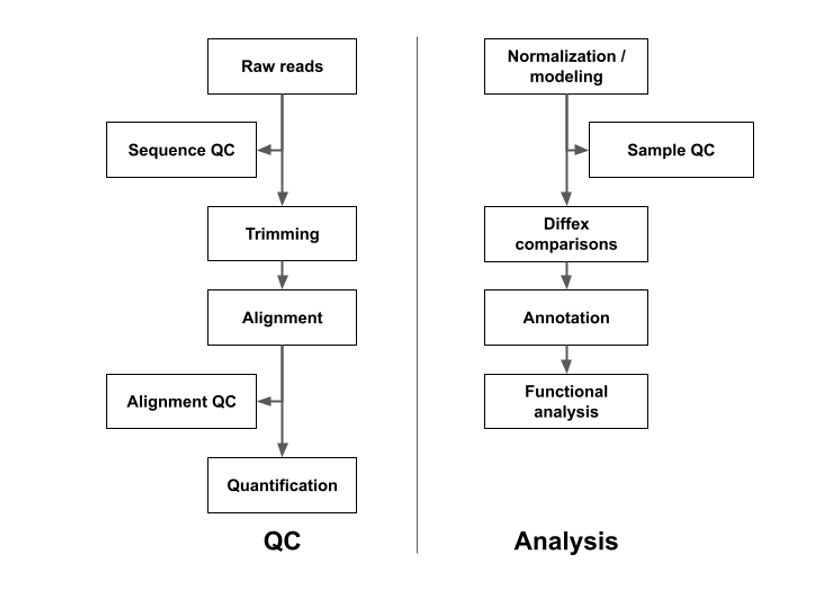
Analysis methods

UM Bioinformatics Core

2023-08-29

Table of Contents

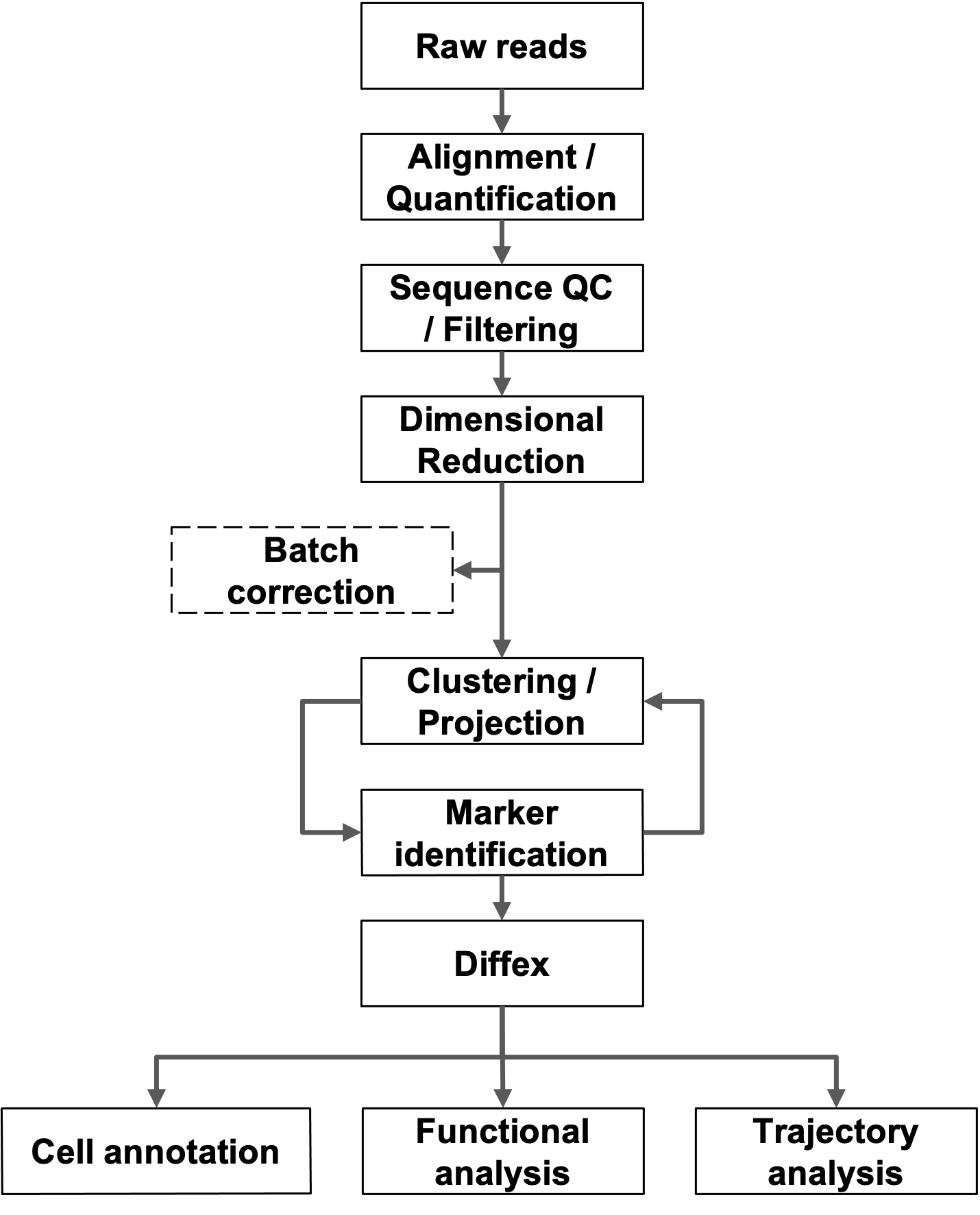
# Bulk RNA-Seq



RNA-Seq workflow

Adapters will be trimmed with CutAdapt (Martin) to remove adapters and low-quality bases and overall quality will assessed with FastQC (FastQC), FastQCScreen (Wingett and Andrews, 2018), and MultiQC (Ewels *et al.*, 2016). Reads will be aligned and quantified using RSEM/STAR (Li and Dewey, 2011; Dobin *et al.*, 2013). Differential expression modeling will use the DESeq2 (Love *et al.*, 2014). Intra and inter group variance will be assessed with Principal Component Analysis. Candidate DEGs are visualized with volcano plots filtered to the subset of DEGs where Benjamini-Hochberg adjusted p-value ≤ 0.05 (Benjamini and Hochberg, 1995). Functional analysis using Advaita iPathwayGuide (Draghici *et al.*, 2007; iPathwayGuide) will assess enrichment of KEGG pathways and Gene Ontology (GO) concepts.

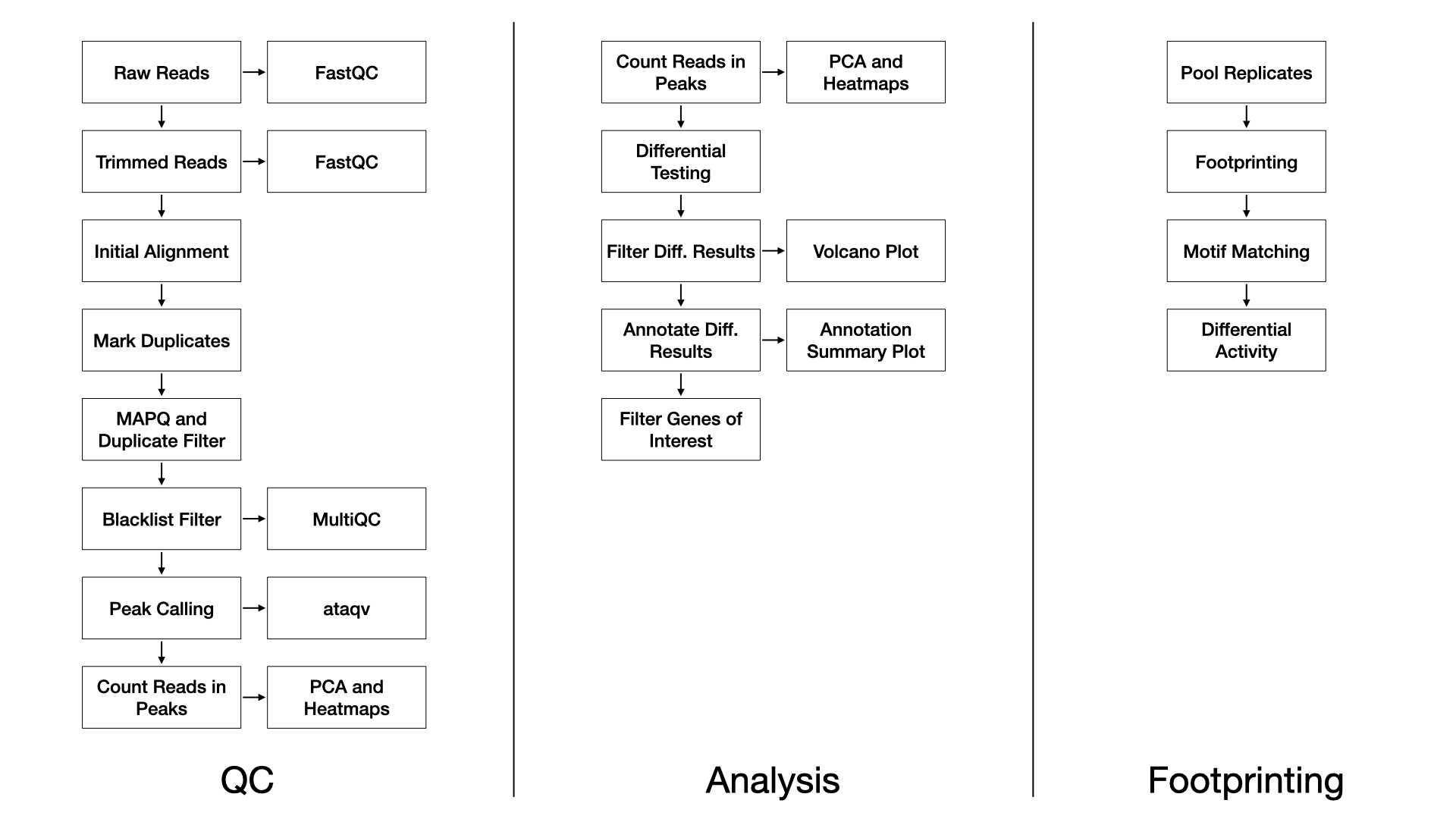
# Single cell/single nuclei RNA-Seq



Single cell/nuclei RNA-Seq Workflow

Raw sequence data is transformed to gene-barcode count matrices using CellRanger (Zheng *et al.*, 2017). Further data analysis is primarily performed using the Seurat package (Stuart *et al.*, 2019). Gene-barcode matrices and metadata for each sample are loaded and further filtering and clustering analyses were performed as described in the Seurat tutorials (Seurat tutorials). Aberrant cells are filtered (low complexity, duplets, or apoptotic cells) and based on detected debris/contamination DecontX (Yang *et al.*, 2020) may be run. Counts are normalized using the default normalization approach and variable features were identified. Where appropriate, anchor points were then generated across related datasets and used for SCTransform data integration. Principal component analysis (PCA) is then performed to identify significant PCA components used to find nearest neighbors followed by graph-based, semi- unsupervised clustering into distinct populations. Projections are generated using uniform manifold approximation (Becht *et al.*, 2018) and marker genes are identified through differential gene expression pairwise comparisons (Wilcoxon rank-sum test for single-cell gene expression; FindAllMarkers function). Cell-type predictions were also generated with scCATCH (Shao *et al.*, 2020). Suitability and approach for trajectory analysis is determined based on experimental design and disease model (Saelens *et al.*, 2019; Lange *et al.*, 2022; Cao *et al.*, 2019; Wolf *et al.*, 2019; La Manno *et al.*, 2018; Bergen *et al.*, 2020).

# Bulk ATAC-Seq



ATAC-Seq workflow

### QC

We use FastQC (FastQC) (v0.11.8) to assess the overall quality of each sequenced sample.

We use TrimGalore (TrimGalore) (v0.4.5) and cutadapt (Martin) (v1.15) with the following parameters: --nextera -e 0.1 --stringency 6 --length 20 --nextseq 20.

We align trimmed reads to the reference genome with Bowtie2 (Langmead and Salzberg, 2012) (v2.3.4.1) with the following parameters: -X 2000 --no-mixed --no-discordant, and defaults multi-seed length of 20bp with 0 mismatches.

Duplicate reads are marked with Picard (Picard) (v2.20.2).

Alignments to autosomes and sex chromosomes are kept (i.e. mitochondrial reads are removed), duplicates marked by Picard are removed, and alignments below a MAPQ threshold are removed. These filtering steps are performed with samtools (Li *et al.*, 2009) (v1.2) and the parameters: -q 10 -F 1024.

Reads completely overlapping blacklisted regions (ENCODE Blacklist Regions) are removed with bedtools (Quinlan and Hall, 2010) (v2.28.0).

Sample-wise peaks are called with macs2 (Zhang *et al.*, 2008) (v2.1.2) with flags: -f BAM --nomodel --shift -100 --extsize 200.

Peaks over all samples are merged with bedops (Neph *et al.*, 2012) (v2.4.36) for the purpose of principal component analysis and unsupervised clustering to assess the similarity of samples.

Finally, MultiQC (Ewels *et al.*, 2016) (v1.7) generates a report combining FastQC, trimming, alignment, and duplicate calling over all the samples. For ATAC specific QC metrics we use ataqv (Ataqv) (v1.0.0).

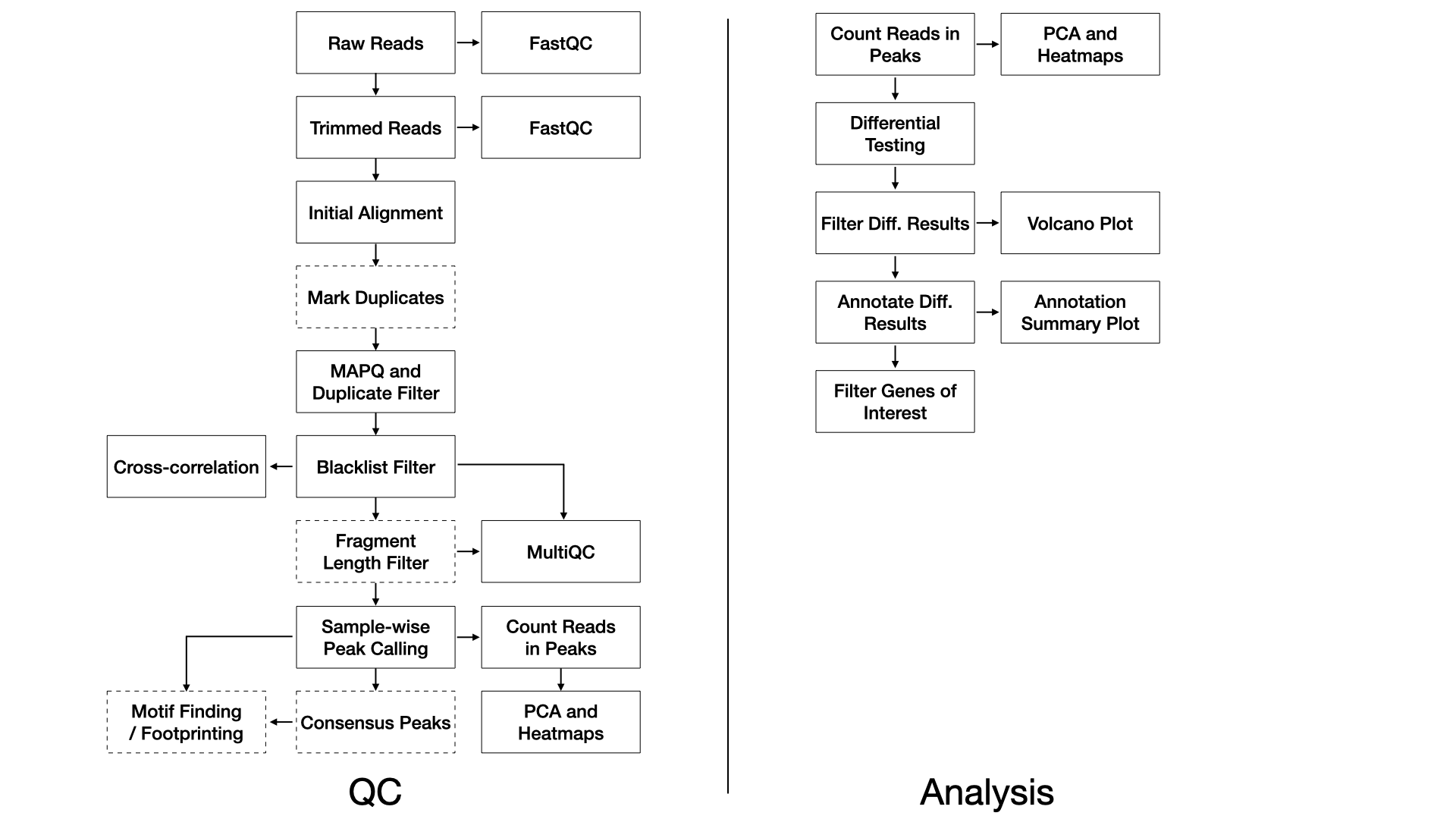
### Differential Testing

We use the edgeR R Bioconductor package (McCarthy *et al.*, 2012) to identify regions of differentially open chromatin (DOC). For each sample, the number of reads in the merged peaks is counted for each sample, and a library size normalization factor is determined. With no replicates, we manually tune the BCV (biological coefficient of variance) parameter. We then fit each model using the glmFit function, and test each contrast with a likelihood ratio test. With replicates, the common, trended, and tagwise negative binomial dispersions are calculated. We then fit each model using the glmQLFit function, and test each contrast with an empirical Bayes quasi-likelihood F-test. The DOC are then annotated to genic and CpG island annotations using the annotatr R Bioconductor package (Cavalcante and Sartor, 2017).

### Footprinting

We use the HINT tool in the Regulatory Genomics Toolbox (rgt-hint) to do footprinting and motif binding analysis (Li *et al.*, 2019). Briefly, for each condition, the reads among the replicates are pooled and tag counts are determined within the merged peaks. Next motifs from the JASPAR database (Khan *et al.*, 2017) are queried against the footprints found within each condition. Finally, a differential score representing the transcription factor binding activity and the openness of the surrounding chromatin is calculated and visualized between pairs of conditions.

# Bulk ChIP-Seq



ChIP-Seq workflow

### QC

We use FastQC (FastQC) (v0.11.8) to assess the overall quality of each sequenced sample.

We use TrimGalore (TrimGalore) (v0.4.5) and cutadapt (Martin) (v1.15) with the following parameters: --nextera -e 0.1 --stringency 6 --length 20 --nextseq 20.

We align trimmed reads to the reference genome with Bowtie2 (Langmead and Salzberg, 2012) (v2.3.4.1) using default parameters with the exception of the following flags: -X 2000.

Duplicate reads may be are marked with Picard (Picard) (v2.20.2).

Alignments are filtered with samtools (Li *et al.*, 2009) (v1.2) using the flags: -q 10 -F 1024.

Alignments completely overlapping blacklisted regions (ENCODE Blacklist Regions) are removed with bedtools (Quinlan and Hall, 2010) (v2.28.0).

Sample-wise peaks are called with macs2 (Zhang *et al.*, 2008) (v2.1.2) with flags: --format BAMPE --gsize mm --qvalue 0.05.

Optionally, consensus peaks for replicates are determined by mutual intersection of sample-wise peaks with bedops –intersect. Consensus peaks are then annotated with annotatr (Cavalcante and Sartor, 2017).

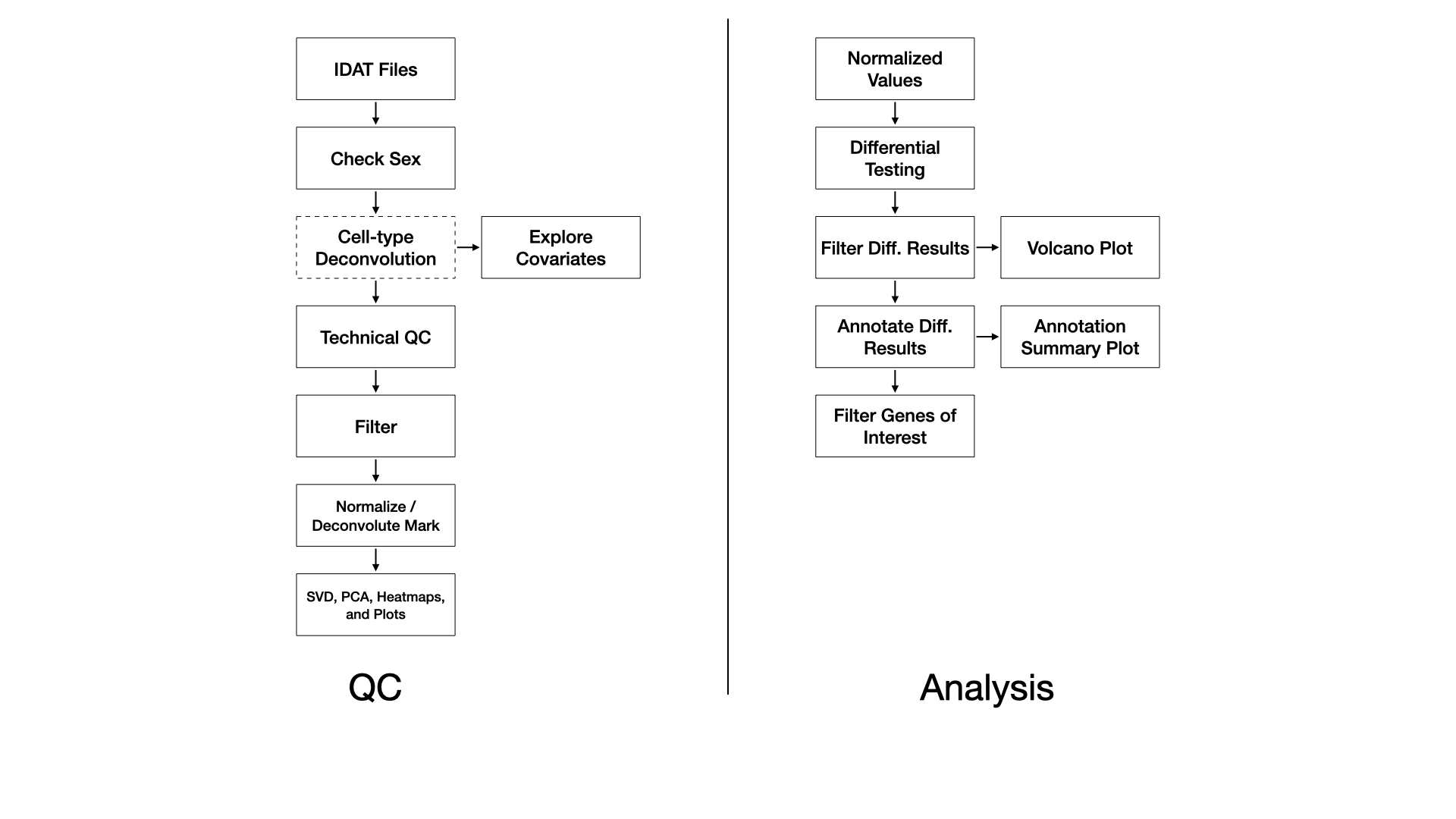
Optionally, motifs for consensus peaks are found with HOMER (Heinz *et al.*, 2010) (v4.11.1).

Peaks over all samples are merged with bedops (Neph *et al.*, 2012) (v2.4.36) for the purpose of principal component analysis and unsupervised clustering to assess the similarity of samples. Read cross-correlations within samples are plotted using phantompeakqualtools (Landt *et al.*, 2012). MultiQC (Ewels *et al.*, 2016) (v1.7) generates a report combining FastQC, trimming, alignment, and duplicate calling over all the samples. The software DeepTools is used to calculate coverage and IP efficiency plots (Ramírez *et al.*, 2016) (v3.3.0).

### Differential Testing

We use the edgeR R Bioconductor package (McCarthy *et al.*, 2012) to identify regions of differential binding. For each sample, the number of reads in the merged peaks is counted for each sample, and a library size normalization factor is determined. With no replicates, we manually tune the BCV (biological coefficient of variance) parameter. We then fit each model using the glmFit function, and test each contrast with a likelihood ratio test. With replicates, the common, trended, and tagwise negative binomial dispersions are calculated. We then fit each model using the glmQLFit function, and test each contrast with an empirical Bayes quasi-likelihood F-test. The differential peaks are then annotated to genic and CpG island annotations using the annotatr R Bioconductor package (Cavalcante and Sartor, 2017).

# DNA Meythlation: EPIC



EPIC methylation array workflow

### QC

Samples are processed with the SeSAMe R Bioconductor package (Zhou *et al.*, 2018). Briefly, red/green IDAT files are read and processed according to the [preparation code](https://zhou-lab.github.io/sesame/v1.18/sesame.html#Preprocessing_Function_Code). Experiment-independent masking occurs with the quality mask code (“Q”), and accounts for probes containing SNPs, with known cross-hybridization issues (Pidsley *et al.*, 2013), and with other issues. Experiment-dependent is done according to the p-value with out-of-bound array hybridization (pOOBAH) algorithm (“P”), which is an improvement on detection p-value filtering (Zhou *et al.*, 2018). Non-linear dye bias correction (“D”) is performed followed by background correction with the NOOB method (“B”) (Fortin *et al.*, 2014). This processing pipeline constitutes a within-array normalization procedure. We note that recent studies have indicated within-array normalization with dye-bias correction and NOOB performs as well or better than between-array normalization procedures (Welsh *et al.*, 2023).

A probe has with pOOBAH masking (p-value < 0.05) in more than 5% of samples it is removed. Similarly, if a sample has more than 10% of probes pOOBAH masked, then that sample is removed.

Cell type deconvolution of whole blood or cord blood may be performed with the FlowSorted.Blood.EPIC or FlowSorted.CordBlood.450k R Bioconductor package, respectively, using a modified version of the Houseman method (Houseman *et al.*, 2012).

In the event of a BS/oxBS or BS/TAB library preparation, methylation mark deconvolution may be performed using the MLML2R R package (Qu *et al.*, 2013). Briefly, methylated and unmethylated channel matrices from bisulfite-only treated samples and oxidative-bisulfte treated samples are extracted and passed to MLML2R::MLML() to determine the levels of methylcytosine (mC), hydroxymethylcytosine (hmC), and cytosine (C) using the exact method provided in the package.

### Differential Testing

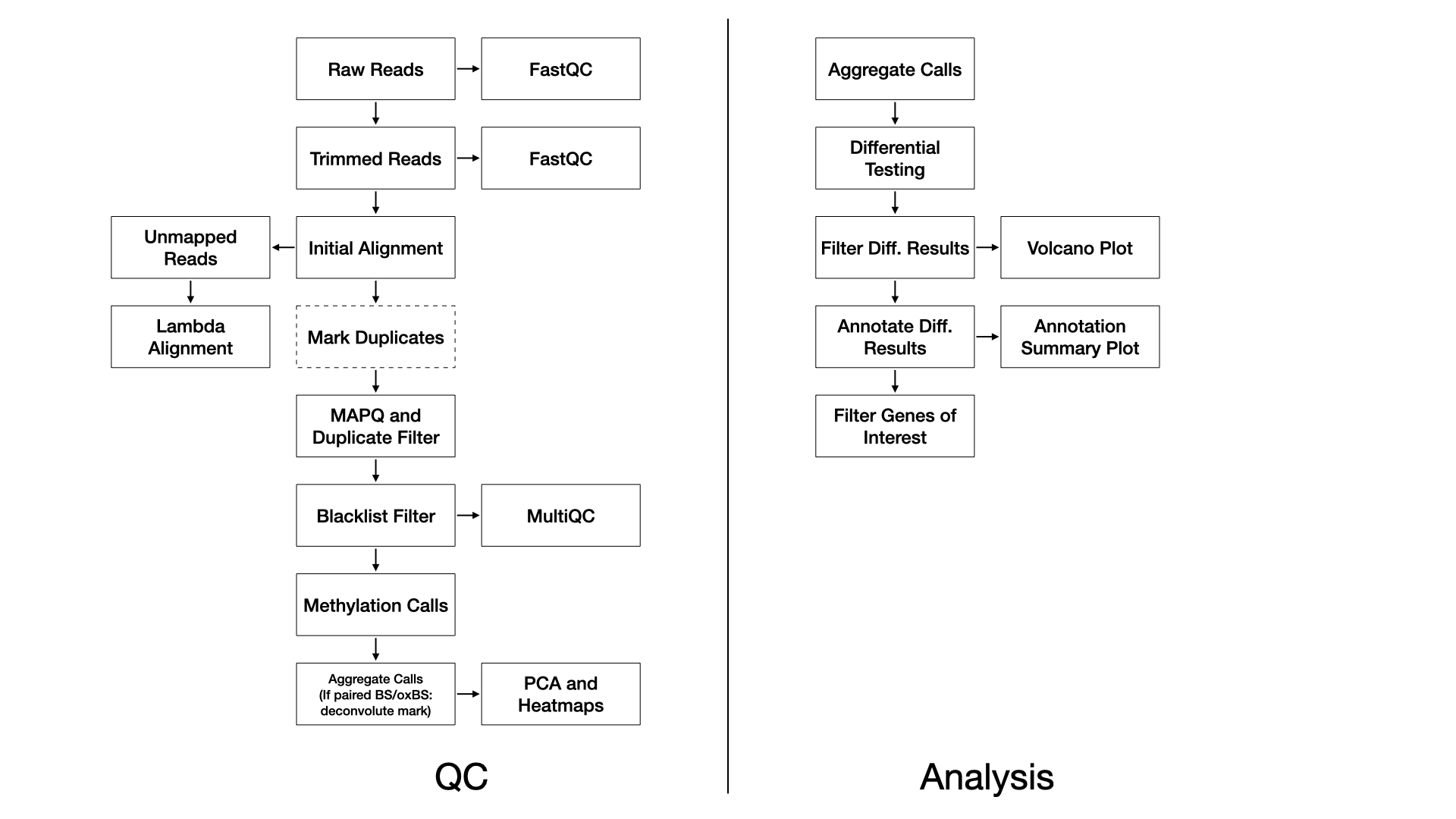
#### Without Mark Deconvolution

For each comparison we use the limma R Bioconductor package to identify differentially methylated probes (DMPs) by fitting a linear model on the M-values whose standard errors are then moderated using an empirical Bayes model (Ritchie *et al.*, 2015). The DMPs are then annotated to CpG island and genic annotations using the annotatr R Bioconductor package (Cavalcante and Sartor, 2017).

#### With Mark Deconvolution

For each comparison we use the gamlss R package to identify differentially methylated probes (DMPs) by fitting a beta-regression model on the beta-values (Stasinopoulos and Rigby, 2007). The DMPs are then annotated to CpG island and genic annotations using the annotatr R Bioconductor package (Cavalcante and Sartor, 2017).

# DNA Meythlation: DNA-Seq



DNA methylation workflow

### QC

We use FastQC (FastQC) (v0.11.8) to assess the overall quality of each sequenced sample.

We use TrimGalore (@ **article?**{noauthor\_trimgalore\_nodate,) (v0.4.5) with the following parameters: --adapter AGATCGGAAGAGC -e 0.1 --stringency 6 --length 20 --nextseq 20, and with the additional --rrbs flag in the ERRBS case.

Reads are aligned to the reference genome with Bismark (Krueger and Andrews, 2011) (v0.22.1) using Bowtie2 (Langmead and Salzberg, 2012) (v2.3.4) with default settings (multi-seed length of 20bp with 0 mismatches).

For WGBS, duplicate reads are marked and removed with Picard (Picard) (v2.20.2). This step is not performed for ERRBS.

Alignments below a MAPQ threshold are removed with samtools (Li *et al.*, 2009) (v1.2) and the parameters: -q 10.

MethylDackel (MethylDackel) (v0.4.0) then calls methylation rates with parameters -d 5 -D 2000 --mergeContext.

In the event of a BS/oxBS or BS/TAB library preparation, methylation mark deconvolution is then performed using the MLML2R R package (Qu *et al.*, 2013). Briefly, methylated and unmethylated counts from bisulfite-only treated samples and oxidative-bisulfte treated samples are extracted and passed to MLML2R::MLML() to determine the levels of methylcytosine (mC), hydroxymethylcytosine (hmC), and cytosine (C) using the exact method provided in the package.

### Differential Testing

#### Without Mark Deconvolution

Differential methylation testing is performed with the MethylSig R Bioconductor package which performs group-versus-group tests using a beta-binomial approach to calculate differential methylation statistics, accounting for variation among replicates within each group (Park *et al.*, 2014). Alternately, differential methylation testing can be performed with the DSS R bioconductor package which is a beta-binomial approach with arcsine link function to test under general experimental design (Wu and Park, 2016).

#### With Mark Deconvolution

We use the gamlss R package to identify differentially methylated probes (DMPs) by fitting a beta-regression model on the beta-values (Stasinopoulos and Rigby, 2007). The DMPs are then annotated to CpG island and genic annotations using the annotatr R Bioconductor package (Cavalcante and Sartor, 2017).

# Spatial Transcriptomics: Nanostring GeoMX

Slide preparation, including any morphology marker staining, and full ROI or segmented AOI selection will be performed by the Advanced Genomic Core (AGC) for an NGS readout workflow. Sequencing depth is estimated by multiplying the total area of all the AOIs (µm2) with the appropriate sequencing depth factor for the selected GeoMx panel (e.g. Mouse Whole Transcriptome Atlas RNA v1.0) and the pooled library will be sequenced. Raw FASTQ files are processed using Nanostring’s proprietary Automated Data Processing Pipeline, which includes adapter trimming and aligning stitched paired-end reads to the barcodes in the reference assay before removal of PCR duplicates based on the Unique Molecular Identifier (UMI) in each read.

The resulting DCC files for each full ROI or segmented AOI are then processed with the GeoMxTools package using the corresponding Nanostring DSP configuration file and the instrument generated LabWorksheet with ROI/AOI identifiers and manually annotated group labels. Initial quality control filtering and third quartile (Q3) normalization per Nanostring’s recommendations for GeoMx DSP data - with minimum average negative control measurements for ROI/AOI retention and minimum counts above background for target (gene/protein) retention.

The resulting filtered and Q3 normalized data are used as inputs for differential expression comparisons, again with the GeoMxTools package. Comparisons within a single slide are generated with an unpaired t-test. An absolute fold-change > 1.5 and an FDR adjusted p-value < 0.05 are used to call differentially expressed targets.

# References

[Ataqv](https://github.com/ParkerLab/ataqv) *https://github.com/ParkerLab/ataqv*.

Becht,E. *et al.* (2018) [Dimensionality reduction for visualizing single-cell data using UMAP](https://doi.org/10.1038/nbt.4314). *Nature Biotechnology*.

Benjamini,Y. and Hochberg,Y. (1995) [Controlling the false discovery rate: A practical and powerful approach to multiple testing](http://www.jstor.org/stable/2346101). *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bergen,V. *et al.* (2020) [Generalizing RNA velocity to transient cell states through dynamical modeling](https://doi.org/10.1038/s41587-020-0591-3). *Nature Biotechnology*, **38**, 1408–1414.

Cao,J. *et al.* (2019) [The single-cell transcriptional landscape of mammalian organogenesis](https://doi.org/10.1038/s41586-019-0969-x). *Nature*, **566**, 496–502.

Cavalcante,R.G. and Sartor,M.A. (2017) [Annotatr: Genomic regions in context](https://doi.org/10.1093/bioinformatics/btx183). *Bioinformatics*, **33**, 2381–2383.

Dobin,A. *et al.* (2013) [STAR: Ultrafast universal RNA-seq aligner](https://doi.org/10.1093/bioinformatics/bts635). *Bioinformatics (Oxford, England)*, **29**, 15–21.

Draghici,S. *et al.* (2007) [A systems biology approach for pathway level analysis](https://doi.org/10.1101/gr.6202607). *Genome Research*, **17**, 1537–1545.

[ENCODE Blacklist Regions](https://sites.google.com/site/anshulkundaje/projects/blacklists) *https://sites.google.com/site/anshulkundaje/projects/blacklists*.

Ewels,P. *et al.* (2016) [MultiQC: Summarize analysis results for multiple tools and samples in a single report](https://doi.org/10.1093/bioinformatics/btw354). *Bioinformatics*, **32**, 3047–3048.

[FastQC](https://github.com/s-andrews/FastQC) *https://github.com/s-andrews/FastQC*.

Fortin,J.-P. *et al.* (2014) [Functional normalization of 450k methylation array data improves replication in large cancer studies](https://doi.org/10.1186/s13059-014-0503-2). *Genome Biology*, **15**, 503.

Heinz,S. *et al.* (2010) [Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities](https://doi.org/10.1016/j.molcel.2010.05.004). *Molecular Cell*, **38**, 576–589.

Houseman,E.A. *et al.* (2012) [DNA methylation arrays as surrogate measures of cell mixture distribution](https://doi.org/10.1186/1471-2105-13-86). *BMC Bioinformatics*, **13**, 86.

[iPathwayGuide](https://advaitabio.com/ipathwayguide/).

Khan,A. *et al.* (2017) [JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework](https://doi.org/10.1093/nar/gkx1126). *Nucleic Acids Research*, **46**, D260–D266.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

La Manno,G. *et al.* (2018) [RNA velocity of single cells](https://doi.org/10.1038/s41586-018-0414-6). *Nature*, **560**, 494–498.

Landt,S.G. *et al.* (2012) [ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia](https://doi.org/10.1101/gr.136184.111). *Genome Research*, **22**, 1813–1831.

Lange,M. *et al.* (2022) [CellRank for directed single-cell fate mapping](https://doi.org/10.1038/s41592-021-01346-6). *Nature Methods*, **19**, 159–170.

Langmead,B. and Salzberg,S.L. (2012) [Fast gapped-read alignment with Bowtie 2](https://doi.org/10.1038/nmeth.1923). *Nature Methods*, **9**, 357–359.

Li,B. and Dewey,C.N. (2011) [RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome](https://doi.org/10.1186/1471-2105-12-323). *BMC bioinformatics*, **12**, 323.

Li,H. *et al.* (2009) [The Sequence Alignment/Map format and SAMtools](https://doi.org/10.1093/bioinformatics/btp352). *Bioinformatics*, **25**, 2078–2079.

Li,Z. *et al.* (2019) [Identification of transcription factor binding sites using ATAC-seq](https://doi.org/10.1186/s13059-019-1642-2). *Genome Biology*, **20**.

Love,M.I. *et al.* (2014) [Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2](https://doi.org/10.1186/s13059-014-0550-8). *Genome Biology*, **15**, 550.

Martin,M. [Cutadapt removes adapter sequences from high-throughput sequencing reads](https://doi.org/10.14806/ej.17.1.200). *EMBnet.journal*, **17**, 3.

McCarthy,D.J. *et al.* (2012) [Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation](https://doi.org/10.1093/nar/gks042). *Nucleic Acids Research*, **40**, 4288–4297.

[MethylDackel](https://github.com/dpryan79/MethylDackel) *https://github.com/dpryan79/MethylDackel*.

Neph,S. *et al.* (2012) [BEDOPS: High-performance genomic feature operations](https://doi.org/10.1093/bioinformatics/bts277). *Bioinformatics*, **28**, 1919–1920.

Park,Y. *et al.* (2014) MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**, 2414–2422.

[Picard](https://github.com/broadinstitute/picard) *https://github.com/broadinstitute/picard*.

Pidsley,R. *et al.* (2013) [A data-driven approach to preprocessing Illumina 450K methylation array data](https://doi.org/10.1186/1471-2164-14-293). *BMC Genomics*, **14**.

Qu,J. *et al.* (2013) [MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation](https://doi.org/10.1093/bioinformatics/btt459). *Bioinformatics*, **29**, 2645–2646.

Quinlan,A.R. and Hall,I.M. (2010) [BEDTools: A flexible suite of utilities for comparing genomic features](https://doi.org/10.1093/bioinformatics/btq033). *Bioinformatics*, **26**, 841–842.

Ramírez,F. *et al.* (2016) [deepTools2: A next generation web server for deep-sequencing data analysis](https://doi.org/10.1093/nar/gkw257). *Nucleic Acids Research*, **44**, W160–W165.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47–e47.

Saelens,W. *et al.* (2019) [A comparison of single-cell trajectory inference methods](https://doi.org/10.1038/s41587-019-0071-9). *Nature Biotechnology*, **37**, 547–554.

[Seurat tutorials](http://satijalab.org/seurat/).

Shao,X. *et al.* (2020) [scCATCH: Automatic annotation on cell types of clusters from single-cell RNA sequencing data](https://doi.org/10.1016/j.isci.2020.100882). *iScience*, **23**, 100882.

Stasinopoulos,D.M. and Rigby,R.A. (2007) [Generalized Additive Models for Location Scale and Shape (GAMLSS) in *r*](https://doi.org/10.18637/jss.v023.i07). *Journal of Statistical Software*, **23**.

Stuart,T. *et al.* (2019) [Comprehensive integration of single-cell data](https://doi.org/10.1016/j.cell.2019.05.031). *Cell*, **177**, 1888–1902.e21.

[TrimGalore](https://github.com/FelixKrueger/TrimGalore) *https://github.com/FelixKrueger/TrimGalore*.

Welsh,H. *et al.* (2023) [A systematic evaluation of normalization methods and probe replicability using infinium EPIC methylation data](https://doi.org/10.1186/s13148-023-01459-z). *Clinical Epigenetics*, **15**, 41.

Wingett,S.W. and Andrews,S. (2018) [FastQ screen: A tool for multi-genome mapping and quality control](https://doi.org/10.12688/f1000research.15931.2). *F1000Research*, **7**, 1338.

Wolf,F.A. *et al.* (2019) [PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells](https://doi.org/10.1186/s13059-019-1663-x). *Genome Biology*, **20**, 59.

Wu,H. and Park,Y. (2016) Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, **32**, 1446–1453.

Yang,S. *et al.* (2020) [Decontamination of ambient RNA in single-cell RNA-seq with DecontX](https://doi.org/10.1186/s13059-020-1950-6). *Genome Biology*, **21**, 57.

Zhang,Y. *et al.* (2008) [Model-based Analysis of ChIP-Seq (MACS)](https://doi.org/10.1186/gb-2008-9-9-r137). *Genome Biology*, **9**, R137.

Zheng,G.X.Y. *et al.* (2017) [Massively parallel digital transcriptional profiling of single cells](https://doi.org/10.1038/ncomms14049). *Nature Communications*, **8**, 14049.

Zhou,W. *et al.* (2018) [SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions](https://doi.org/10.1093/nar/gky691). *Nucleic Acids Research*, **46**, e123–e123.