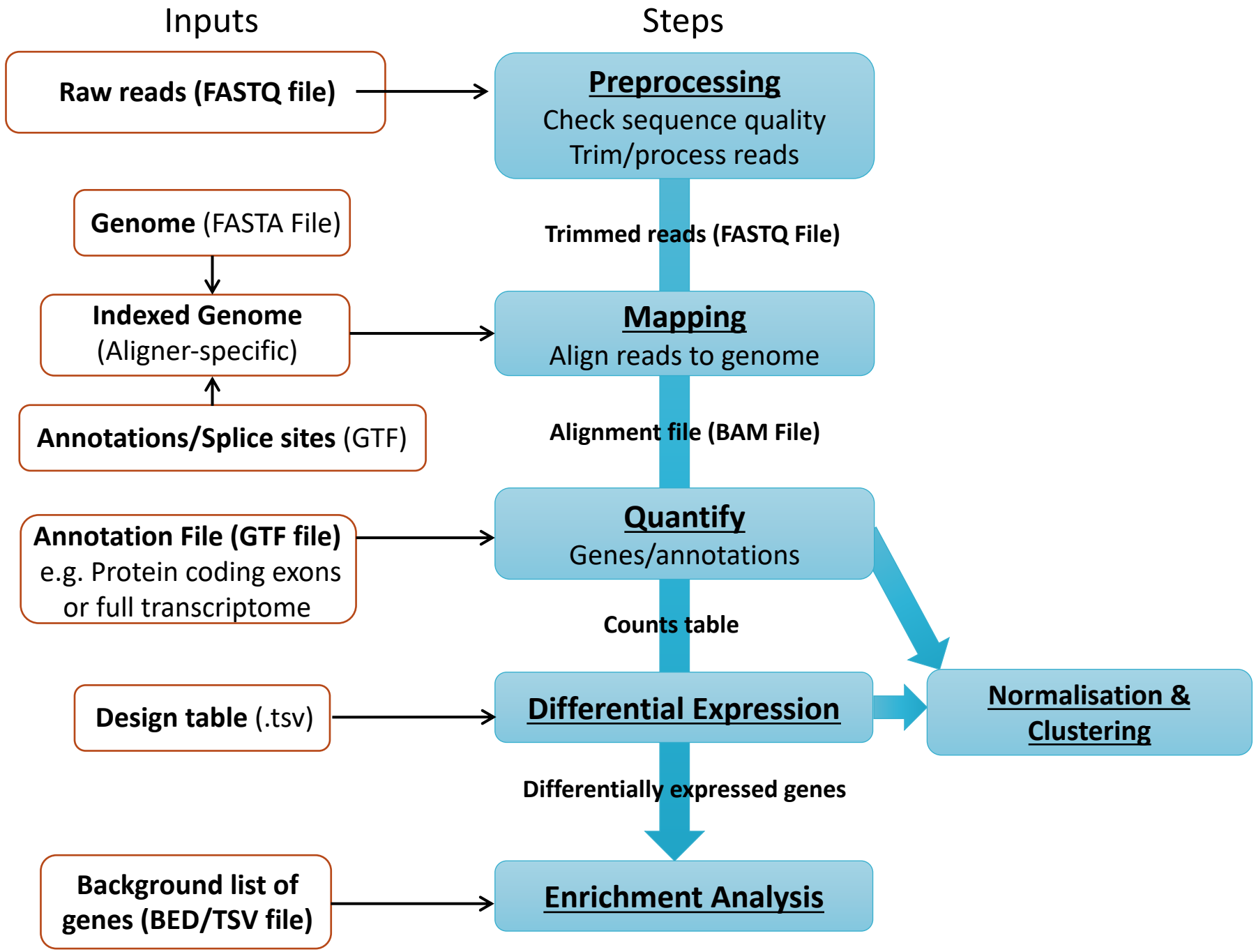
RNA-seq on the Linux command line

Oxford Biomedical Data Science Training Programme

RNA-seq analysis

Two basic workflows:

1. Alignment method – FASTQ → BAM → counts
2. Pseudoalignment – FASTQ → counts


Inputs
Steps
Raw reads (FASTQ file)
Preprocessing
Check sequence quality
Trim/process reads
Genome (FASTA File)
Trimmed reads (FASTQ File)
Indexed Genome
(Aligner-specific)
Mapping
Align reads to genome
Annotations/Splice sites (GTF)
Alignment file (BAM File)
Annotation File (GTF file)
e.g. Protein coding exons
or full transcriptome
Quantify
Genes/annotations
Counts table
Design table (.tsv)
Differential Expression
Normalisation &
Clustering
Differentially expressed genes
Background list of
genes (BED/TSV file)
Enrichment Analysis

Choosing appropriate analysis tools

How to decide what tool to use for each step of the workflow?

- Ask around
- What's most recent – is it a major advance?
- What do reviews/blogs comparing the tools say?
- What is the most appropriate for your question/problem? READ the manual!
- Test multiple tools – compare results

RNA-seq group exercise

- Think about the workflow and tools to use at each stage
- Find documentation for tools
- Download data & verify download
  - European Nucleotide Archive PRJEB18572
- Run tools on the Linux command line
- Final result: counts matrix

Information about the RNA-seq dataset

- Mouse $CD4^+$ and $CD8^+$ T cells from GFP-Egr2 knockin (Egr2 Kin) and $Egr2^{loxP/loxP}$ hCD2-Cre $Egr3^{-/-}$ (Egr2/3 DKO) mice 7 days after infection with vaccinia virus
- 3 biological replicates per group (12 samples total)
- TruSeq Stranded mRNA Library Prep Kit

Useful resources

- https://www.rna-seqblog.com
- https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-016-0881-8
- https://rnaseq.uoregon.edu
- https://rnabio.org
- https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing