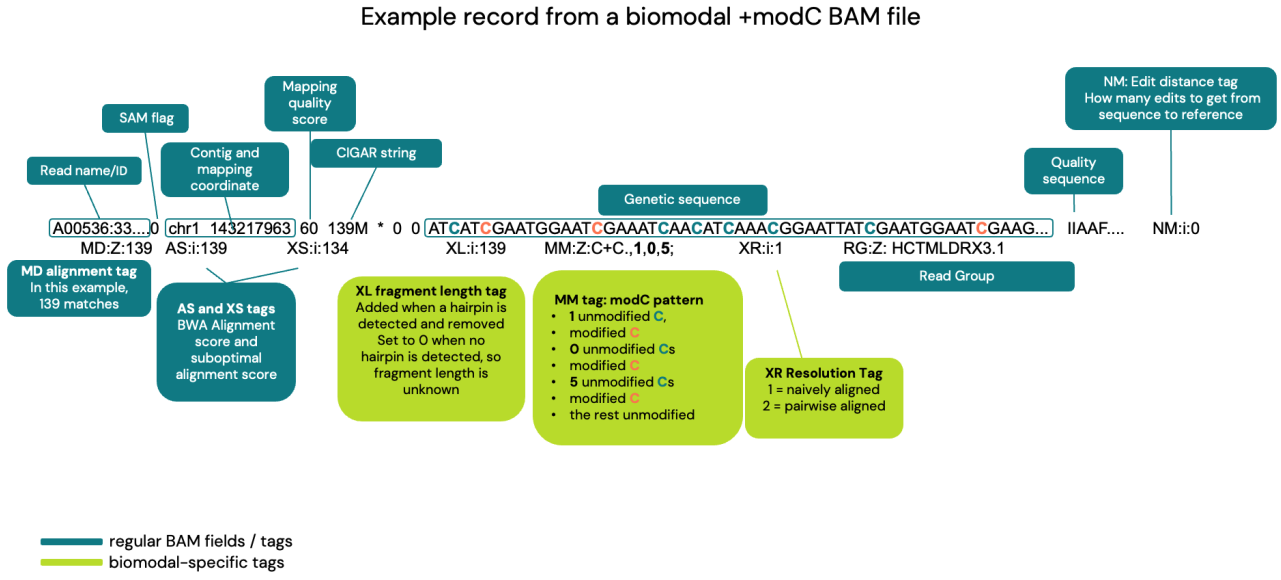


Fig. 3: Figure 3. Breakdown of a biomodal duet +modC BAM file

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 format described in the SAM file specification. It is interpreted as follows: starting at the beginning of the sequenced read, 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. Note that to interpret the MM tag of a reverse-strand aligned read, the reverse complement of the bam-reported sequence needs to be used.
- 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).

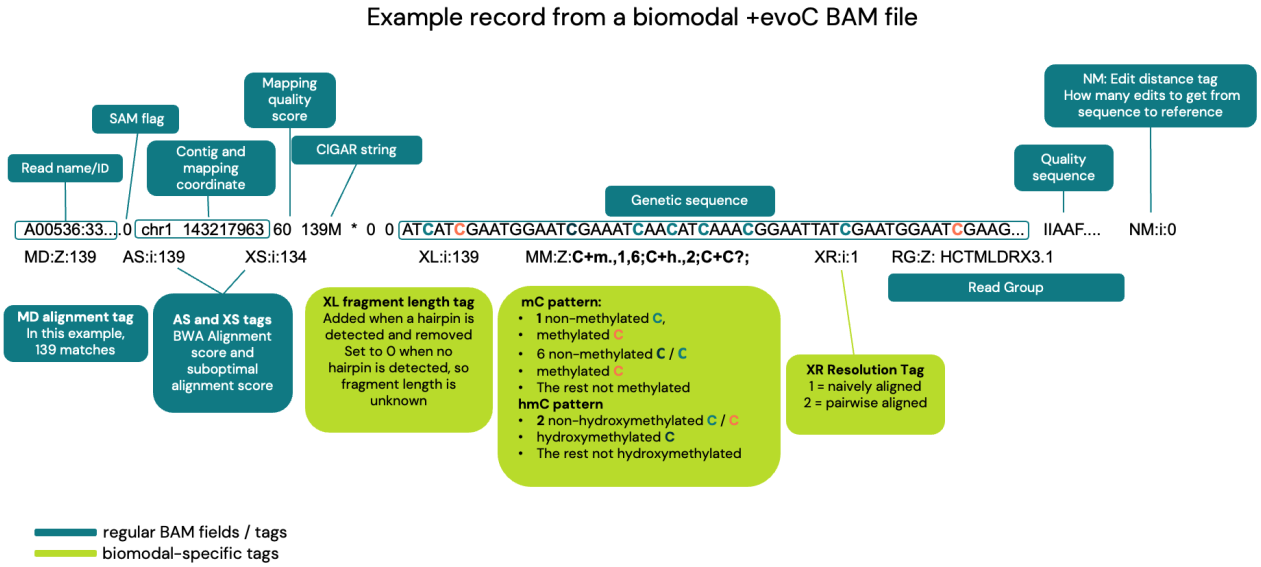


Fig. 4: Figure 4. Breakdown of a biomodal duet evoC BAM file

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. For visualisation of methylation status in IGV, we recommend using `v2.18.0` and right-click menu and toggling the ‘color alignments by -> base modification 2-colour’ setting.

Note: Some downstream tools that parse BAM files may also require the corresponding reference FASTA and an associated index, which can be obtained from the reference directory. If the reference genome you are working with includes decoy contigs (as is the case for the *GRCh38_Decoy* reference) then these decoy contigs will be present in the reference FASTA file, but will have been filtered out of the final BAM file generated by the pipeline. For compatibility with downstream tools that require the reference FASTA, you will need to use the primary assembly bed file to filter the FASTA file to generate a FASTA file that only contains the primary assembly contigs. This can be done with common tools such as *bedtools* or *samtools faidx*. For example, to generate a new FASTA file for the *GRCh38Decoy* reference to only include the primary assembly contigs using *bedtools*, you can run:

```
bedtools getfasta -fi GRCh38Decoy.fa.gz -bed GRCh38Decoy_primary_assembly.bed | fold -w 60␣
↪| bgzip > GRCh38Decoy_primary_assembly.fa.gz
```

...where *GRCh38_Decoy.fa.gz* is the reference FASTA file and *GRCh38Decoy_primary_assembly.bed* is the primary

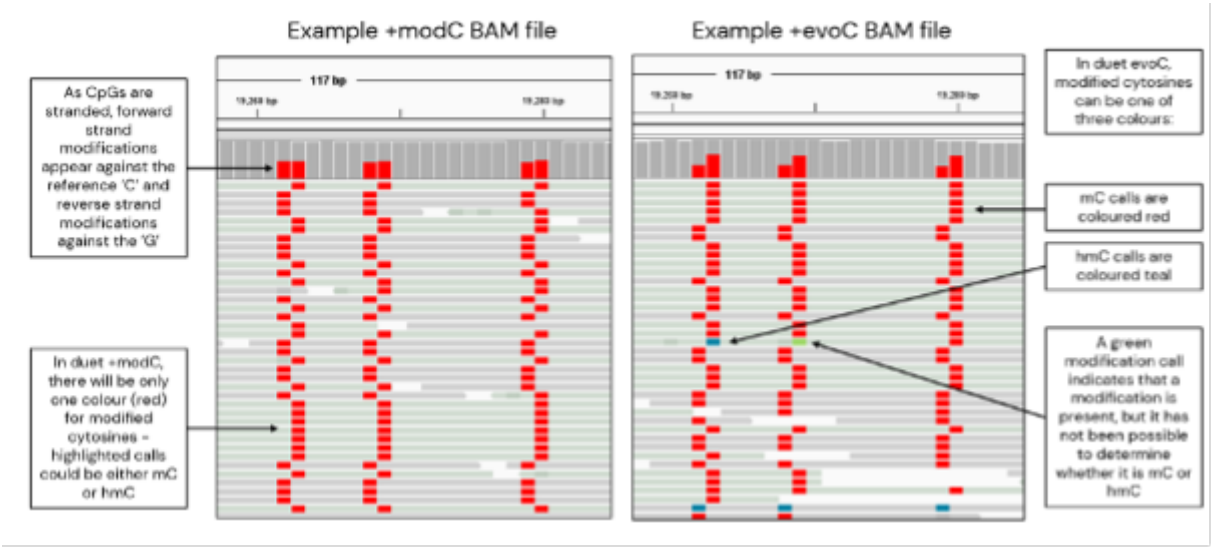


Fig. 5: Figure 5. viewing a biomodal BAM file in IGV

assembly bed file that contains the primary assembly contigs, both of which can be found in the *reference* directory. The *fold -w 60* command sets the length of the lines in the FASTA file to 60 characters. Note that to execute the above command, you will need to have *bedtools* and *bgzip* installed on your system.

CRAM output 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 approximately 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
```

5.3.2 Variant calling output files

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.

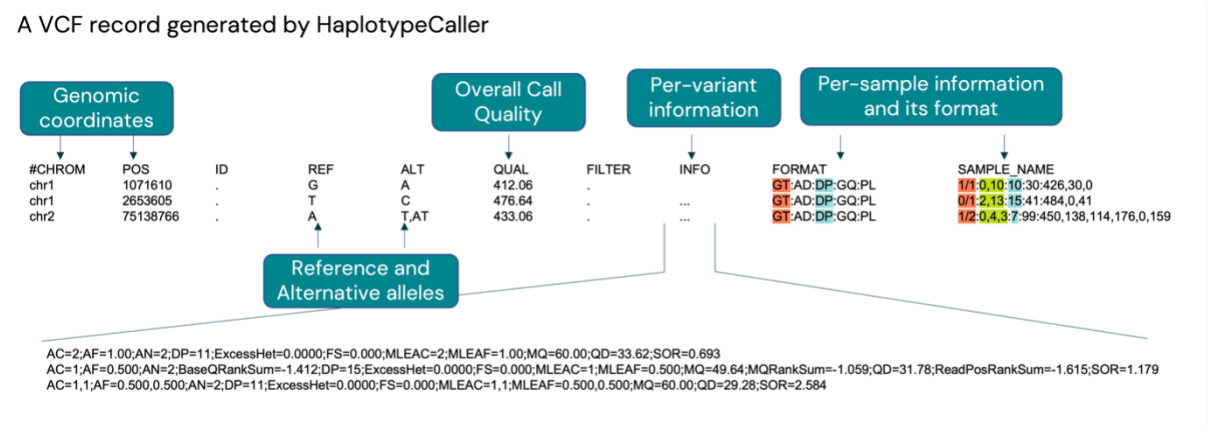


Fig. 6: Figure 6. Breakdown of a biomodal VCF file

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, for the third variant, 0 reads support the REF allele and 4 reads support the first ALT allele, and 3 reads support the second ALT allele).
- 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.

5.3.3 Epigenetic quantification output files

There are four types of plain-text epigenetic quantification files that can be generated from the pipeline:

- Cytosine Report
- BedMethyl
- Bedgraph
- Bismark

These file types are all similar with each file type containing one row for each stranded CpG and with columns that quantify the number of modified and unmodified cytosine calls at that CpG. Which file type you choose to generate may depend upon:

- Which downstream tools you plan to use to process them (some downstream methylation analysis tools, such as those from the `methylKit` R library, or those from the Bismark suite of tools may require plain-text quantification files in a specific format)
- Whether you want to load and view them in IGV (only the BedMethyl and the Bedgraph files can be viewed in IGV)
- Which file type, if any, you are familiar with and have existing scripts for processing

For each quantification file type requested:

- In duet +modC, you will get a single quantification file per sample, which will contain the modC calls.
- In duet evoC, you will get three quantification files per sample: one containing mC calls; one containing hmC calls; and one containing modC calls. In evoC, modC calls are the union of the mC calls, the hmC calls, and any undifferentiated modC calls (where a modification was detected but it was not possible to determine whether the modification was an mC or a hmC).

Whichever of these file types you choose to generate from the pipeline, they will all share the following properties:

- Each file will feature quantification of one type of modification (modC, mC, or hmC); all remaining calls at that position will be grouped together as 'not' having been called as that type of modification
- The data in each row will correspond to one **stranded** CpG in the **reference genome** - i.e. the reference genome is used the determine the set of sites at which to quantify and a CpG pair will have its forward strand cytosine quantified separately from its reverse strand cytosine
- CpGs with no coverage will be excluded from the file by default (but can be included via a parameter override)
- The coverage counts will by default exclude any non-C genetic bases (G, A, or T) called at the CpG (but these can be included via a parameter override).

  Note: If this is changed at the time of launching the pipeline, this will adjust any of fields marked as (*) below to include non-C genetic bases

  ---

- The coverage counts will exclude bases called/resolved as N
- In evoC, the modC outputs will always contain the combination of mC, hmC, and any undifferentiated modC calls

By default, only cytosines in a reference CpG context are reported, although reporting cytosines in reference CHG/CHH contexts is also possible. With CHG/CHH calling, additional plain text quantification files will be generated for each sample for CHG calls and for CHH calls. These will only contain modC calls, even with the evoC assay, as duet evoC can only differentiate mC and hmC in CpG contexts, not at CH contexts.

Cytosine report duet +modC quantification file

Fields featured in the **duet +modC** cytosine report are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Position	The genomic position of the cytosine (one-indexed)
Strand	Whether the CpG is on the forward or reverse strand
modC	Count of modified cytosines
C	Count of unmodified cytosines (*)
Dinucleotide context	The reference dinucleotide
Trinucleotide context	The reference trinucleotide

Note: The genomic positions in a cytosine report are one-indexed.

Cytosine report duet evoC quantification files

Fields featured in the **duet evoC mC** cytosine report are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Position	The genomic position of the cytosine (one-indexed)
Strand	Whether the CpG is on the forward or reverse strand
mC	Count of methylated cytosines
C	Count of cytosines that are **not** called as mC (including those called as hmC and those called as undifferentiated modC)(*)
Dinucleotide context	The reference dinucleotide
Trinucleotide context	The reference trinucleotide

Fields featured in the **duet evoC hmC** cytosine report are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Position	The genomic position of the cytosine (one-indexed)
Strand	Whether the CpG is on the forward or reverse strand
hmC	Count of hydroxymethylated cytosines
C	Count of cytosines that are **not** called as hmC (including those called as mC and those called as undifferentiated modC)(*)
Dinucleotide context	The reference dinucleotide
Trinucleotide context	The reference trinucleotide

Fields featured in the **duet evoC modC** cytosine report are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Position	The genomic position of the cytosine (one-indexed)
Strand	Whether the CpG is on the forward or reverse strand
modC	Count of modified cytosines including those that are called as mC, hmC, and undifferentiated modC
C	Count of unmodified cytosines (*)
Dinucleotide context	The reference dinucleotide
Trinucleotide context	The reference trinucleotide

Note: The genomic positions in a cytosine report are one-indexed.

BedMethyl duet +modC quantification file

Fields featured in the **duet +modC** bedMethyl file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
modC	Modification
Score	A score that represents the coverage, but capped at 1000 (*)
Strand	Whether the CpG is on the forward or reverse strand
Rendering start position	The genomic start position used for rendering in IGV
Rendering end position	The genomic end position used for rendering in IGV
RGB Code	RGB (Red/Green/Blue) encoding for rendering in IGV
Coverage	Coverage of cytosines at this position (*)
Modification percentage	The percentage of C (*) calls reported as modified

Note: The genomic positions in a bedMethyl file are zero-indexed.

The following table shows how the colour rendering of a modification will appear when loaded in IGV:

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

BedMethyl duet evoC quantification file

Fields featured in the **duet evoC mC** bedMethyl file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
mC	Modification
Score	A score that represents the coverage, but capped at 1000 (*)
Strand	Whether the CpG is on the forward or reverse strand
Rendering start position	The genomic start position used for rendering in IGV
Rendering end position	The genomic end position used for rendering in IGV
RGB Code	RGB (Red/Green/Blue) encoding for rendering in IGV
Coverage	Coverage of cytosines at this position (*)
Modification percentage	The percentage of C (*) calls with the given modification (i.e. mC)

Fields featured in the **duet evoC hmC** bedMethyl file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
hmC	Modification
Score	A score that represents the coverage, but capped at 1000 (*)
Strand	Whether the CpG is on the forward or reverse strand
Rendering start position	The genomic start position used for rendering in IGV
Rendering end position	The genomic end position used for rendering in IGV
RGB Code	RGB (Red/Green/Blue) encoding for rendering in IGV
Coverage	Coverage of cytosines at this position (*)
Modification percentage	The percentage of C (*) calls with the given modification (i.e. hmC)

Fields featured in the **duet evoC modC** bedMethyl file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
modC	Modification
Score	A score that represents the coverage, but capped at 1000 (*)
Strand	Whether the CpG is on the forward or reverse strand
Rendering start position	The genomic start position used for rendering in IGV
Rendering end position	The genomic end position used for rendering in IGV
RGB Code	RGB (Red/Green/Blue) encoding for rendering in IGV
Coverage	Coverage of cytosines at this position (*)
Modification percentage	The percentage of C (*) calls with the given modification (i.e. mC or hmC or undifferentiated modC)

Note: The genomic positions in a bedMethyl file are zero-indexed.

Bedgraph duet +modC quantification file

Fields featured in the **duet +modC** bedgraph file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls (*) reported as modified

Note: Forward strand modifications are given positive modification percentage and reverse strand modifications are given negative modification percentage. This helps to visualise the difference between forward strand and reverse strand methylation in IGV.

Note: The genomic positions in a bedgraph file are zero-indexed.

Bedgraph duet evoC quantification file

Fields featured in the **duet evoC mC** bedgraph file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls (*) reported as mC

Fields featured in the **duet evoC hmC** bedgraph file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls (*) reported as hmC

Fields featured in the **duet evoC modC** bedgraph file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls (*) reported as modC (i.e. mC or hmC or undifferentiated modC)

Note: Forward strand modifications are given positive modification percentage and reverse strand modifications are given negative modification percentage. This helps to visualise the difference between forward strand and reverse strand methylation in IGV.

Note: The genomic positions in a bedgraph file are zero-indexed.

Bismark duet +modC quantification file

Fields featured in the **duet +modC** Bismark file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls reported as modified
modC	Count of modified cytosines
C	Count of unmodified cytosines (*)

Note: The genomic positions in a Bismark file are zero-indexed.

Bismark duet evoC quantification file

Fields featured in the **duet evoC mC** Bismark file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls reported as mC
mC	Count of modified cytosines
C	Count of cytosines that are **not** called as mC (including those called as hmC and those called as undifferentiated modC) (*)

Fields featured in the **duet evoC hmC** Bismark file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls reported as hmC
hmC	Count of modified cytosines
C	Count of cytosines that are **not** called as hmC (including those called as mC and those called as undifferentiated modC) (*)

Fields featured in the **duet evoC modC** Bismark file are:

Field	Description
Contig	The contig / chromosome featuring the CpG
Start position	The genomic start position of the site (zero-indexed)
End position	The genomic end position of the site (zero-indexed)
Modification percentage	The percentage of C calls reported as modified
modC	Count of modified cytosines including those that are called as mC, hmC, and undifferentiated modC
C	Count of unmodified cytosines (*)

Note: The genomic positions in a Bismark file are zero-indexed.

Zarr store

A Zarr store is a chunked, compressed storage format optimised for storing and accessing large, multidimensional arrays. By default, the Duet pipeline generates a single multi-sample Zarr store that contains epigenetic quantification data for all samples in a pipeline run. This file is output to the `sample_outputs/zarr_store/` subdirectory.

5.3.4 Allele-specific methylation (ASM) file format

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

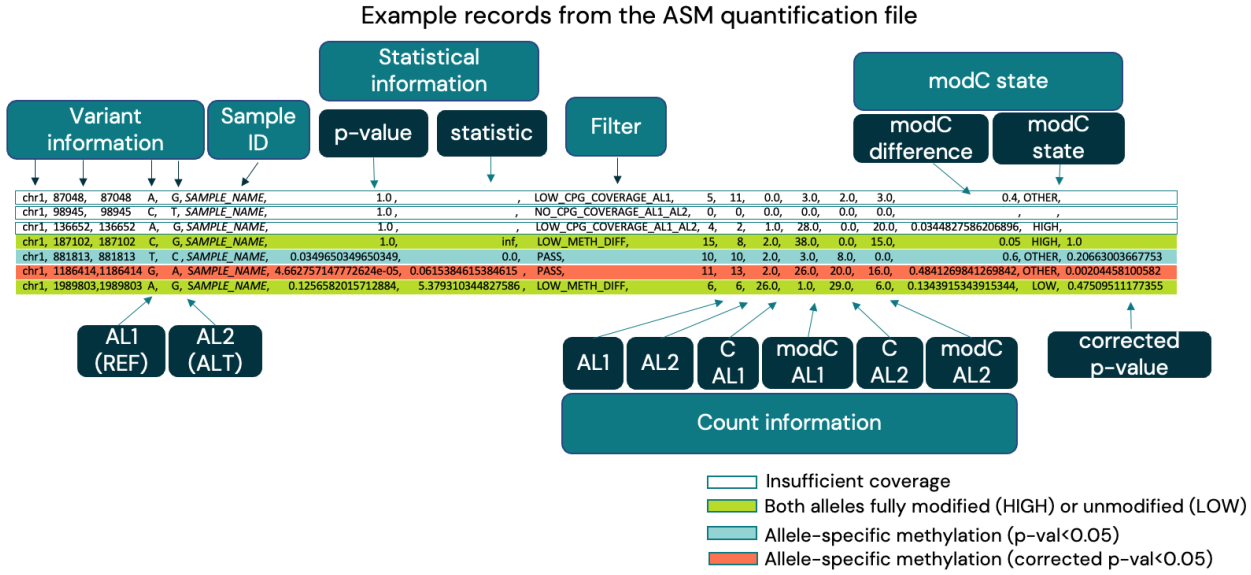


Fig. 7: Figure 10. Breakdown of a biomodal ASM file

- **contig**: The genomic contig (chromosome) where the SNV or haplotype is located
- **pos_first_variant**: The genomic position of the SNV, or of the first variant in the haplotype
- **pos**: In the case of a single SNV, this is the same as `pos_first_variant`; in the case of a haplotype, this is a list of positions of all the variants in the haplotype
- **ref_allele**: The reference allele (individual SNV or haplotype) from the VCF file
- **alt_allele**: The alternate allele (individual SNV or haplotype) from the VCF file
- **sample_id**: The ID of the sample from which the variant was called
- **p_value and test_statistic**: These fields report the results of a statistical test measuring the confidence of the ASM call. By default, the statistical test is the odds ratio of the Fisher's Exact Test. This value is then corrected for multiple hypothesis testing using Benjamini/Hochberg correction, to generate a corrected p-value.
- **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 - it is recommended that only sites with sufficient coverage are analysed ("PASS" or "LOW_METH_DIFF"):

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 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).
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 modified Cs, and no CpG contains unmodified C; in the 11 reads supporting allele G, a total of 2 CpGs contain unmodified Cs and 3 CpGs contain modified Cs.

  Note: 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 a methylation state of 'OTHER' is a pre-requisite for it being considered a possible ASM site.

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 will more stringently reduce false discoveries.

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 command-line tools such as `grep` and `awk`.

5.3.5 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 output 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 output 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.

5.4 Metrics

5.4.1 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 ID 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.

Additionally, there is an interactive HTML report with a filename `{run_name}_duet-{A}_Summary.html` (where `{A}` is either `modC` or `evoC`) in the `reports/summary_reports/` subdirectory that allows you to plot any of the metrics from the Excel file across all samples on your sequencing run.

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

5.4.2 Modified cytosine accuracy: control DNA (duet +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
modC sensitivity on fully methylated lambda control	Sensitivity for measuring methylated CpGs calculated from a fully methylated lambda control. The value should be >= 95%
modC specificity on fully unmethylated pUC19 control	Specificity for measuring unmodified CpGs calculated from a fully non-methylated pUC19 control. The value should be >= 99%
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.

5.4.3 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. The value should be >= 95%
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. The value should be >= 99%
Methylated lambda control mC sensitivity	Sensitivity for measuring mC calculated from a fully methylated lambda control. The value should be >= 93%
Methylated lambda control hmC specificity	Specificity for measuring hmC calculated from a fully methylated lambda control.
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.

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. The value should be >= 95%

5.4.4 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. The value should be >= 99.9%
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.
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.

5.4.5 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. This 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.

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

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

5.4.6 Genome duplication and 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.
Mean coverage including Ns	The mean genome-wide coverage (including Ns)
Mean coverage excluding Ns	The mean genome-wide coverage (excluding Ns)
Genome percentage no coverage	Percentage of the genome with no coverage (including Ns)
Genome percentage 1x	Percentage of the genome covered at 1X or above (including Ns)
Genome percentage 2x	Percentage of the genome covered at 2X or above (including Ns)
Genome percentage 5x	Percentage of the genome covered at 5X or above (including Ns)
Genome percentage 10x	Percentage of the genome covered at 10X or above (including Ns)
Genome percentage 25x	Percentage of the genome covered at 25X or above (including Ns)
Genome percentage 30x	Percentage of the genome covered at 30X or above (including Ns)
CpG to genome-wide coverage ratio	Ratio of mean coverage at CpGs to mean coverage genome-wide (including Ns). 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.

5.4.7 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	Total number of read-pairs (R1/R2 read-pairs) that remain after trimming
Bases after trimming	Total number of bases that remain after trimming
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.
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.

5.4.8 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.
Reads discarded during trimming	Number of read-pairs filtered out after trimming as too short or having too many N
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	The fraction of the total input reads in the initial input FASTQ files that is remaining after initial trimming in the prelude module.
Percentage of input bases remaining after trimming	The fraction of the total input bases in the initial input FASTQ files that is remaining after initial trimming in the prelude module.

5.4.9 Targeted metrics

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, within 500np, of a baited region.
Fold-80 base penalty	A measure of capture uniformity, the fold increase in coverage necessary to raise 80% of target bases to the mean target coverage level.
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.

5.5 Reports

5.5.1 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:

5.5.2 Qualimap coverage histogram

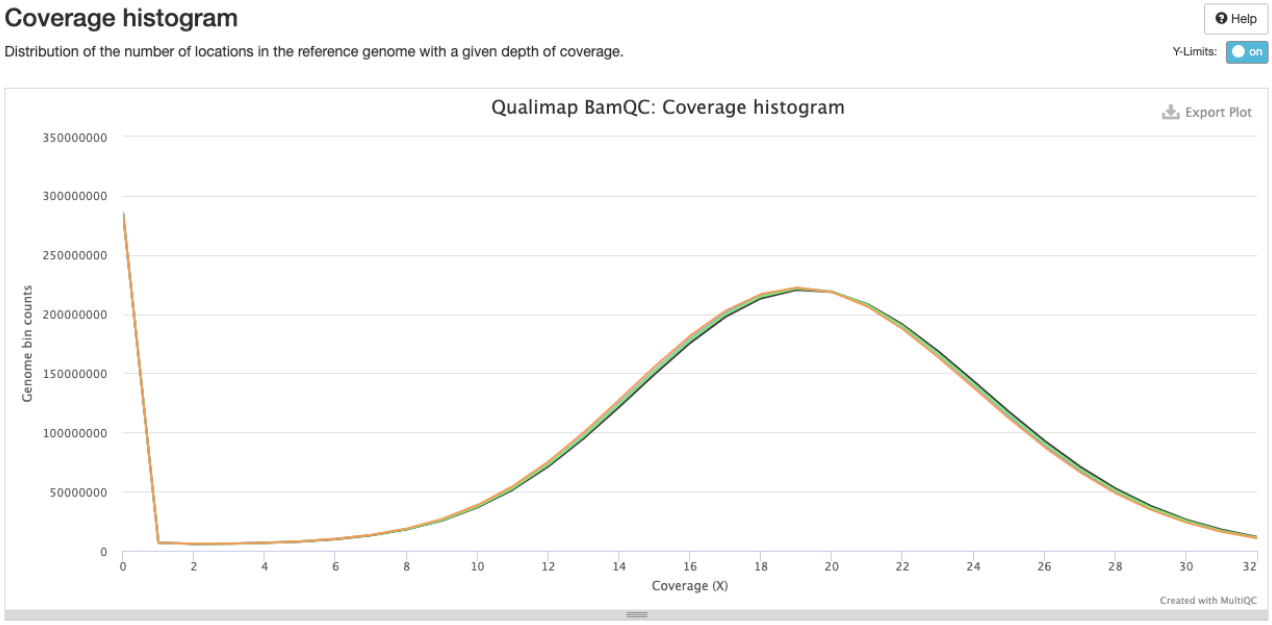


Fig. 8: Qualimap coverage histogram

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

5.5.3 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%).

5.5.4 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%.

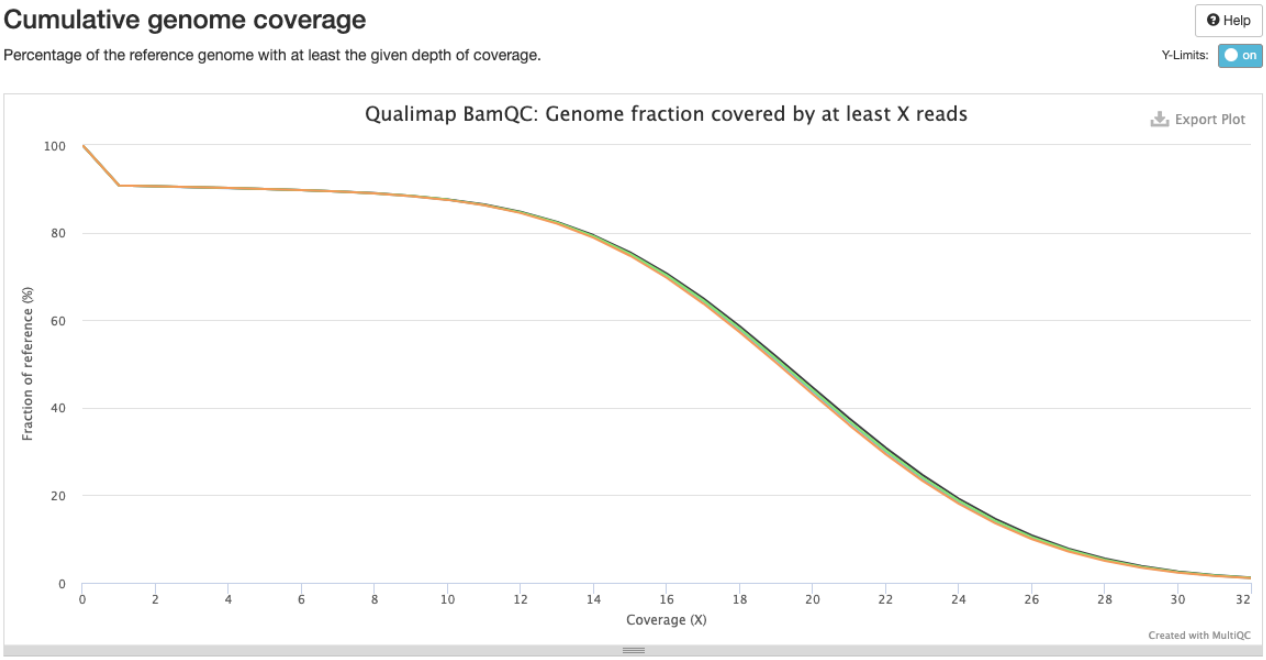


Fig. 9: Cumulative genome coverage

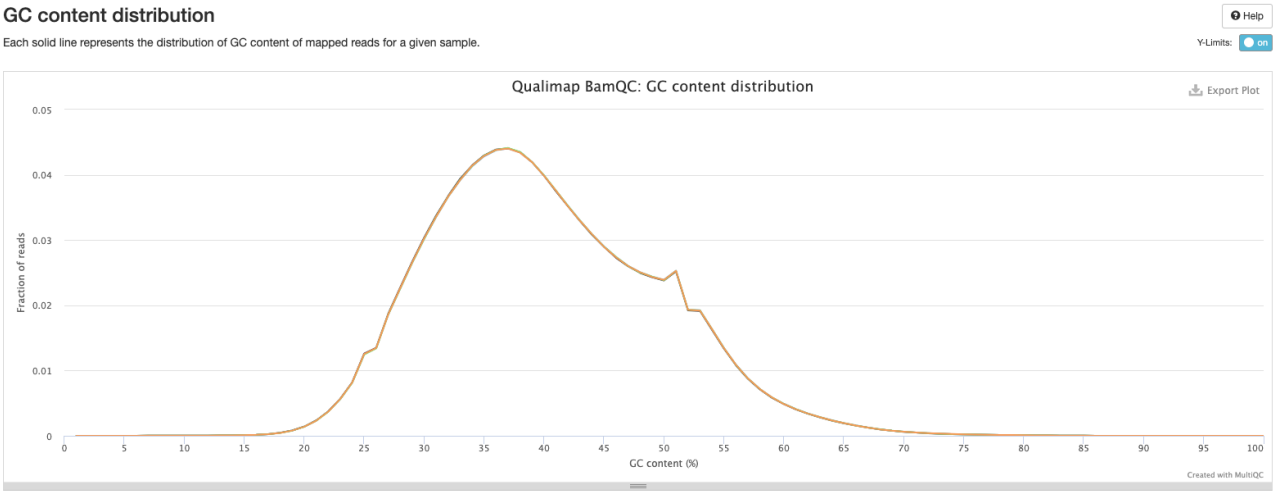


Fig. 10: GC content distribution

5.5.5 Picard GC coverage bias

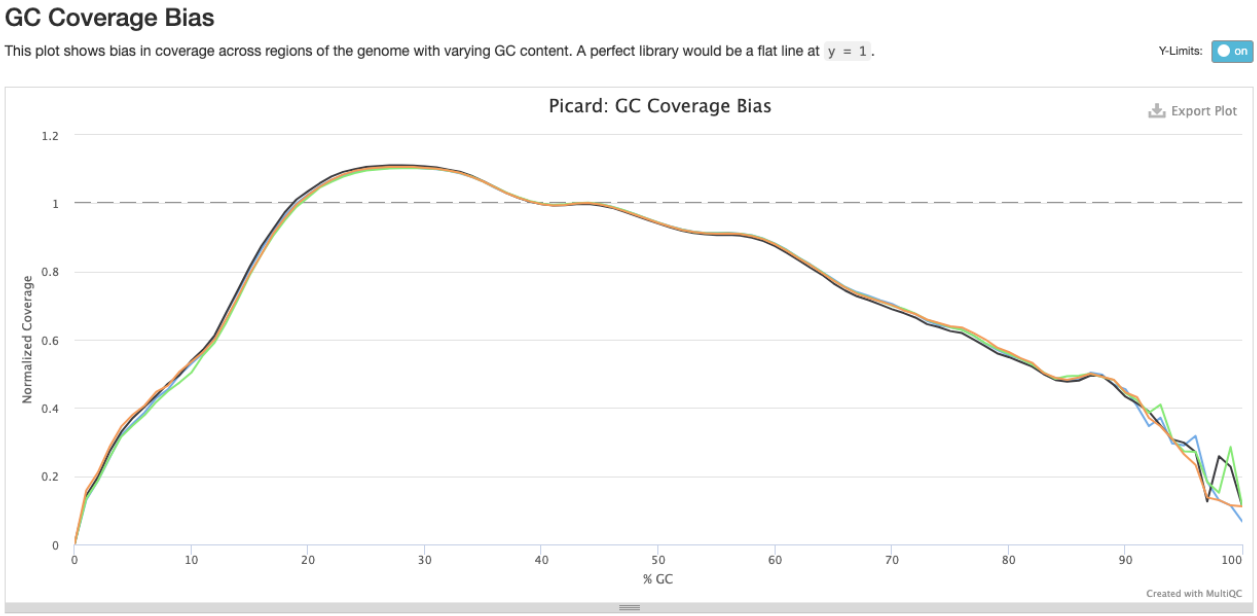


Fig. 11: GC coverage bias

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.

5.5.6 Resolution and alignment stats

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

- **Filtered Short Reads:** This shows the proportion of reads that were filtered out due to being too short after the trimming of potential artefacts
- **Discarded Unresolved 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 assay
- **Non-primary Assembly Reads:** This shows the proportion of reads that were filtered out because they aligned to contigs that were in the reference genome, but did not feature in the primary assembly. Contigs that feature in the reference genome, but not in the primary assembly are typically decoy contigs, such as the sequence of the Epstein-Barr Virus which is commonly included as a decoy in human reference genomes.
- **Duplicate Reads Mapped to Genome:** This shows the proportion of all reads that mapped to the genome and were identified and removed as duplicates (duplicates may be either PCR duplicates or sequencing duplicates).
- **Reads Uniquely 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.

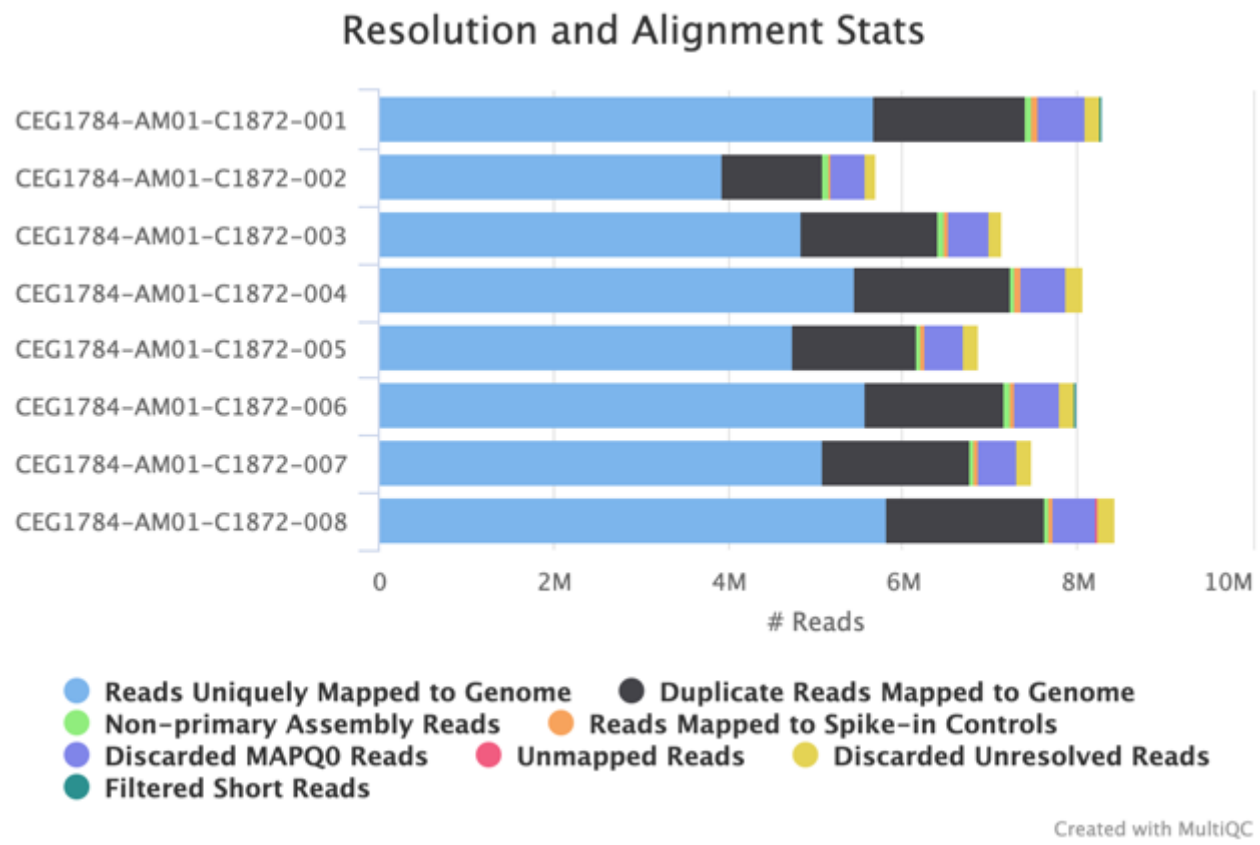


Fig. 12: Resolution and Alignment Stats

5.5.7 Epigenetic accuracy on spike-in controls

duet +modC

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 of base calls at CpG sites that were 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%.

duet evoC

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 of base calls at CpG sites that were reported as:

- Unmodified C
- Modified C (the sum of mC, hmC and undifferentiated modC calls)
- mC
- hmC

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

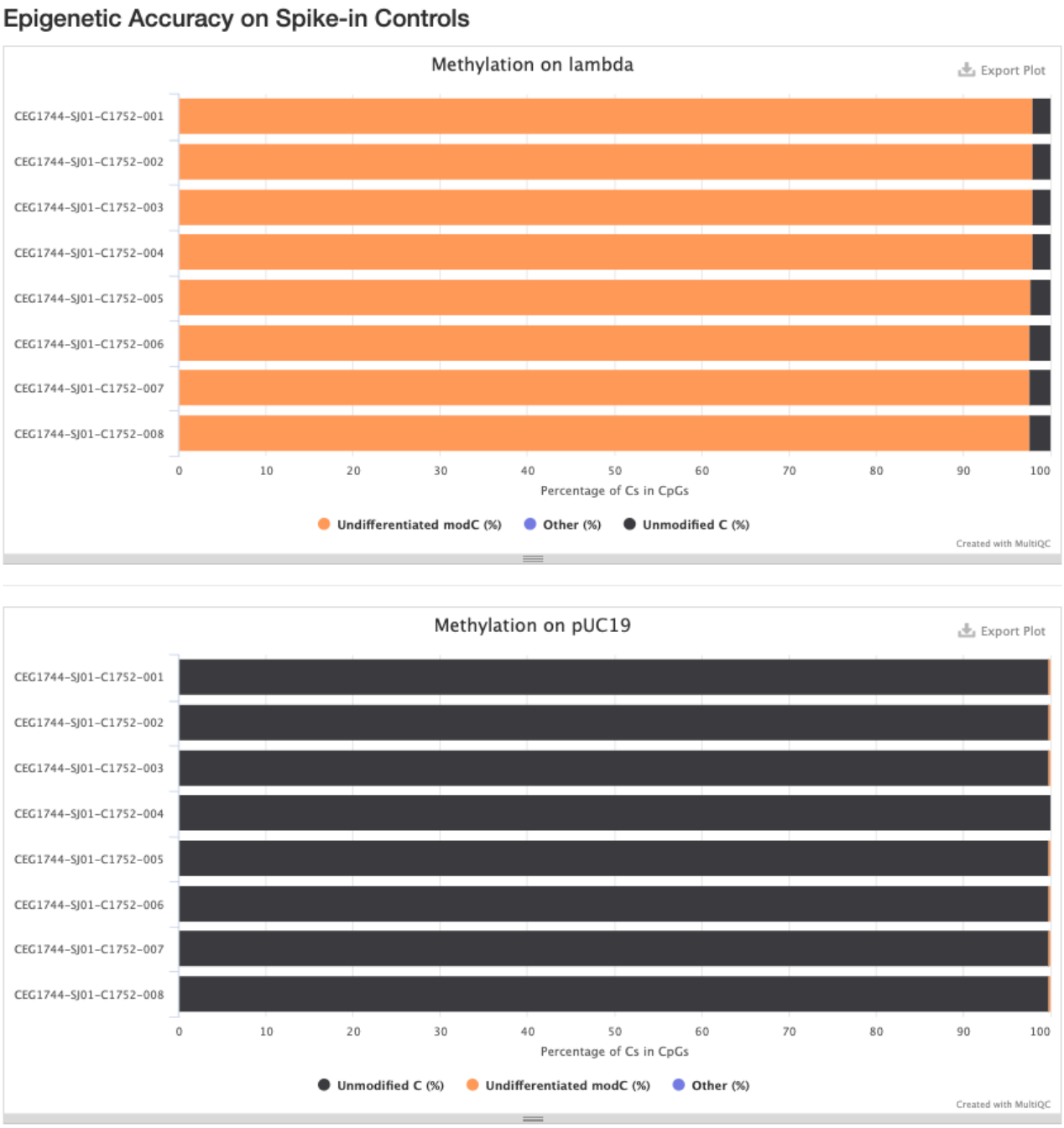
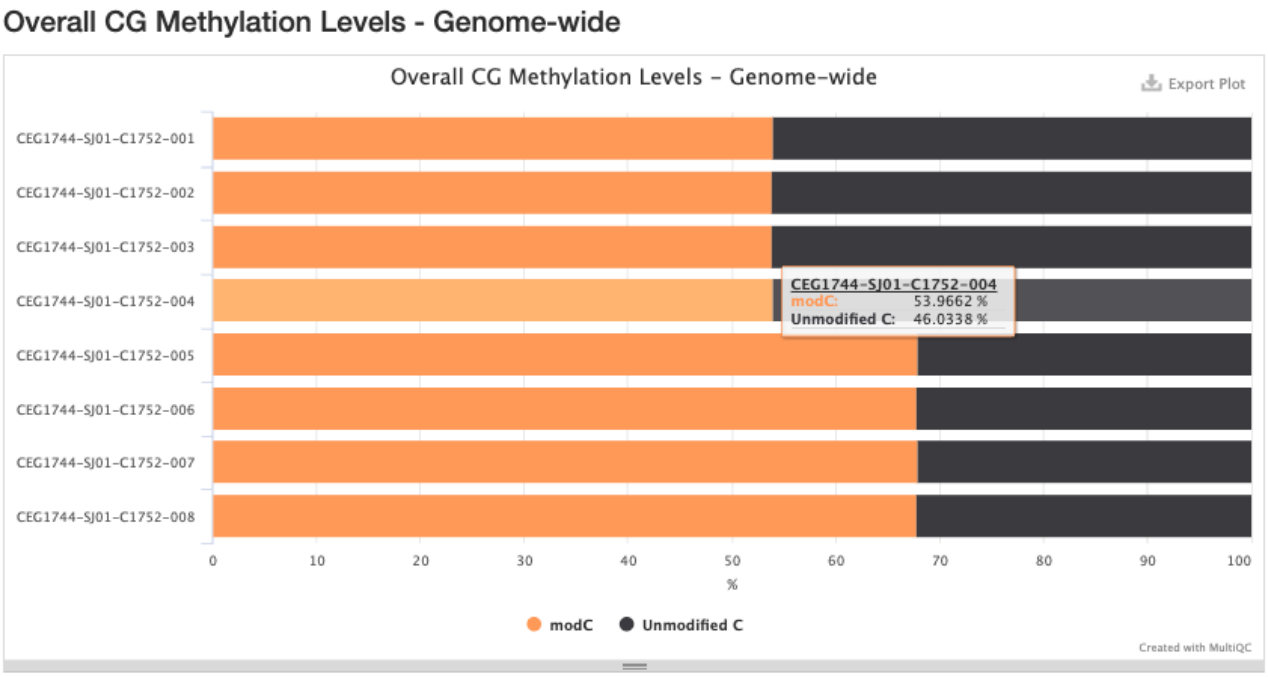


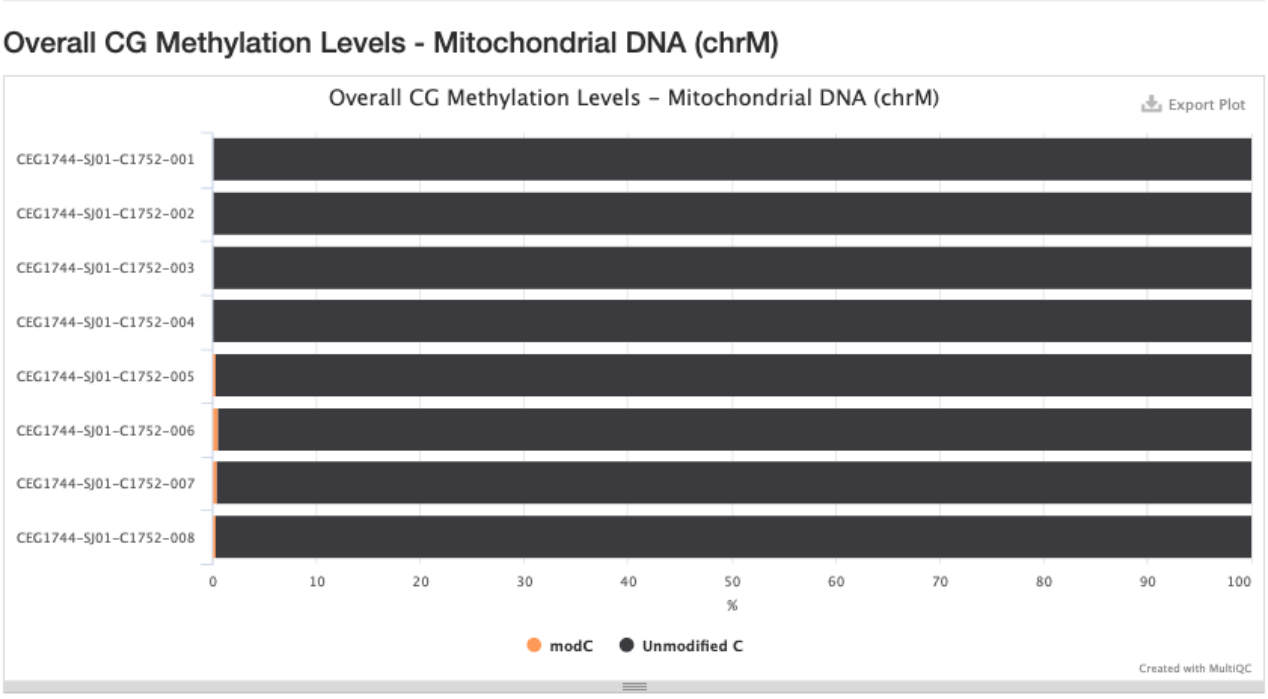
Fig. 13: duet +modC

Fig. 14: duet evoC

5.5.8 Overall CG methylation levels

duet +modC



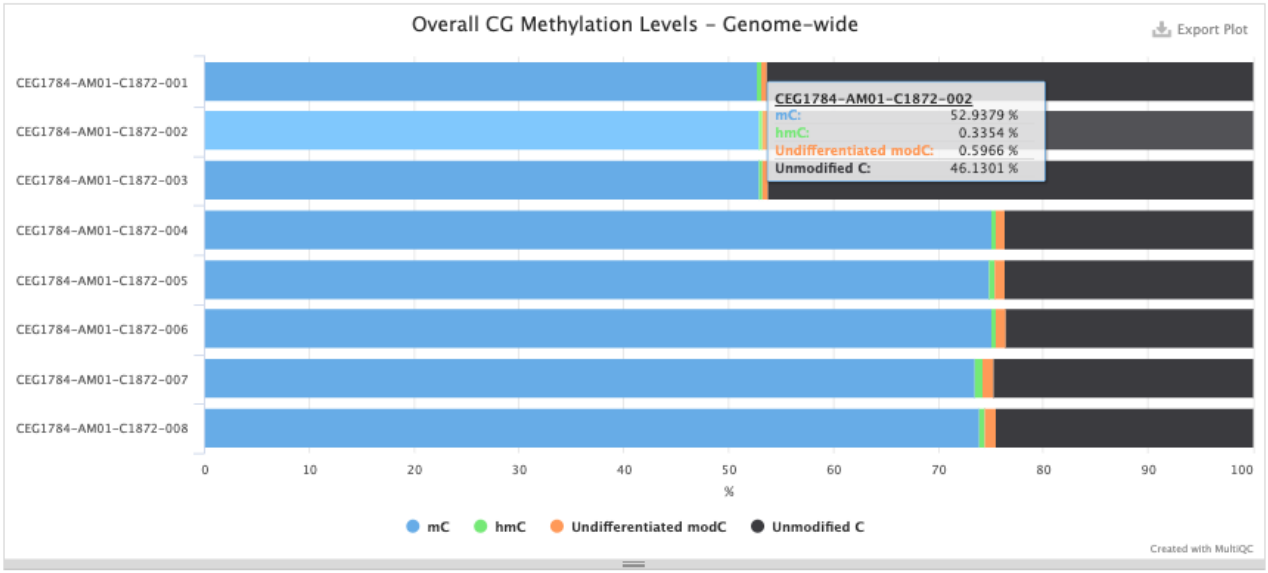


duet evoC

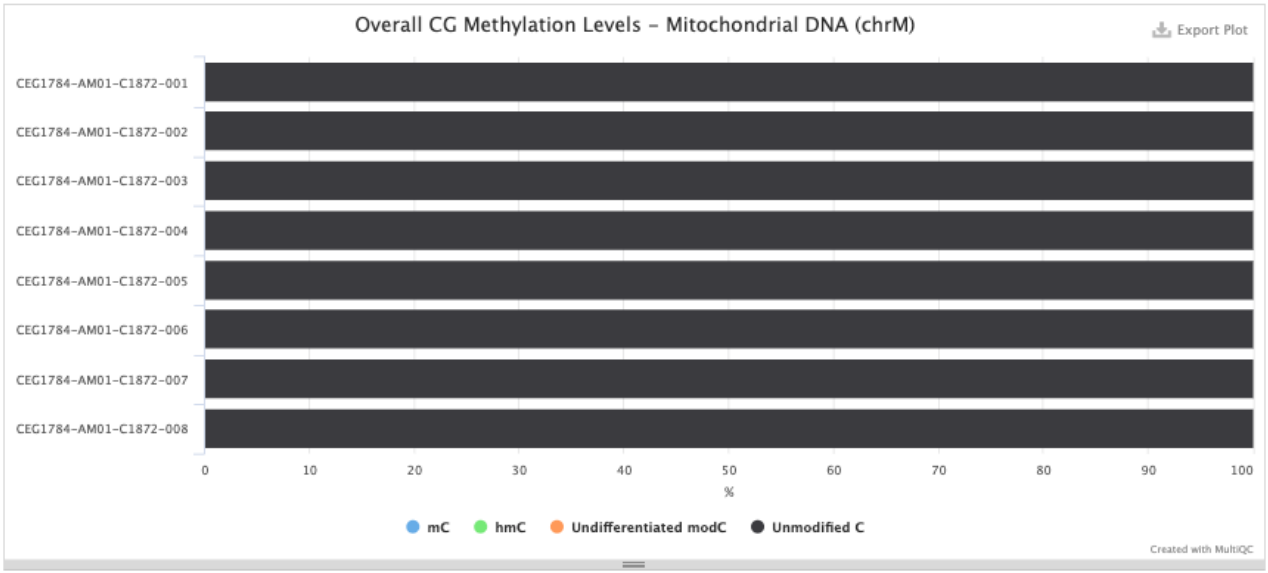
The genome-wide plot shows the proportion of CpGs in the autosomes that are reported as modified or as unmodified. This gives an indication of the overall methylation rate across the genome. The allosomes are excluded to avoid any gender-bias.

The mtDNA plot 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.

Overall CG Methylation Levels - Genome-wide



Overall CG Methylation Levels - Mitochondrial DNA (chrM)



5.5.9 M-bias

This plot is the mC plot from a duet evoC run. In duet evoC, there will be also be an m-bias plot for hmC. In duet +modC, there will only be one plot, which will represent modC calls.

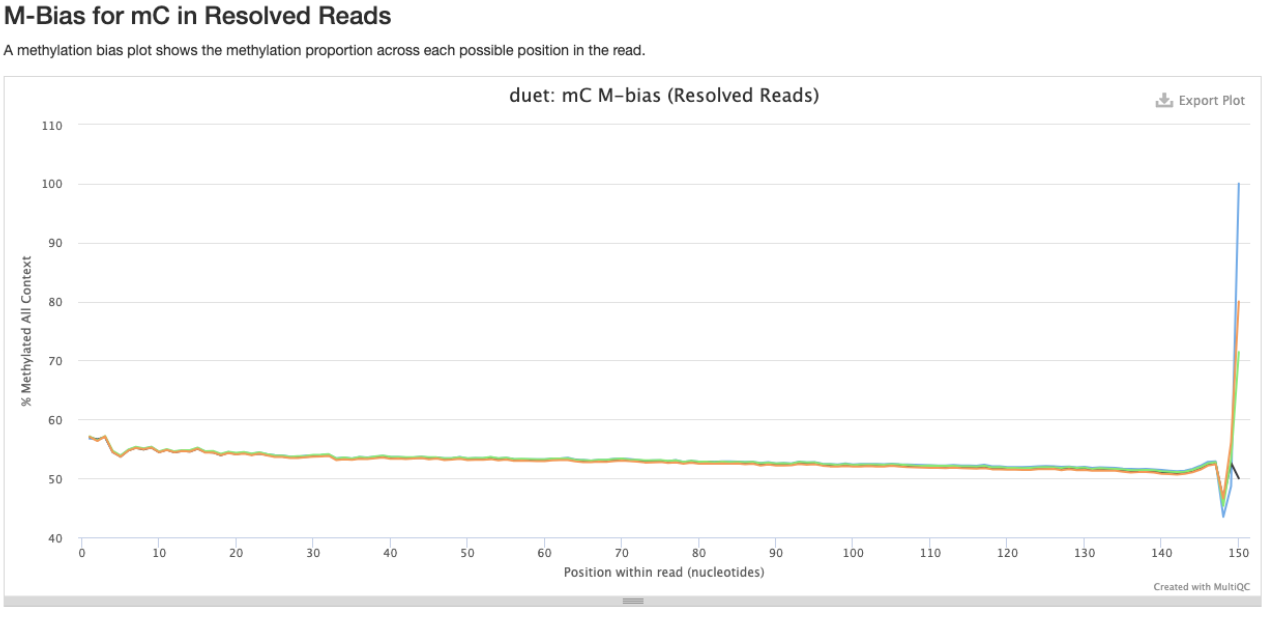


Fig. 15: M-bias for mC in Resolved Reads

On this plot, a line for each sample presents the rate of mC (left-hand y-axis) at CpG sites by read cycle (x-axis) in the genome-aligned reads.

5.5.10 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/` subdirectory. 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:

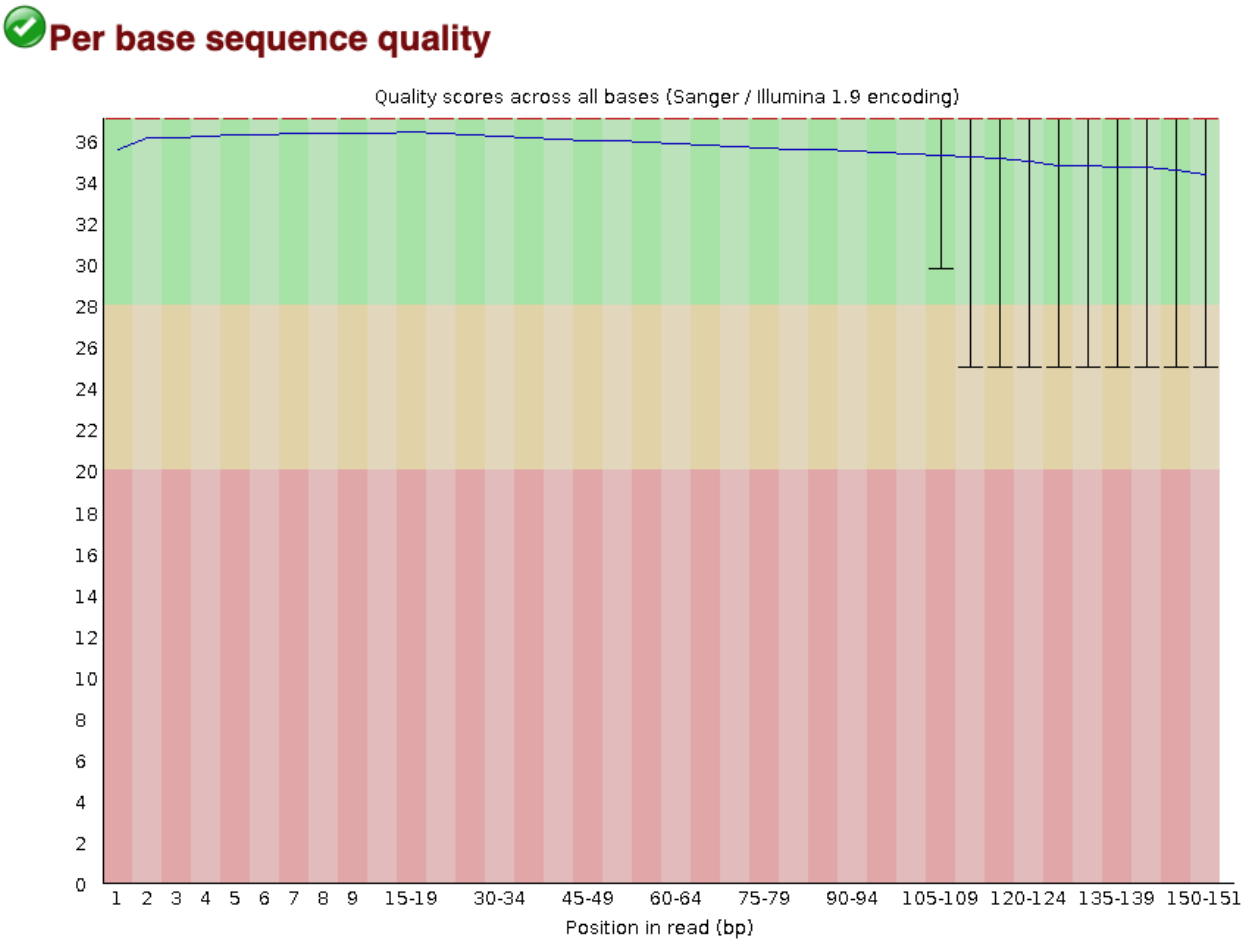


Fig. 16: Per base sequence quality

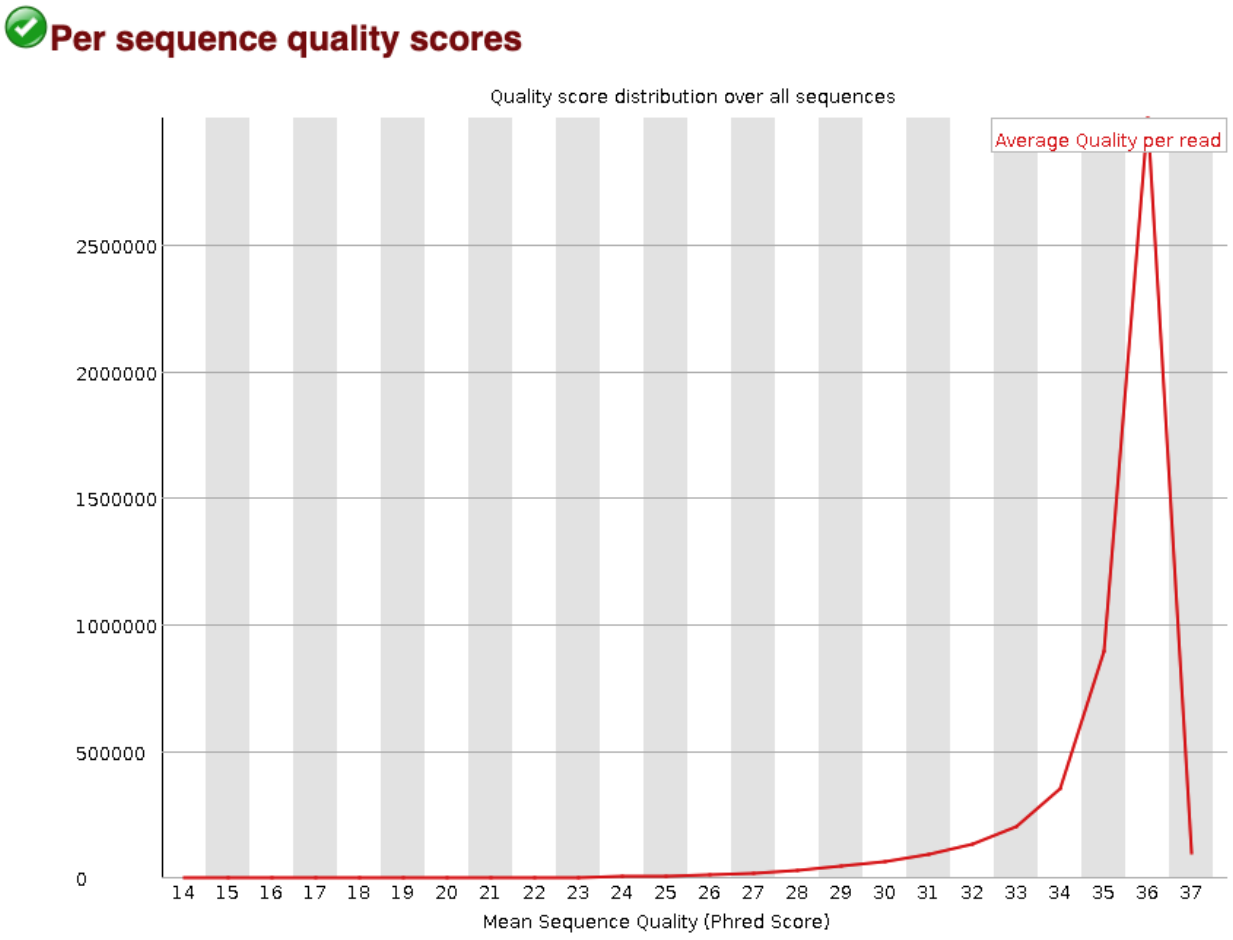


Fig. 17: Per sequence quality scores

Note: 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:

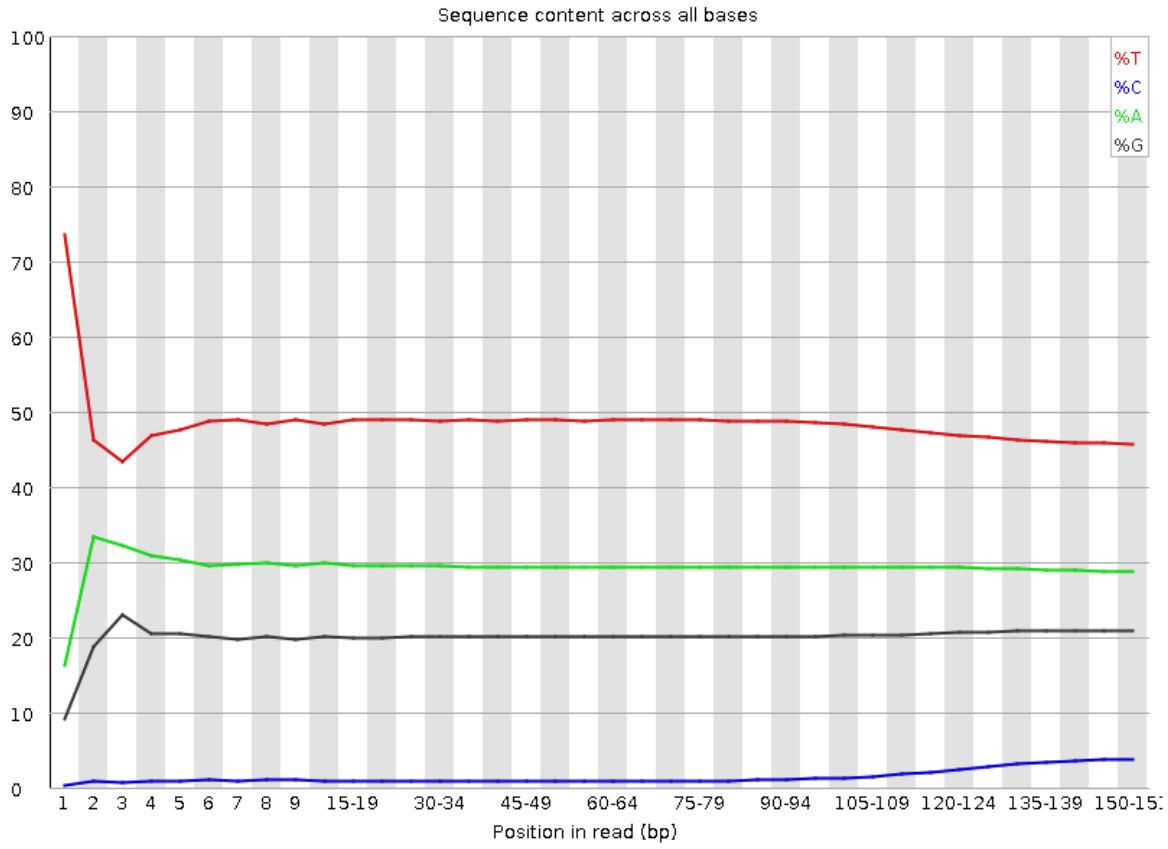


Fig. 18: Per base sequence content 1

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.

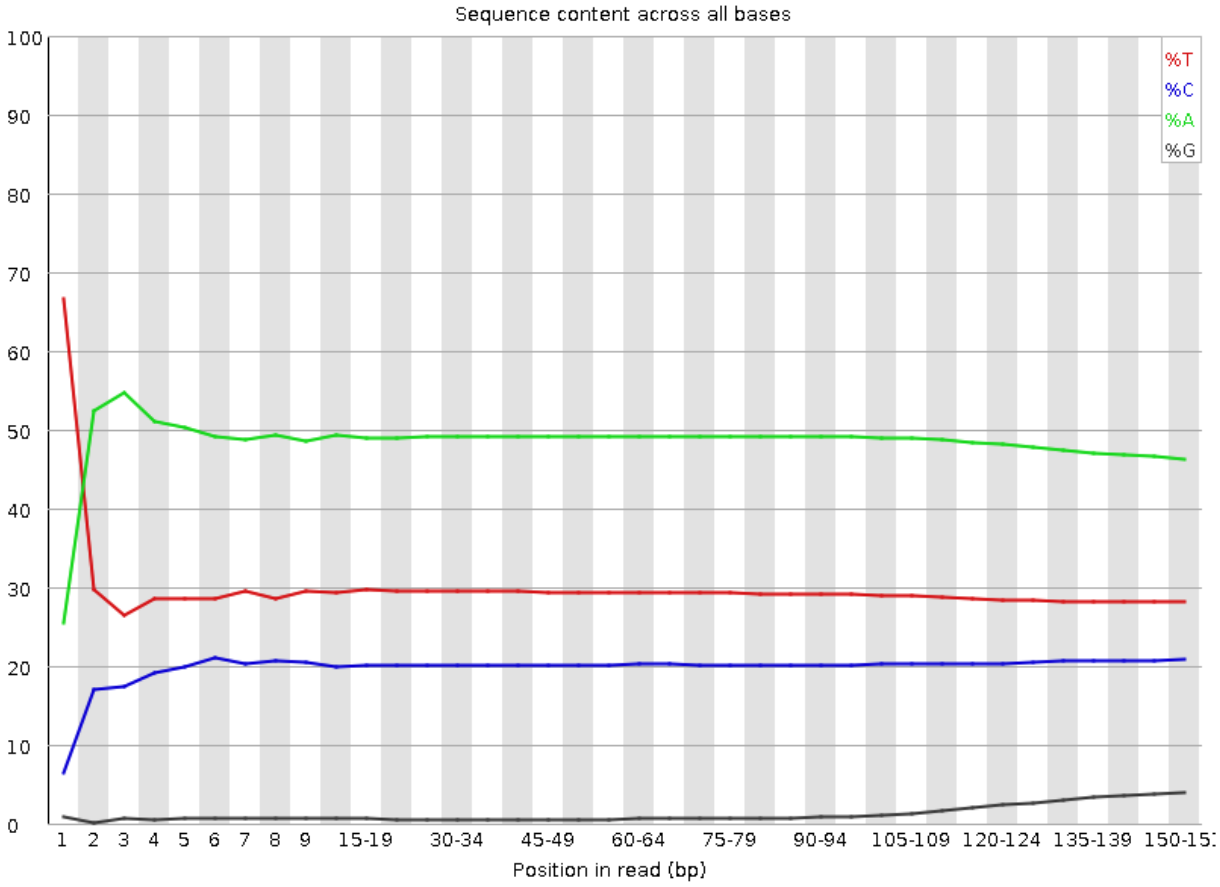


Fig. 19: Per base sequence content 2

CHAPTER

SIX

LEGACY DOCUMENTATION

For recent versions of the duet pipeline, legacy documentation is available through the "flyout" menu in the bottom right hand corner of the screen.

Older versions are available here:

- `Pipeline v1.3.0`
- `Pipeline v1.2.2`

CHAPTER

SEVEN

RELEASE NOTES

Links to release notes for pipeline and biomodal CLI versions are linked below.

Release version	Date
`pipeline 1.5.0 / CLI 2.0.0`	December 2025
`pipeline 1.4.2 / CLI 1.1.3`	March 2025
`pipeline 1.4.1 / CLI 1.1.2`	November 2024
`pipeline 1.4.0 / CLI 1.1.1`	November 2024
`pipeline 1.3.0 / CLI 1.1.0`	June 2024
`pipeline 1.2.2 / CLI 1.0.5`	May 2024
`pipeline 1.2.1 / CLI 1.0.5`	April 2024
`pipeline 1.2.0 / CLI 1.0.4`	February 2024
`pipeline 1.1.2 / CLI 1.0.3`	December 2023