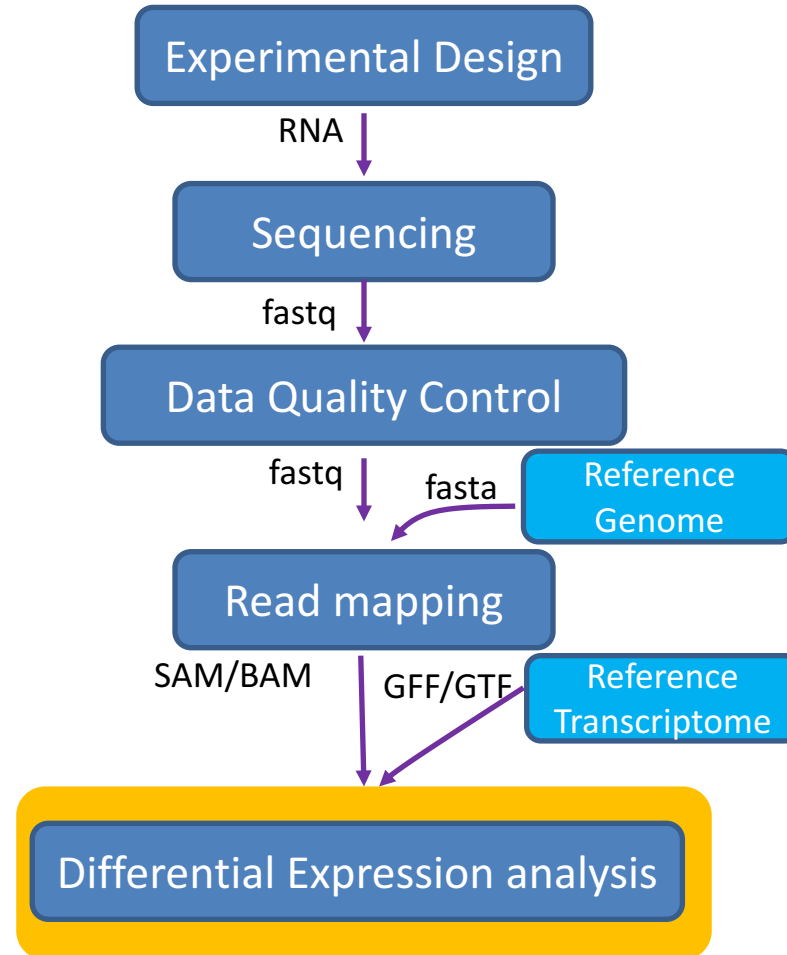
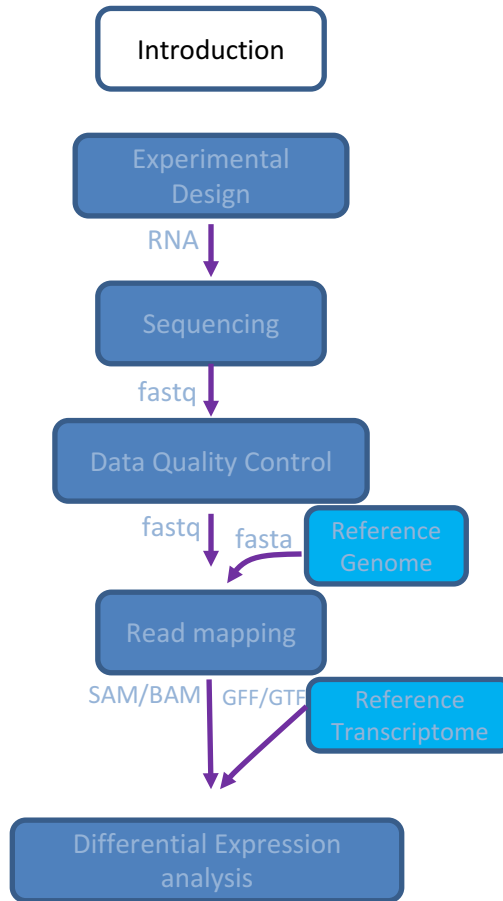


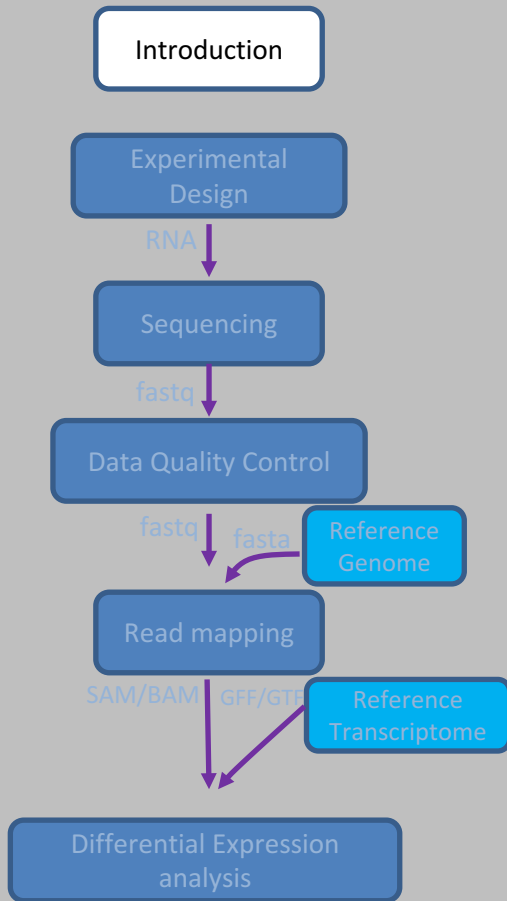
RNA-Seq Analysis



Quality Control checks

- Reproducibility
- Reliability

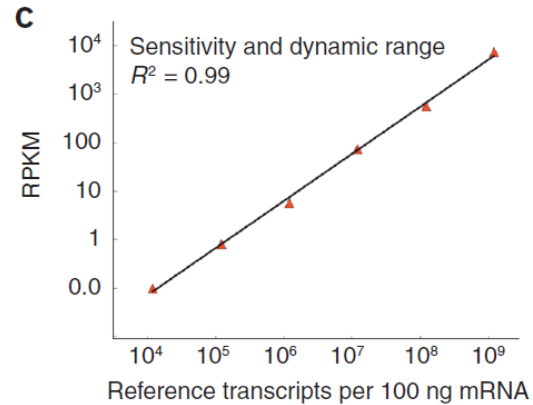
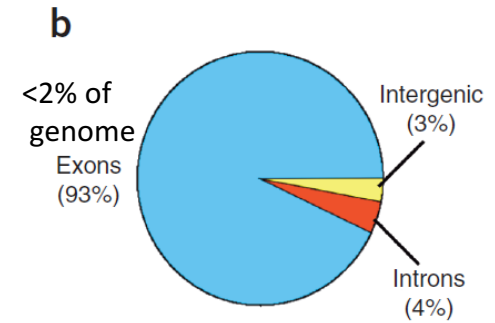
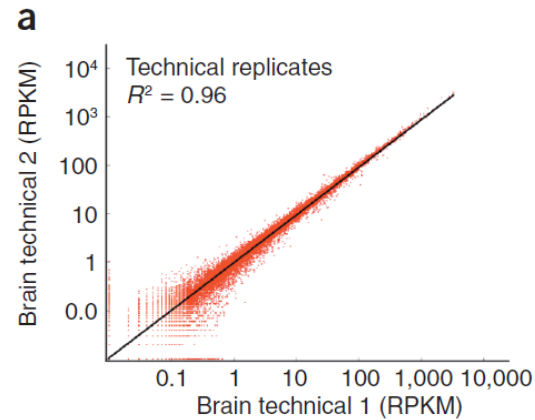
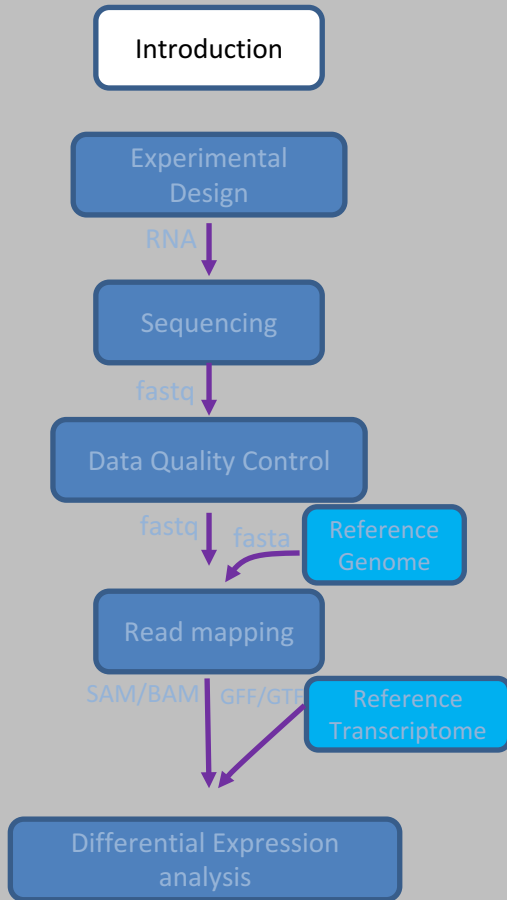




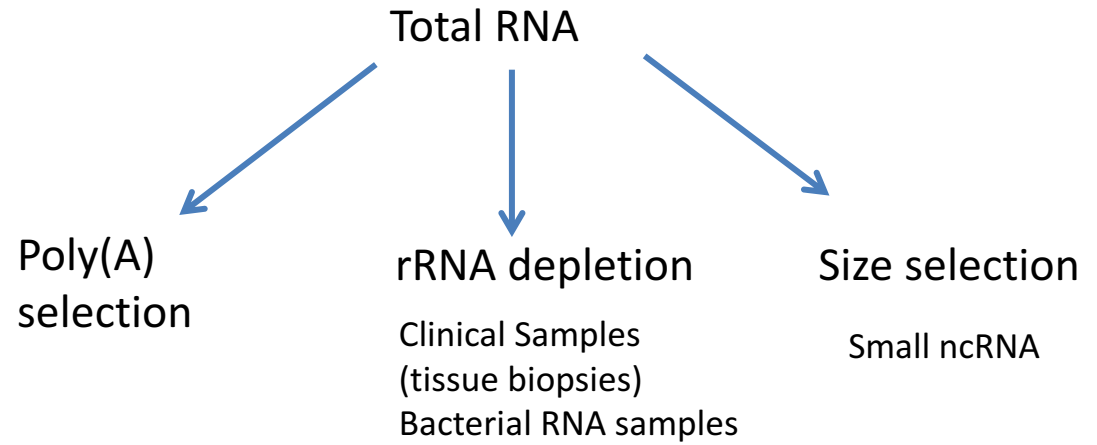
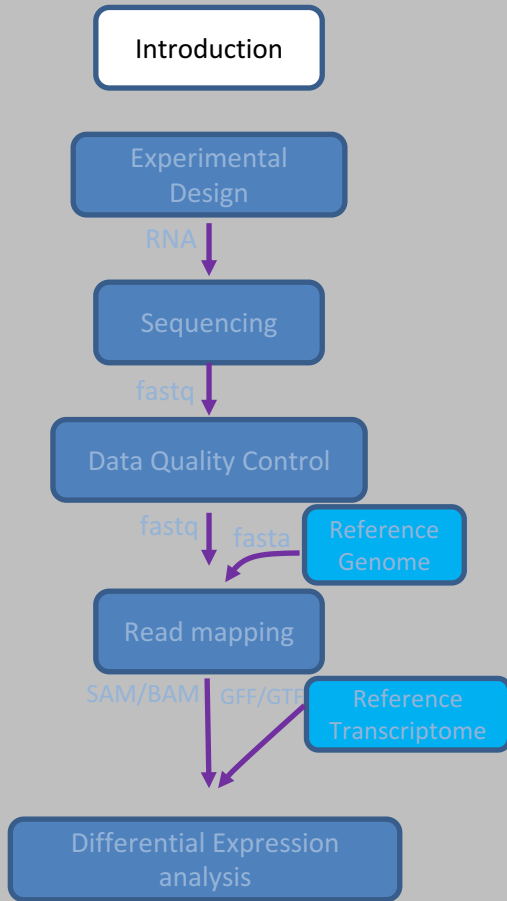
RNA-seq vs Microarray

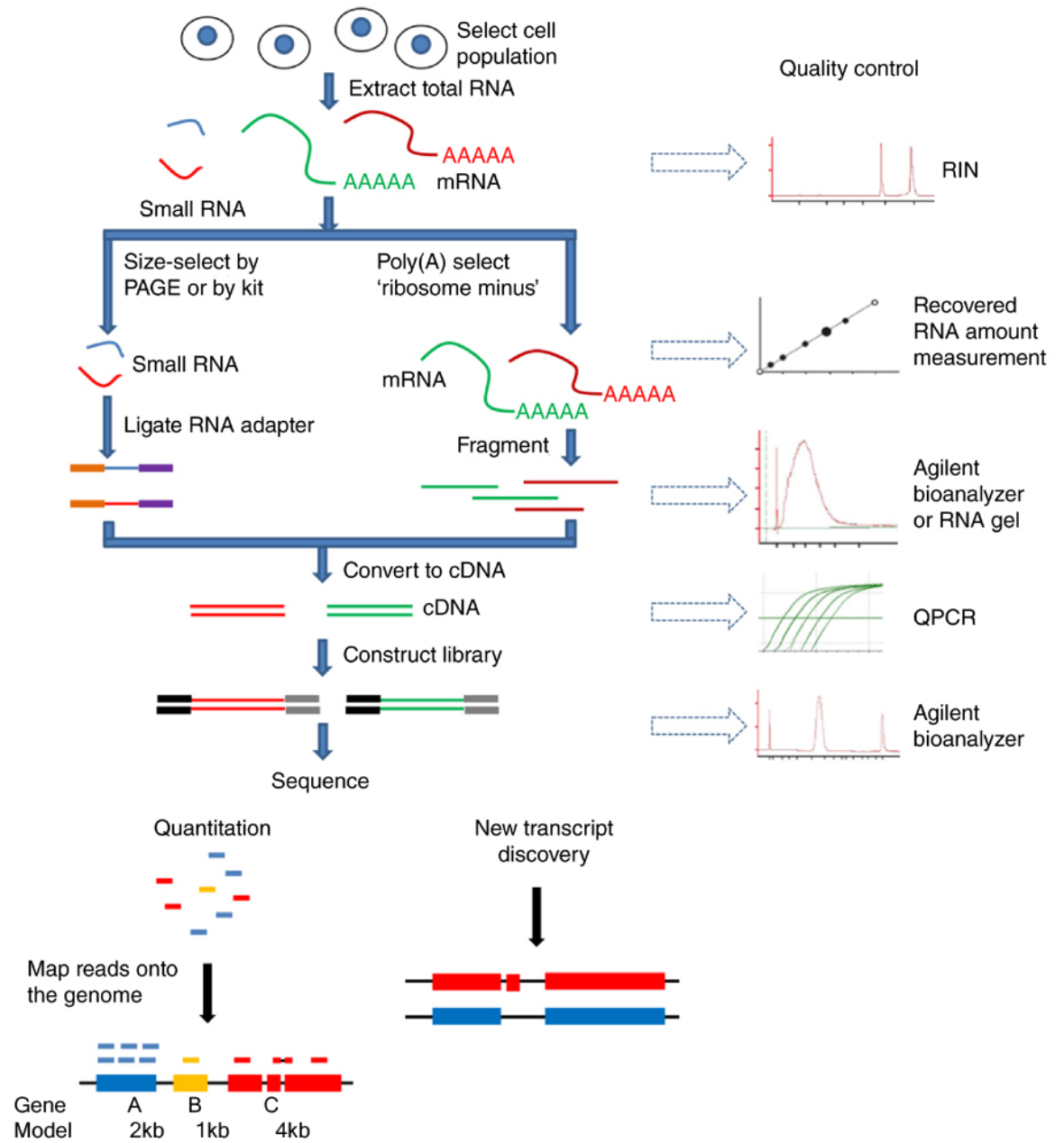
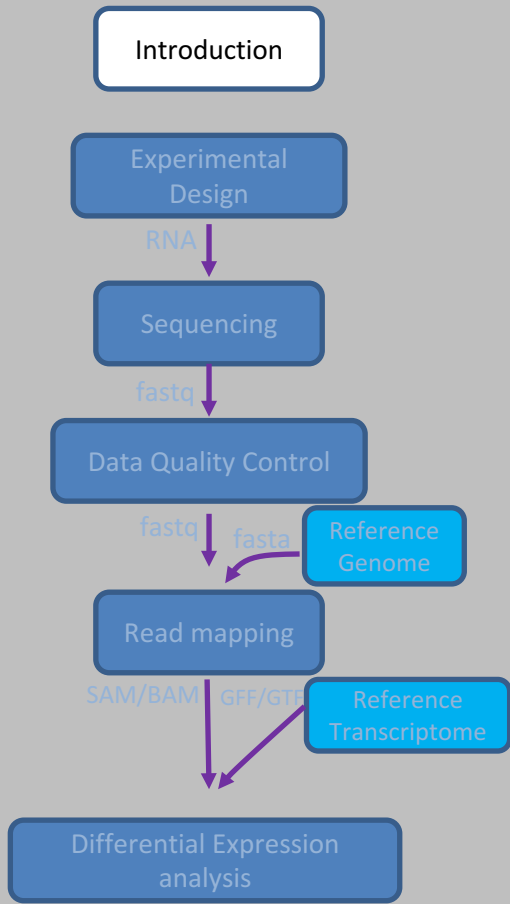
- Higher sensitivity and dynamic range
- Lower technical variation
- Available for all species
- Novel transcript identification
- Alternate splicing
- Allele specific expression
- Fusion genes
- **Higher Informatics Cost**

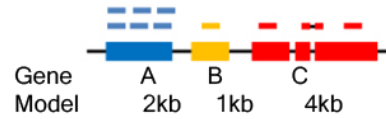
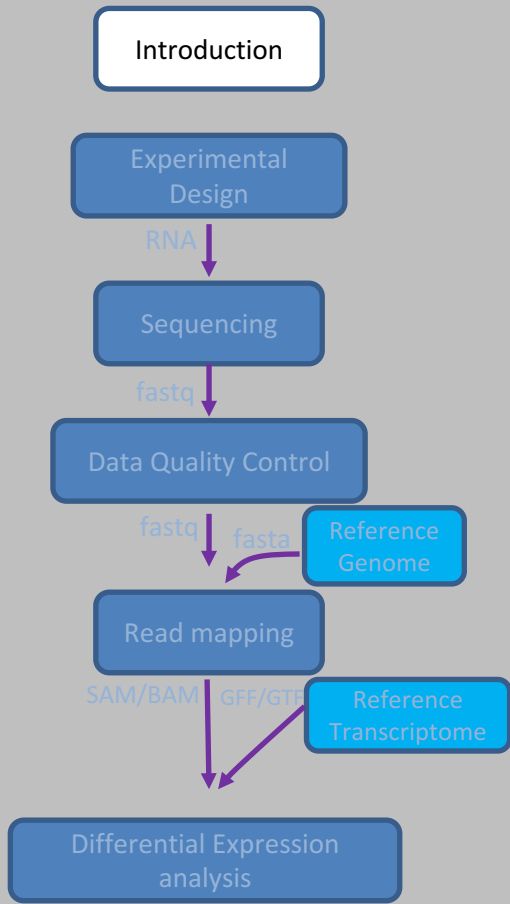
Reproducibility, Linearity and Sensitivity



RNA isolation







	Gene A	Gene B	Gene C
Sample 1	6	1	4

of reads

	Gene A	Gene B	Gene C
Sample 1	3	1	1

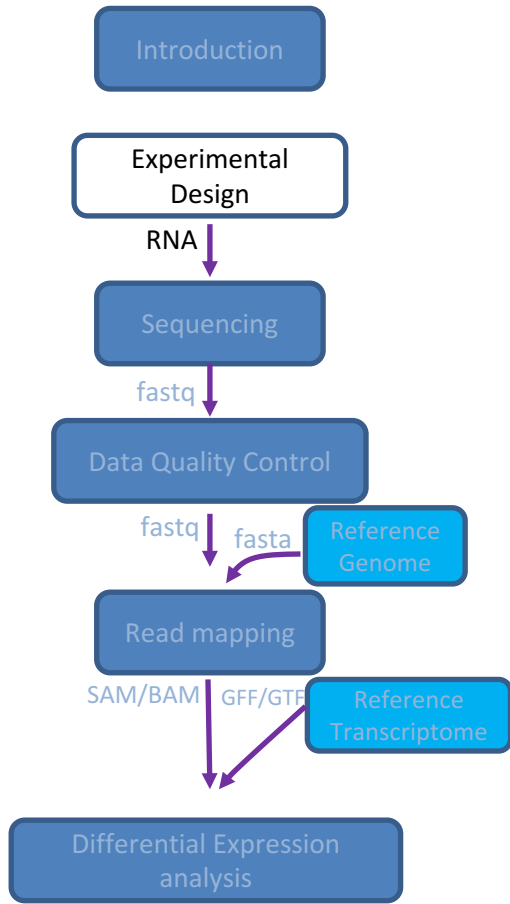
Reads per kb of exon

	Gene A	Gene B	Gene C	Total
Sample 1	3	1	1	5
Sample 2	6	3	6	15

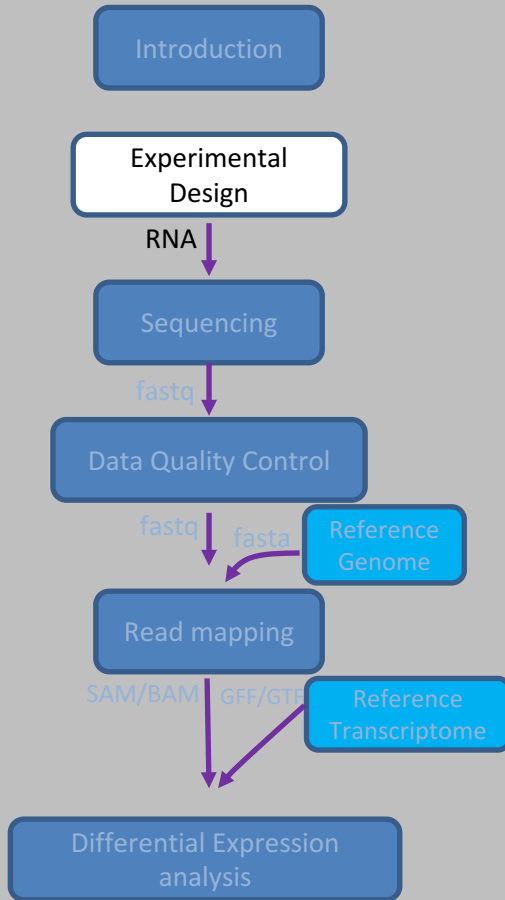
Reads per kb of exon

	Gene A	Gene B	Gene C	Total
Sample 1	0.6	0.2	0.2	5
Sample 2	0.4	0.2	0.4	15

Reads per kb of exon per million mapped reads - **RPKM**



Experimental Design

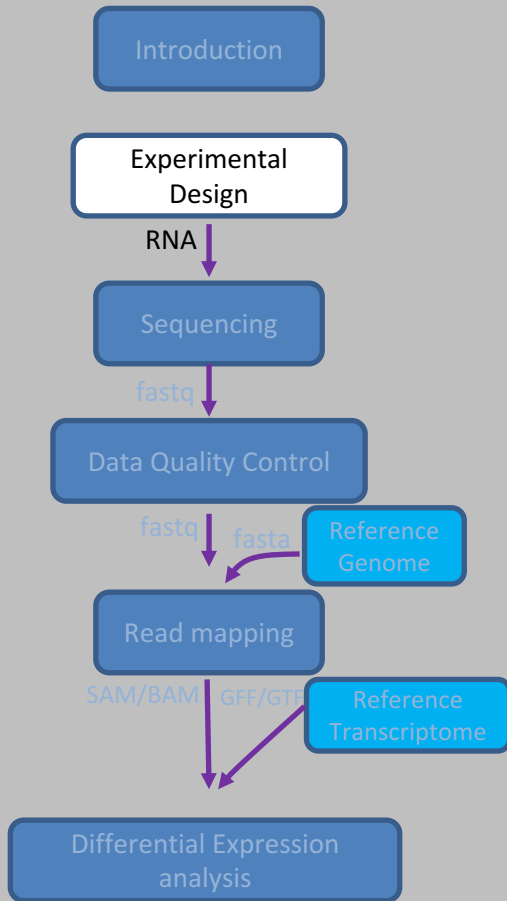


What are my goals?

- Transcript assembly
- Differential Expression analysis
- Identify new/rare transcripts

What are Characteristics of my system?

- Large and complex genome
- Introns and high degree of alternative splicing
- No reference genome or transcriptome.



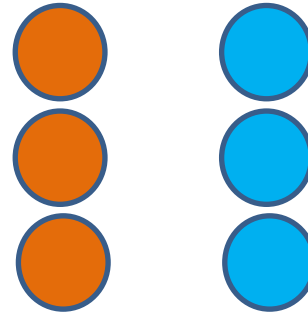
Experimental Design

- Biological Comparison(s)
- Paired End vs Single end
- Read depth
- Read length
- Replicates

Experimental Design

Simple Design- Pairwise comparison

Two Groups



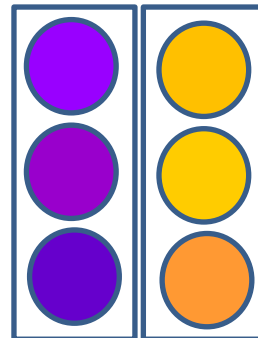
Control

Experimental treatment

Complex design

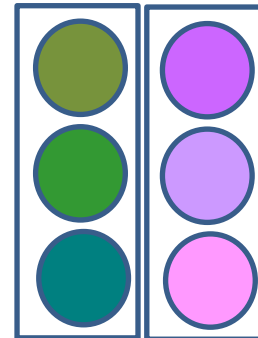
Cancer Subtype A

Cancer Subtype B



-

+drug

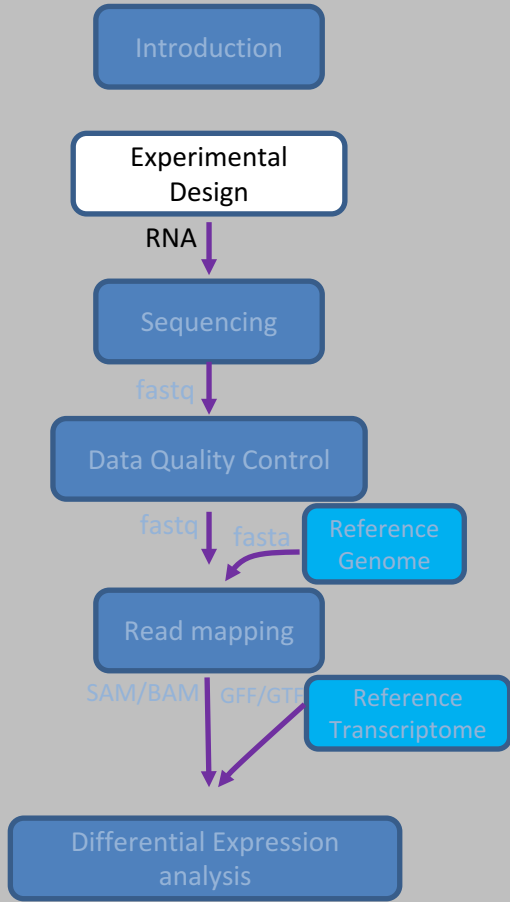


-

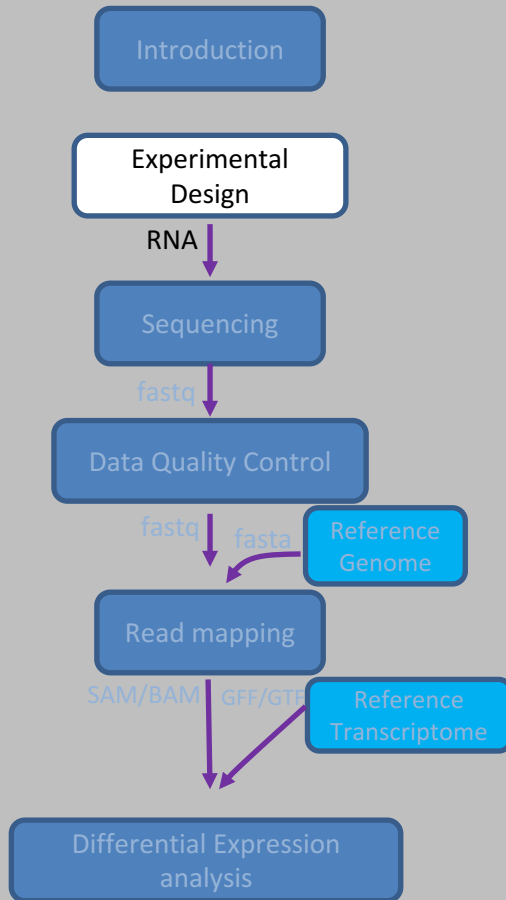
+drug



Consult a statistician



Experimental Design



Read depth and read length

Small genome with no alternate splicing (yeast)

Annotated transcriptome

10million reads per sample, 50bp single-end reads

Mammalian genomes

(Large transcriptome, alternative splicing, gene duplication)

30million reads per sample,

Transcriptome assembly (100X coverage of transcriptome)

50-200million reads per sample, 100bp paired end reads

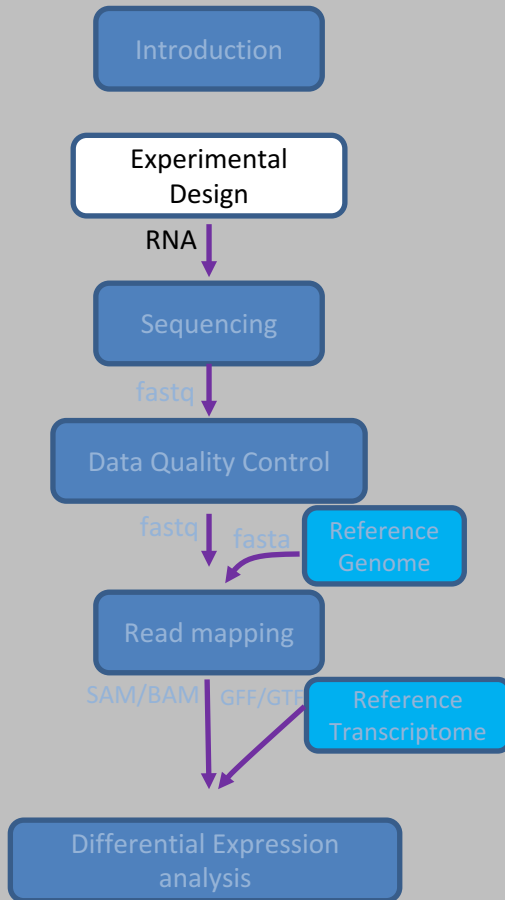
Nature of samples.

- What is the expected purity of your sample?
- Is there contamination or heterogeneity expected?

If yes, then

High coverage to detect variants at lower frequency due to impurity or because they come from minor (but possibly still interesting) subpopulations of your sample.

Experimental Design



Replicates:

Factors determining number of replicates:

- Variability in measurements (Technical noise and Biological variation)
- Statistical power analysis

Technical Replicates

Not Needed : High reproducibility at sequencing step

Error prone steps

RNA fragmentation, cDNA synthesis, adapter ligation, PCR amplification, bar-coding, lane loading

Spike Ins: Quality control and library-size normalisation

Minimize batch effects

Randomize samples at library preparation and sequencing runs

Biological Replicates

Not required for transcription assembly

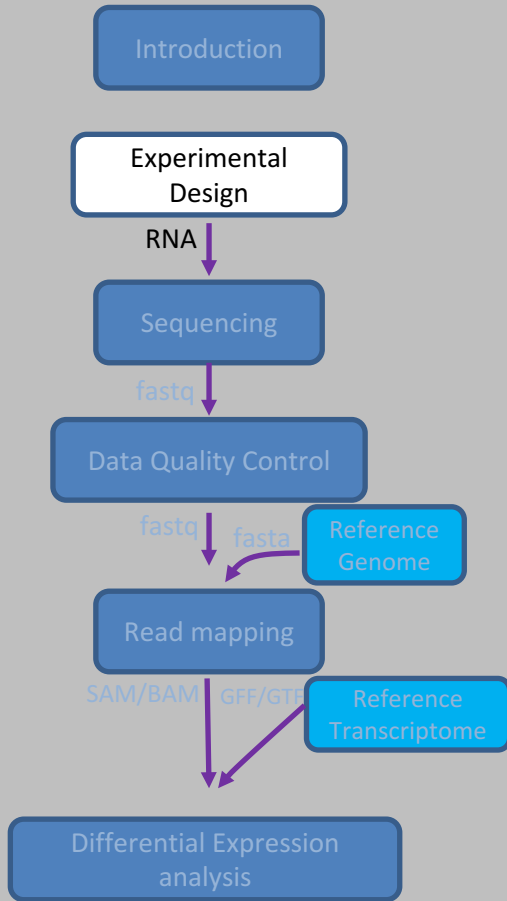
Essential for differential expression analysis

Complex designs:

- 3+ for cell lines
- 5+ for inbred lines
- 20+ for human samples

Experimental Design

Scotty: <http://scotty.genetics.utah.edu/>



The screenshot shows the Scotty web application interface. At the top, there is a navigation bar with several tabs. The main content area is titled "The Matlab code that runs background calculations is available on [github](#). Please contact us if your require assistance."

Inputs

Pilot Data: Upload your own pilot data or used a stored dataset as a model for your experiment. (?)

CAUTION
Power analysis results will not be predictive of the actual results unless the power analysis is performed on data that closely matches the experiment. Please read about [generating pilot data](#) and [selecting preloaded datasets](#) before continuing.

Upload Data

Upload a file containing the number of reads per gene for pilot data as a tab delimited text file. [See format info.](#)

No file selected.

Number of Replicates in Control:

Number of Replicates in Test (enter 0 if none):

Use a stored dataset(?)

Choose a model dataset (*Less Accurate*): [Dataset Descriptions](#)

Cost Data (?)

Cost per replicate, excluding reads:

Control:

Test:

Cost per million reads sequenced: (?)

Alignment Rate (to genes or transcripts): % ([How to calculate?](#))

Constraints for Power Optimization(?)

Experimental Configurations to Test:

Maximum number of biological replicates per condition:

Assess the power of sequencing depths between and reads aligned to genes per replicate

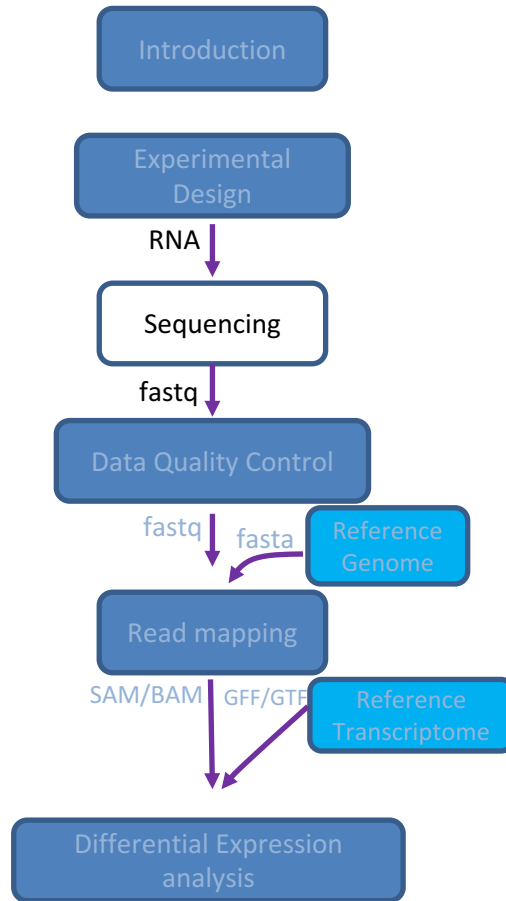
Leave the following fields blank to leave parameters unconstrained:

Detect at least % of expressed genes that are differentially expressed by a X fold change at $p < 0.01$

Experiment will cost no more than \$ (?)






Limit measurement bias by measuring at least % of genes with at least % of maximum power (?)

Results processing usually takes about 5 minutes.



Sequencing

Illumina sequencing by synthesis

					
	MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million [†]	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

SOLID

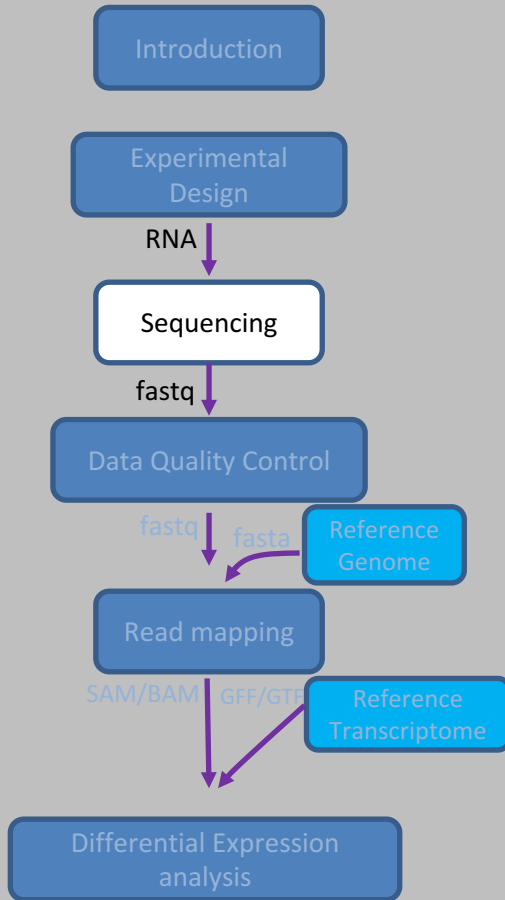
“Color-Space” reads
Low error rate

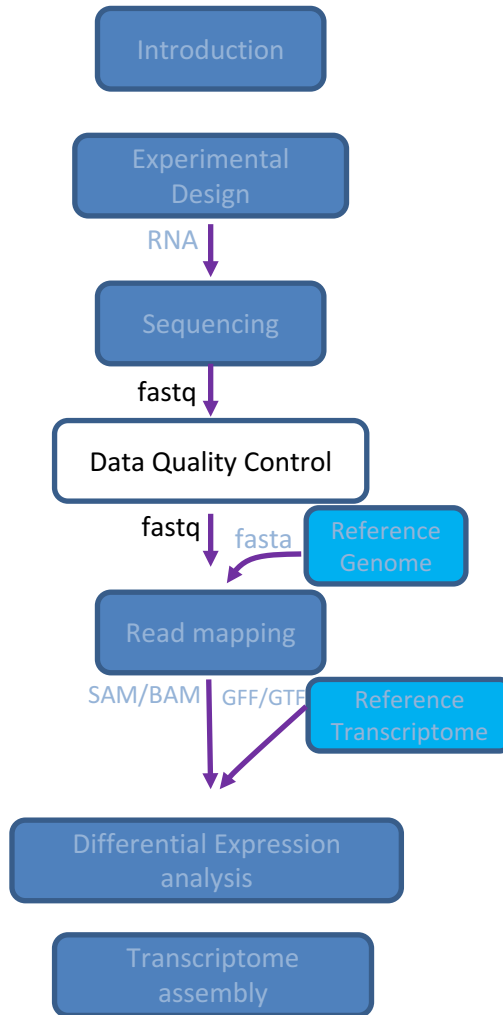
454 pyrosequencing

Longer reads, low throughput

Pacific-Bioscience (pacBio)/ Oxford Nanopore

Longer read (Recovery of full length transcripts)

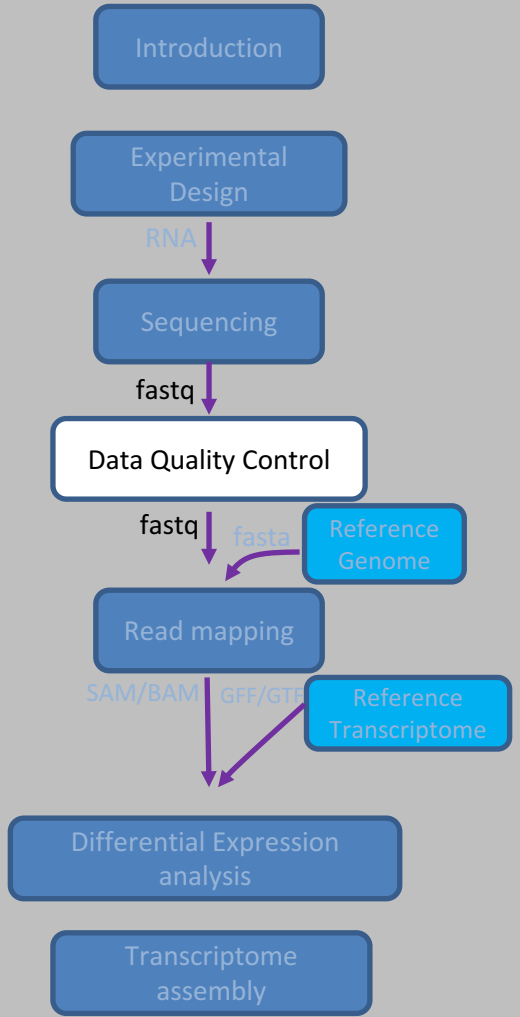




Sequence Data Format .FASTQ



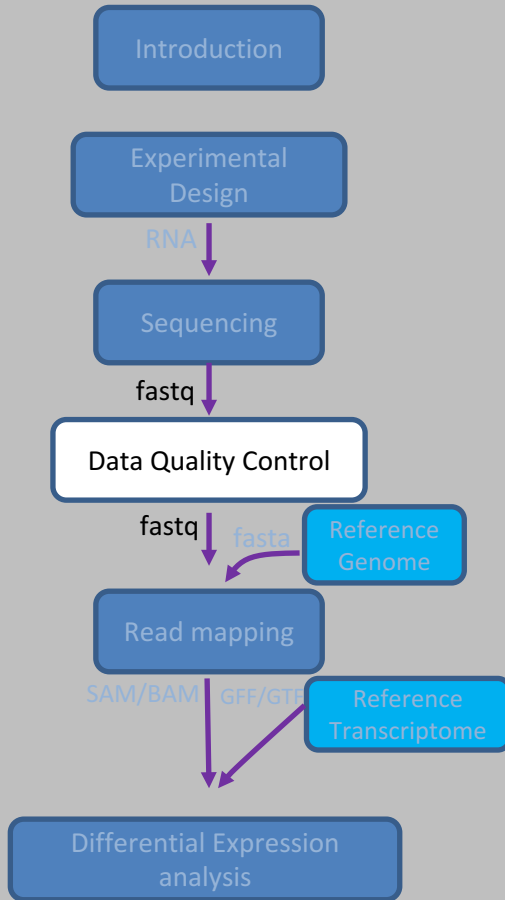
Sample1_R1.fastq
Sample1_R2.fastq



```

Machine ID
Read ID → @NS500650:24:HGVNFBGXX:1:11101:23736:1051 1:N:0:2
Sequence → TGCACNTTCATTATTGACGCTAACAAAGGATTTGAAGACTACAGATTCTGTGAGTGTCAACAAATTGGTTCCTGTTT
Quality Score → +
Phred +33 → AAAAA#EEEEEEAEEEE6EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
QC Filter flag → @NS500650:24:HGVNFBGXX:1:11101:25377:1054 1:N:0:2
Y=bad → TACCTNATAATATAACACAAGTACCGACAAATAAAAATTCCTACAGAAATACTCTTAGACAATTCTTCTACTCCAAT
N=good → +
Sample ID/Barcode → A/AAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
Read pair # 1 → @NS500650:24:HGVNFBGXX:1:11101:17837:1054 1:N:0:2
CGTCCNCTCTTCATCTTCTCTTCATCTTCTCTATCGTTTCATATCCGCTTACCGCATACCAAGTAATGTCAT
Read pair # 2 → +
A/AAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
Read pair # 2 → @NS500650:24:HGVNFBGXX:1:11101:26874:1092 2:N:0:2
TGAAGCCGAGAAAAACAGAAATACCTGCAGCCAGTGTTTCGAAAATCGCAATTGACGCTAGCTCTACAGCGCATTTT
+
AAAAAEEEEAAE66EEAEEEE6//E/EEEEEEEEEEEE/EEEEEE/EE/EE//EAE/EEE6E/E/<EA6EE/A<EAE
@NS500650:24:HGVNFBGXX:1:11101:3459:1094 2:N:0:2
AGAGTGTCCAGAAATAGTTGAGGTAACAGCACAAATCCTTCATTTTAGTGCAACAACCTTTTAGAACAAAACAAAATCA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<EEEEEEAEEEEEAEEEEEEEEEE/
@NS500650:24:HGVNFBGXX:1:11101:1188:1095 2:N:0:2
AGAAGTTAAAGATGCTTATGAAAATTTGTAAAGATTTATGCAATTTCTTCTGATAATTATTAATGTTTCATTGGAA
+
A/AA6/EEE6EE66/<<EE6//<A/EEA/EEAE/EE//E//E//E<E</EE/AEEEA/EE/E//E6/AAE</AA
    
```

Data Quality Assessment



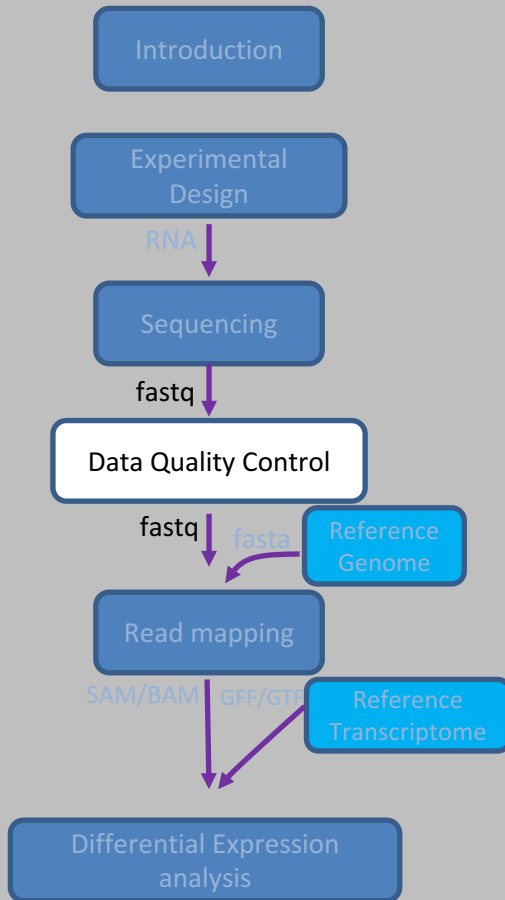
Evaluate raw read library quality

- Sequence quality
- GC content for biases
- Adapter Contamination
- *K*-mer over representation
- Duplicate reads
- PCR artifacts

Software/Tools

- FASTQC (Command line)
Illumina read files
- NGSQC
Support reads from any platform
Support quality based read trimming and filtering
- SAMSat (Command line)
Also work with Bam alignment files

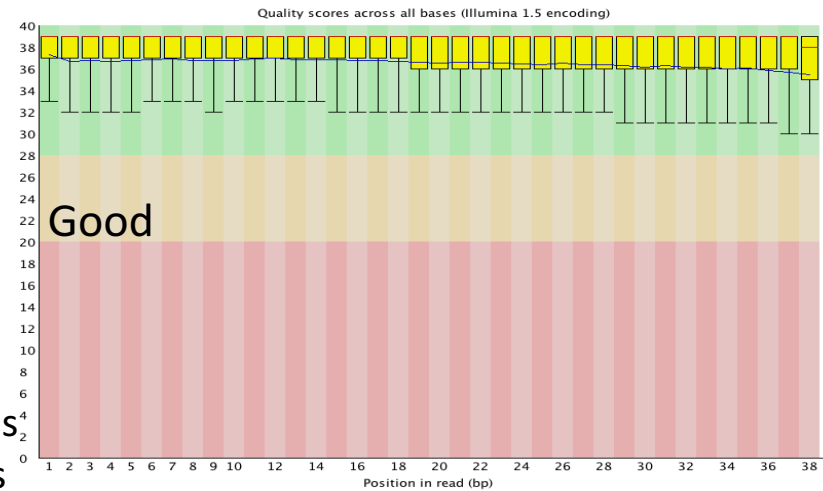
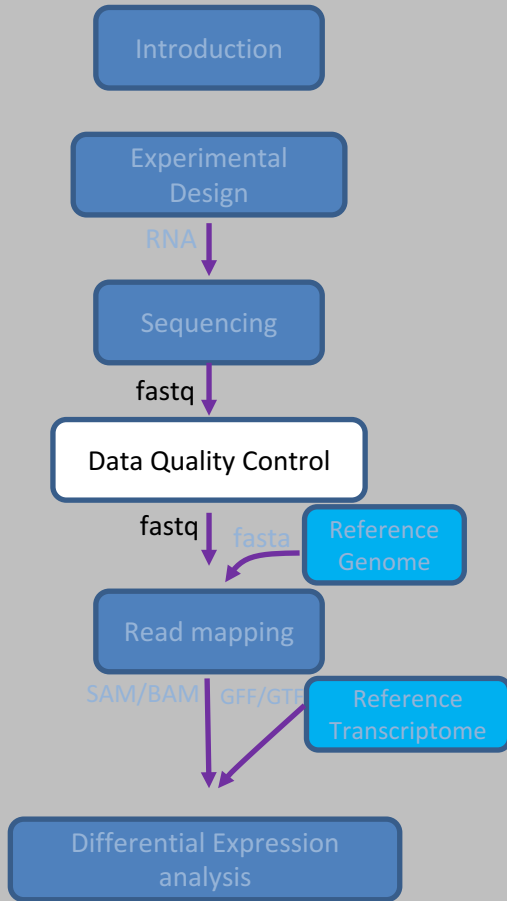
Data Quality Assessment



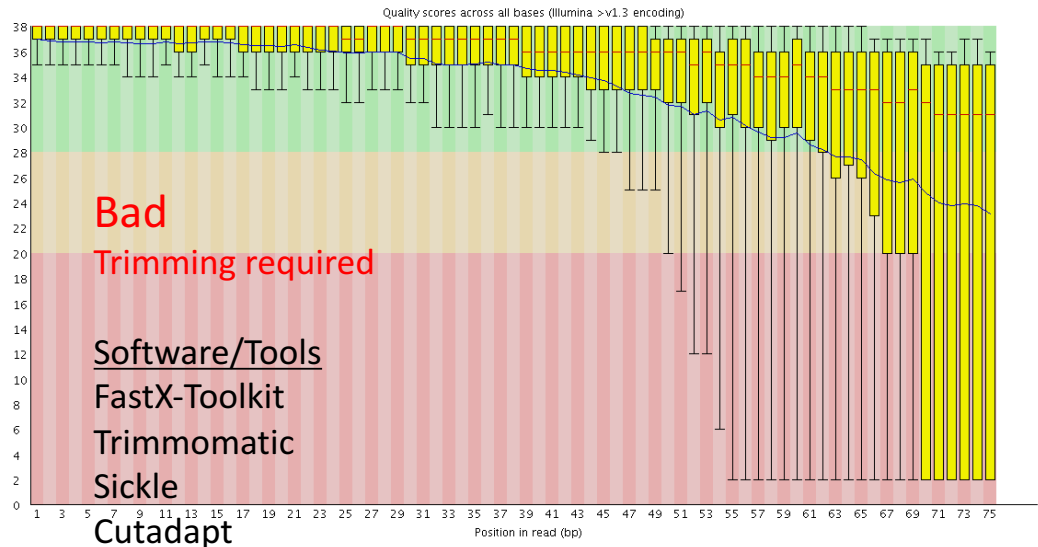
- Trimming: remove bad bases from (end of) read
 - Adaptor sequence
 - Low quality bases
- Filtering: remove bad reads from library
 - Low quality reads
 - Contaminating sequence
 - Low complexity reads (repeats)
 - Short reads
 - Short (< 20bp) reads slow down mapping software
 - Only needed if trimming was performed
- Software
 - Galaxy, many options (NGS: QC and manipulation)
 - Tagdust
 - Many others: <http://seqanswers.com/wiki/Software/list>

Data Quality Assessment

Sequence quality : Quality scores over bases

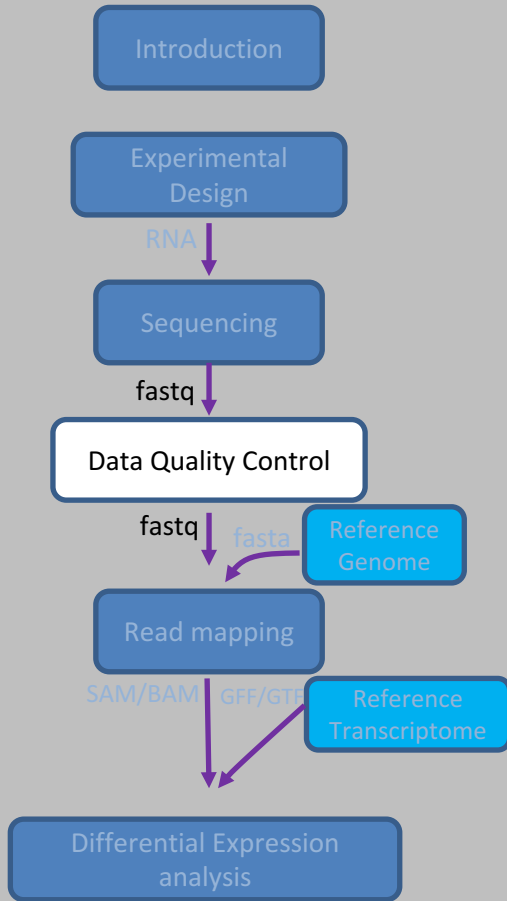


Phred 30 = 1 error/1000bases
Phred 20 = 1 error/100 bases

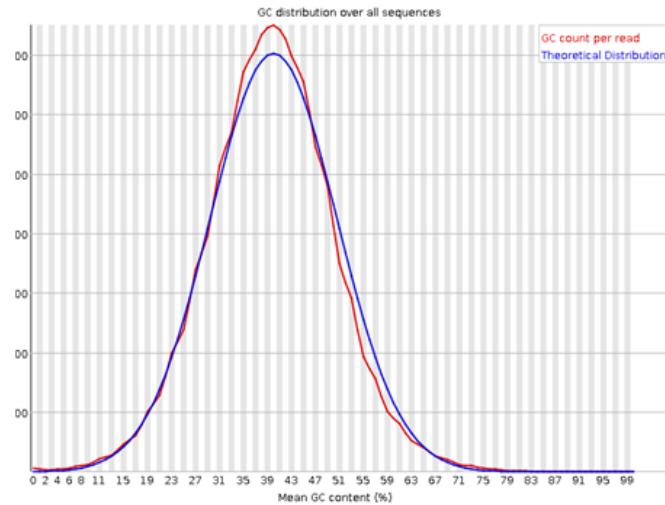


Data Quality Assessment

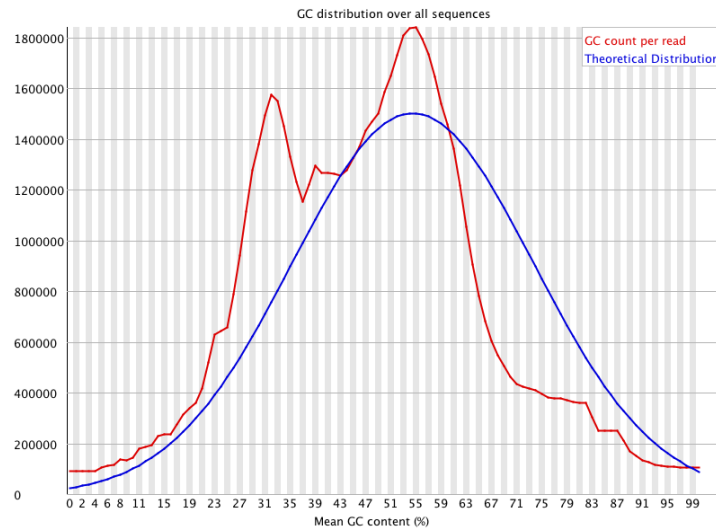
GC Distribution: Acceptable levels depend on Source of sample



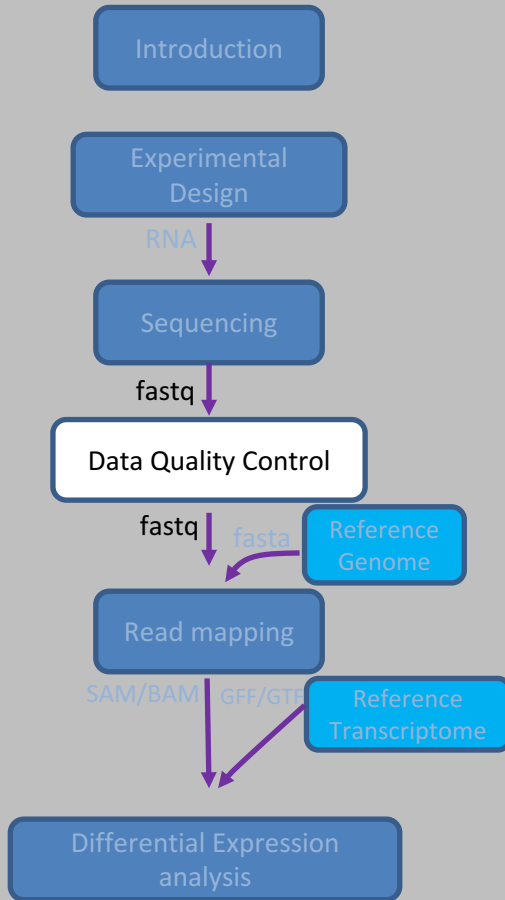
Good



Bad



Data Quality Assessment



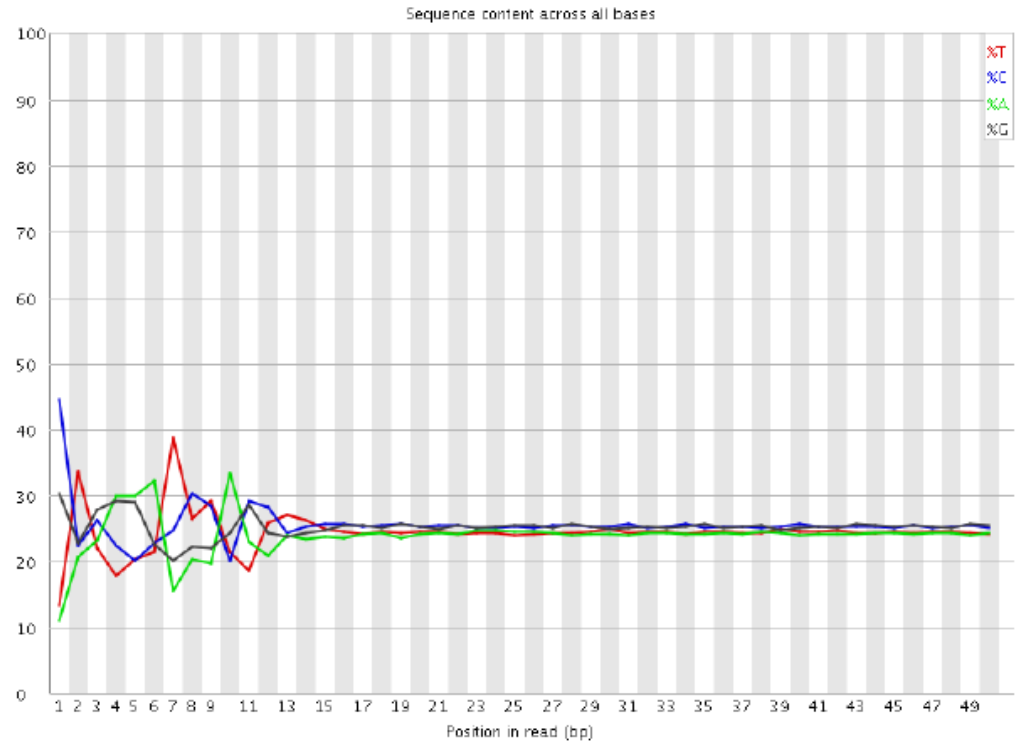
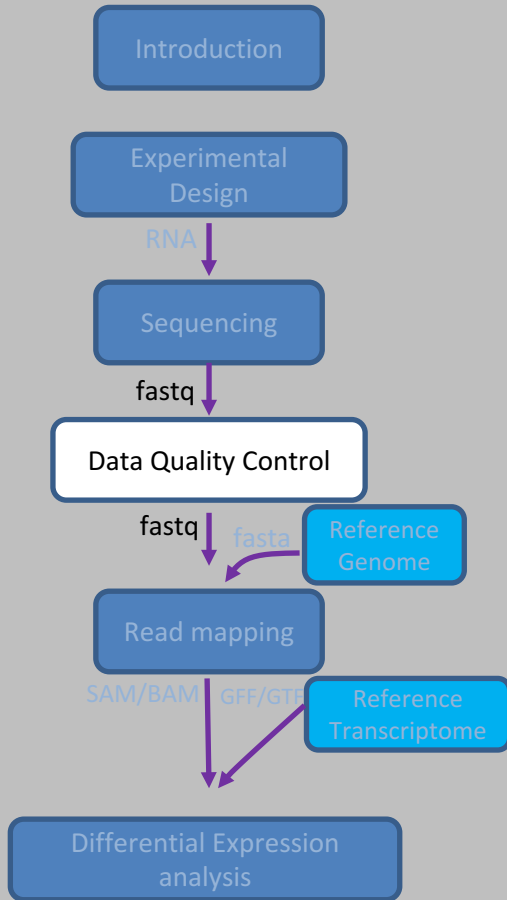
High level of sequencing adapter contamination, trimming needed

Overrepresented sequences

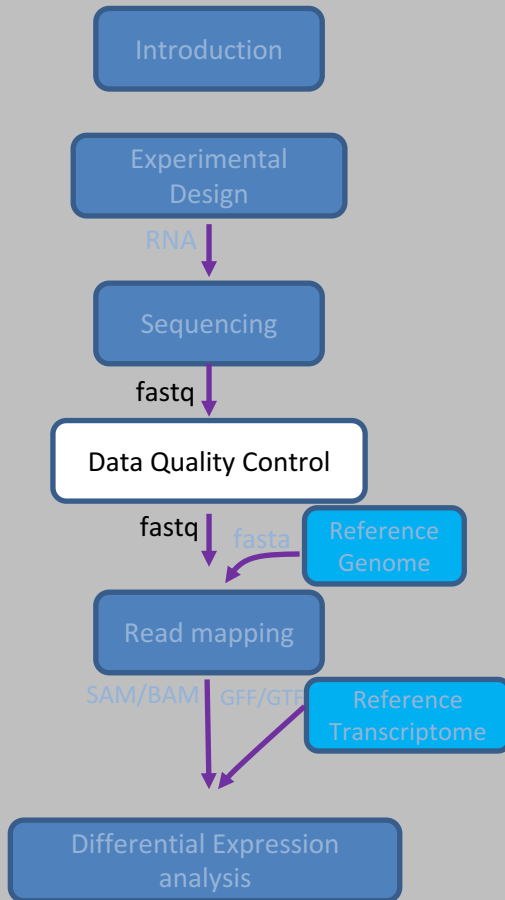
Sequence	Count	Percentage	Possible Source
GTATTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	820428	2.8366639370528275	Illumina Paired End PCR Primer 2 (100% over 43bp)
GTATACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	749728	2.5922157461699773	Illumina Paired End PCR Primer 2 (100% over 44bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCG	648852	2.243432780066747	Illumina Paired End Adapter 2 (100% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAG	176765	0.6111723403310748	Illumina Paired End PCR Primer 2 (97% over 36bp)
ACGTCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	143840	0.4973327832615156	Illumina Paired End PCR Primer 2 (100% over 43bp)
GTATTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	124281	0.42970672717272257	Illumina Paired End PCR Primer 2 (100% over 44bp)
GTATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA	99207	0.34301232917842867	Illumina Paired End PCR Primer 2 (100% over 45bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCCGT	96289	0.33292322279941655	Illumina Paired End PCR Primer 2 (100% over 50bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAG	93842	0.3244626185124245	Illumina Paired End PCR Primer 2 (96% over 33bp)
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	75370	0.26059491013918545	Illumina Paired End PCR Primer 2 (100% over 43bp)
CGTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	63691	0.22021428183196043	Illumina Paired End PCR Primer 2 (100% over 44bp)
ACGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	56765	0.19626734873359242	Illumina Paired End PCR Primer 2 (100% over 46bp)
TACTGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	42991	0.14864317078139472	Illumina Paired End PCR Primer 2 (100% over 43bp)

Data Quality Assessment

Normal sequence bias at beginning of reads due to non-random hybridization of random primers

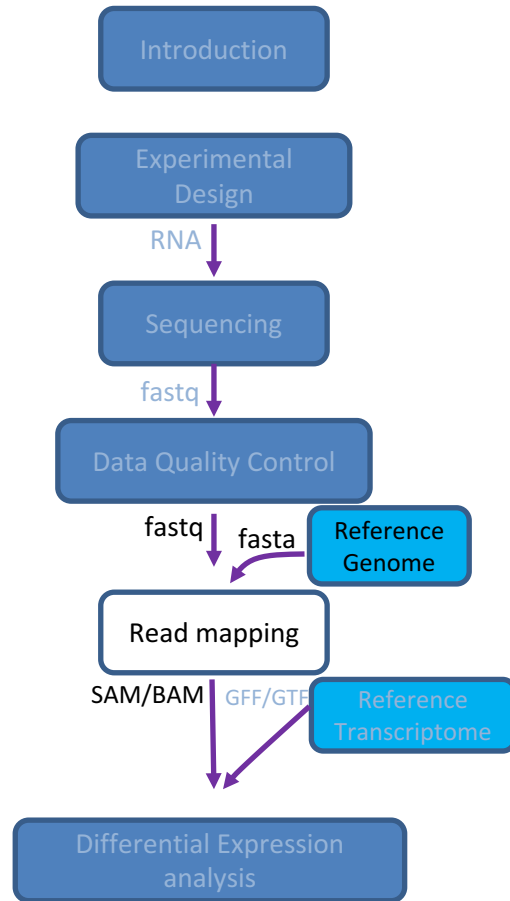


Data Quality Assessment

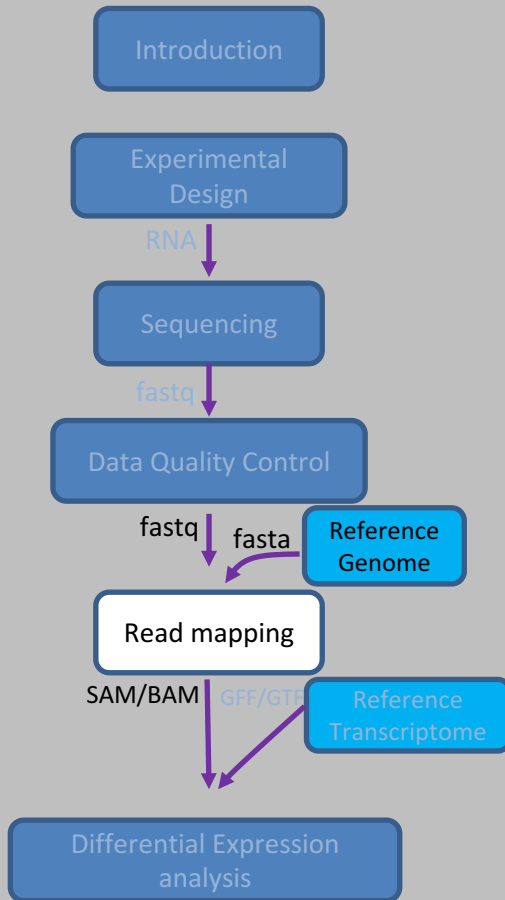


Recommendations:

- Generate quality plots for all read libraries
- Trim and/or filter data if needed
 - Always trim and filter for de novo transcriptome assembly
- Regenerate quality plots after trimming and filtering to determine effectiveness
- Acceptable duplication K-mer or GC content levels are experiment and organism specific but the values should be homogeneous for samples in the same experiment.
- Outliers with >30% disagreement should be discarded.

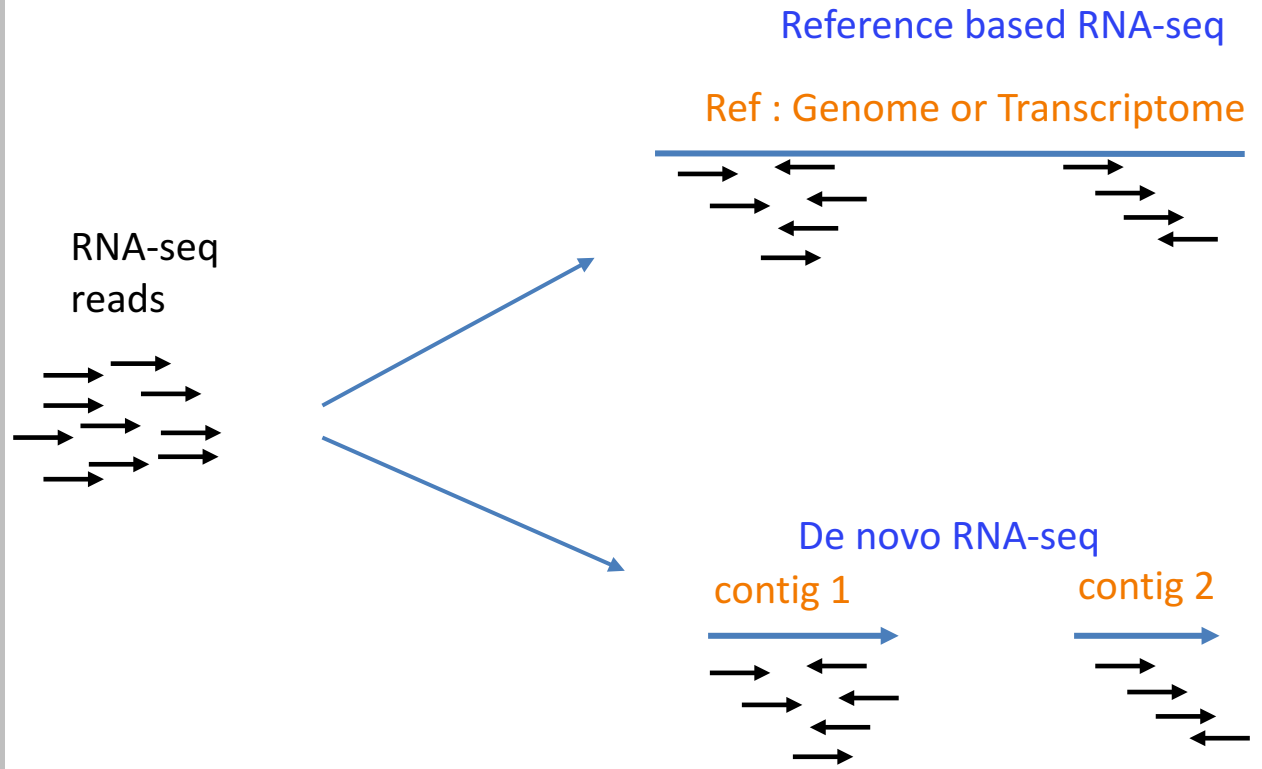
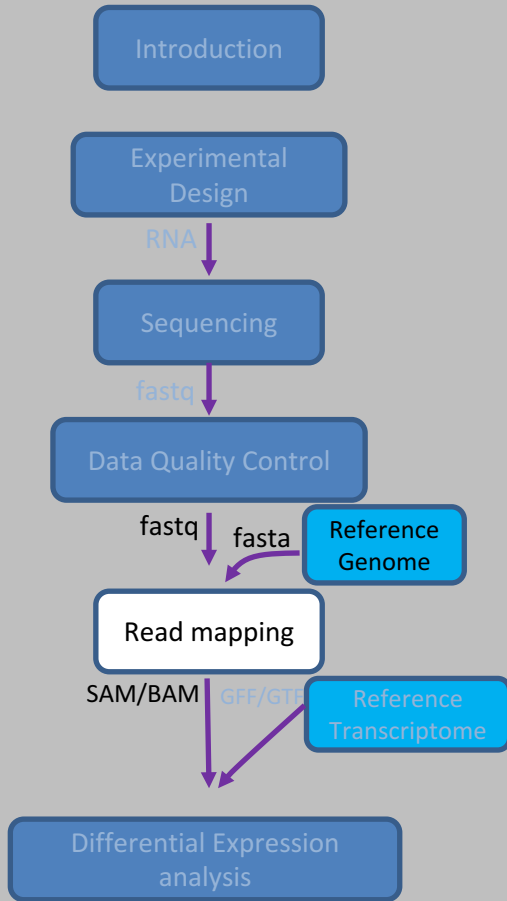


Read Mapping

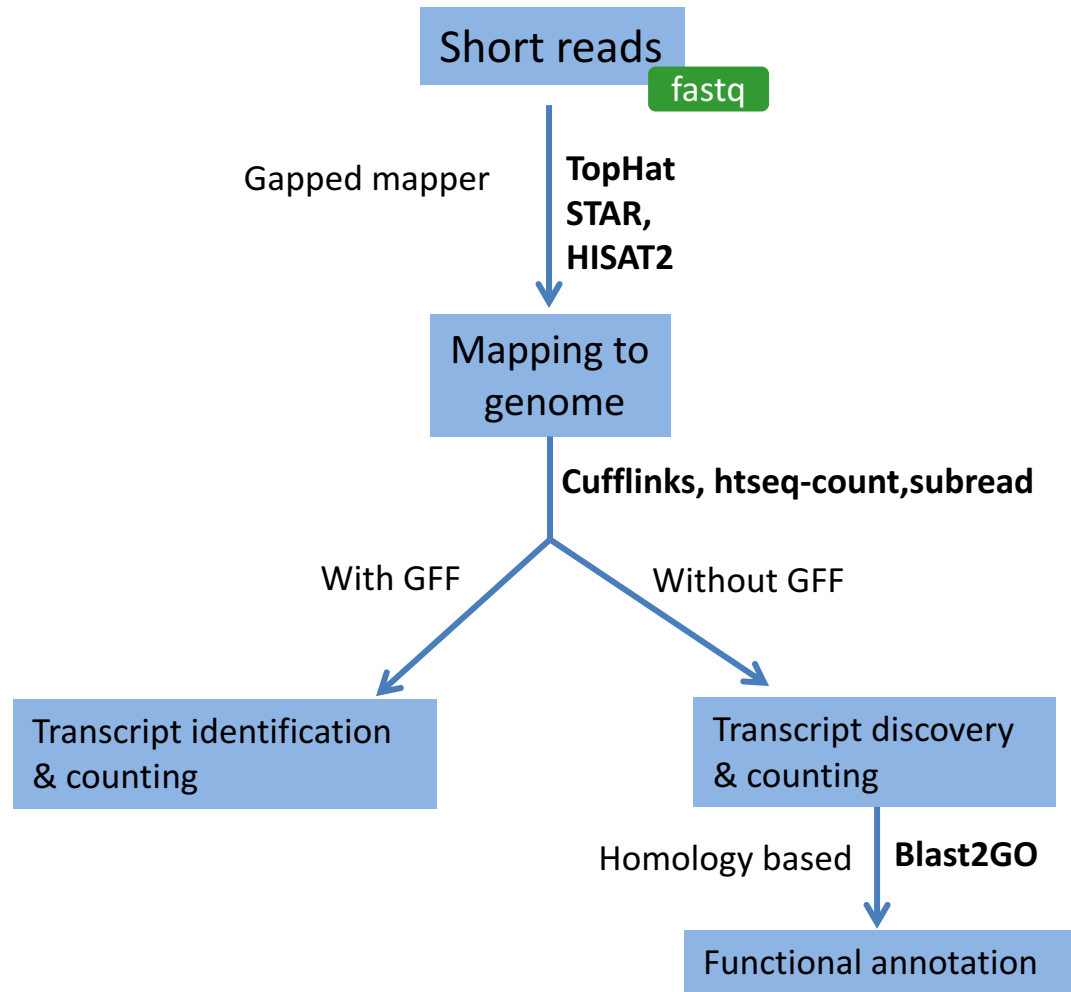
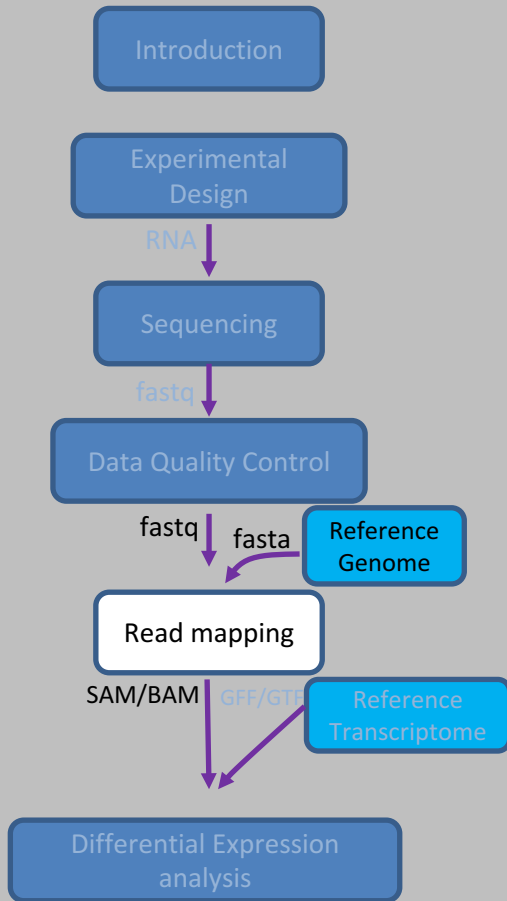


- (1) With reference genome (with/without transcriptome).
- (2) With reference transcriptome.
- (3) Reference free assembly.

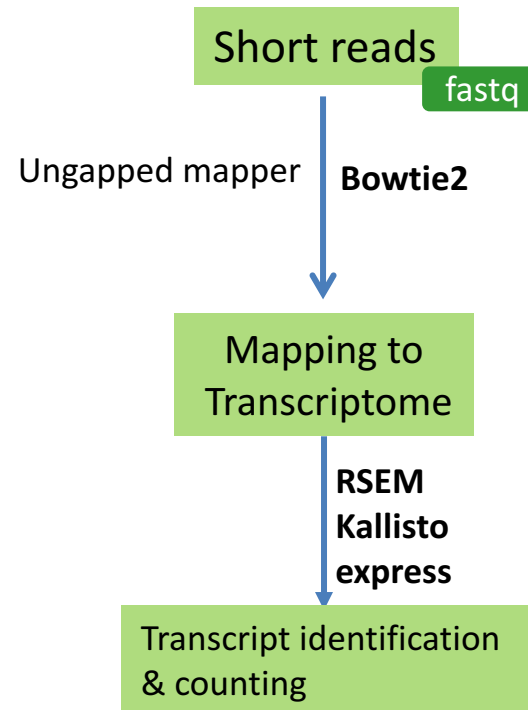
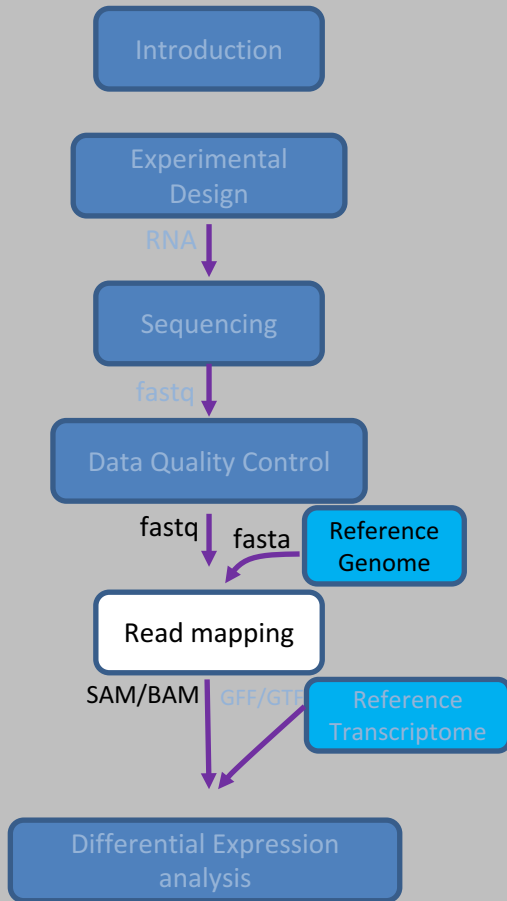
RNA-seq: Assembly vs Mapping



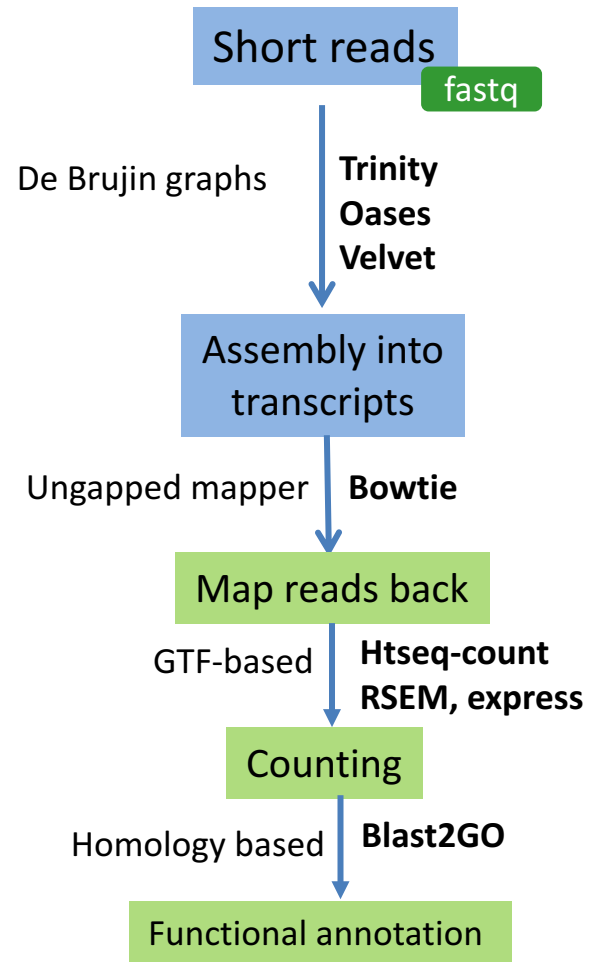
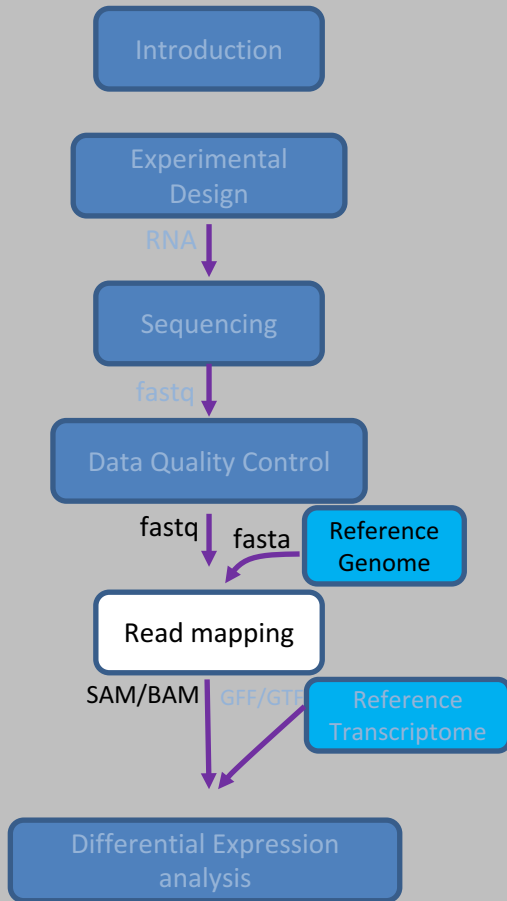
Mapping with reference genome



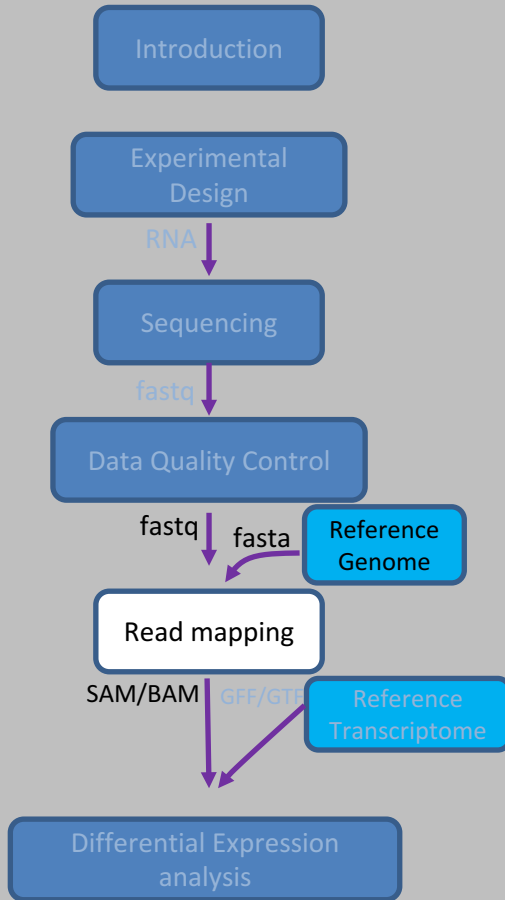
Mapping with reference Transcriptome



Mapping without reference



Alignment tools



- Alignment algorithm must be
 - Fast
 - Able to handle SNPs, indels, and sequencing errors
 - Allow for introns for reference genome alignment (spliced alignment)
- Burrows Wheeler Transform (BWT) mappers
 - Faster
 - Few mismatches allowed (< 3)
 - Limited indel detection
 - Spliced: Tophat, MapSplice
 - Unspliced: BWA, Bowtie
- Hash table mappers
 - Slower
 - More mismatches allowed
 - Indel detection
 - Spliced: GSNAP, MapSplice
 - Unspliced: SHRiMP, Stampy

Alignment tools

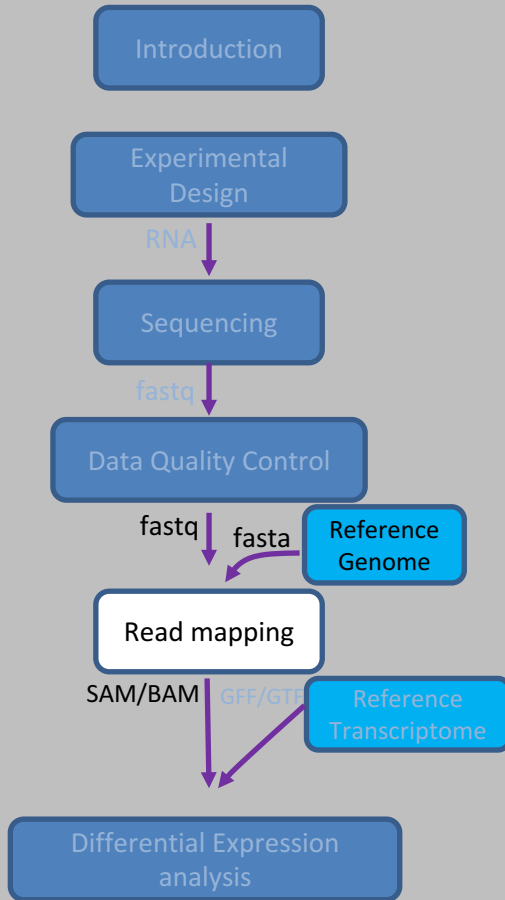
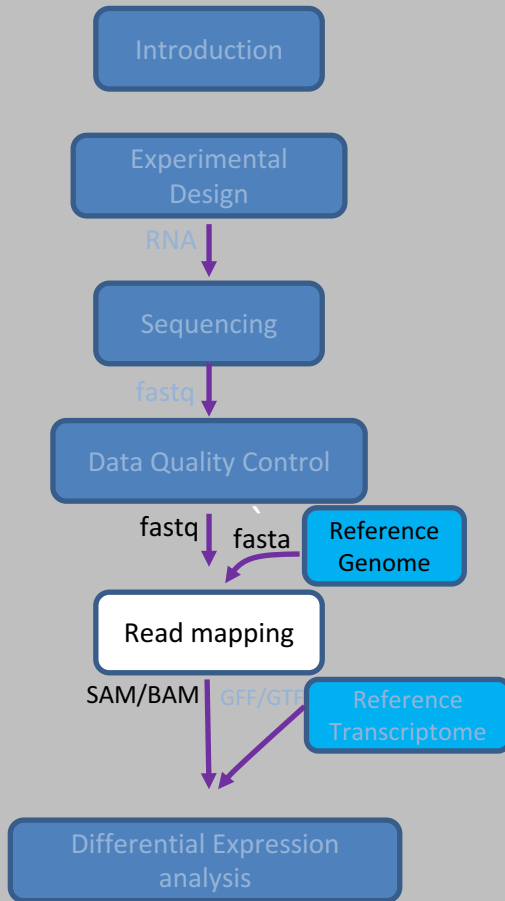


Table 1.1 Overview of common alignment tools

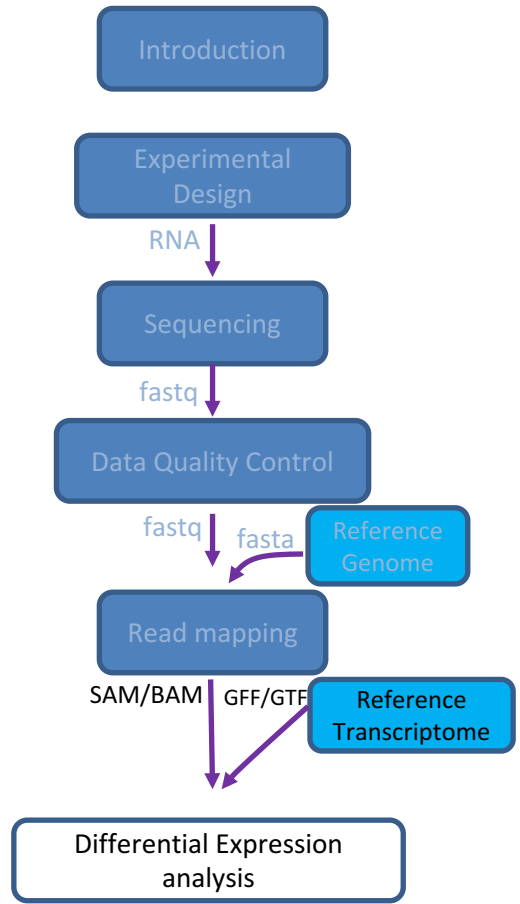
ALIGNERS	Operating system	Language	Alignment algorithm	Input	Output	Paired-end mapping	Splice junction	Read length range
BOWTIE	Unix-based, windows	C++	FM-index based on BWT	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–1 k
BOWTIE2	Unix-based, windows	C++	FM-index based on BWT, dynamic programming	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–5000 k
PALMapper	Unix-based, web interface	C++	Reference indexing	FAST(A/Q)	SAM, BED (x), SHORE	Yes	Yes	12 bp–12 k
STAR	Unix-based	C++	Reference indexing	FAST(A/Q)	SAM	Yes	Yes	15 bp–10 k
BFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, TSV	Yes	No	25–100 bp
GENOME-MAPPER	Unix-based	C	Reference indexing	FAST (A/Q), SHORE	BED, SHORE	No	No	12 bp–2 k
NOVAALIGN	Unix-based	C++	Reference indexing	FAST (A/Q), CSFASTA	SAM	Yes	Yes	1–250 bp
SHRIMP2	Unix-based	Python	Reference indexing	FAST(A/Q)	SAM	Yes	No	30 bp–1 k
SOAP2	Unix-based	C++	BWT + reference indexing	FAST(A/Q)	SAM/BAM	Yes	No	27 bp–1 k
MrFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, DIVET	Yes	No	25 bp–1 k
MAQ	Unix-based	C, C++, Perl	Hashing reads	FAST(A/Q)	TSV	Yes	No	8–63 bp
Mosaik	Unix-based, windows	C++	Reference indexing	FAST(A/Q)	BAM	Yes	No	15 bp–1 k
BWA	Unix-based, windows	C, C++	FM-index based on BWT	FAST(A/Q)	SAM	Yes	No	4–200 bp

Read Mapping



- **Output**

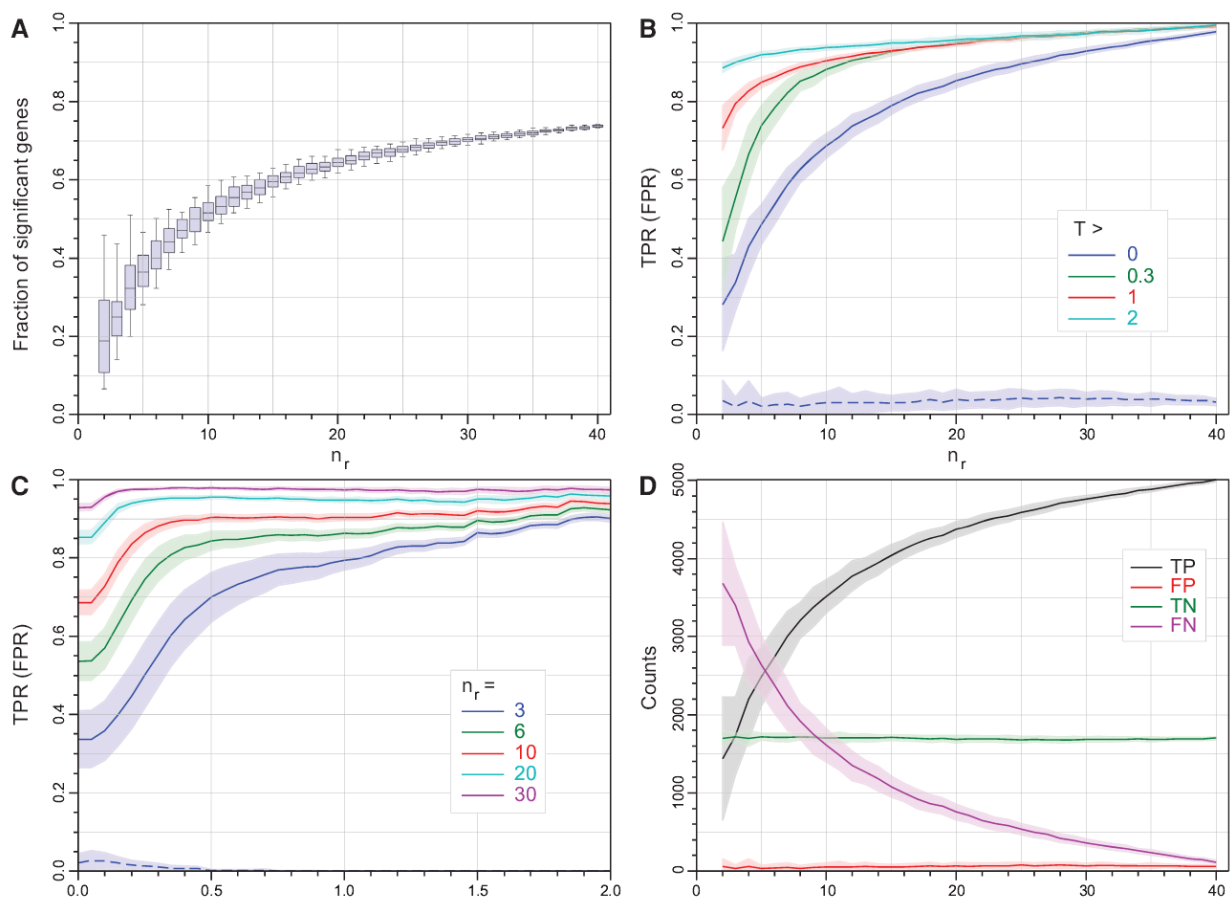
- SAM (text) / BAM (binary) alignment files
 - SAMtools – SAM/BAM file manipulation
- Summary statistics (per read library)
 - % reads with unique alignment
 - % reads with multiple alignments
 - % reads with no alignment
 - % reads properly paired (for paired-end libraries)



How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

NICHOLAS J. SCHURCH,^{1,6} PIETÁ SCHOFIELD,^{1,2,6} MAREK GIERLIŃSKI,^{1,2,6} CHRISTIAN COLE,^{1,6}
 ALEXANDER SHERSTNEV,^{1,6} VIJENDER SINGH,² NICOLA WROBEL,³ KARIM GHARBI,³
 GORDON G. SIMPSON,⁴ TOM OWEN-HUGHES,² MARK BLAXTER,³ and GEOFFREY J. BARTON^{1,2,5}

¹Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
²Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
³Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom
⁴Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
⁵Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom



Name
<i>t</i> -test
log <i>t</i> -test
Mann-Whitney
Permutation
<i>Bootstrap</i>
<i>baySeq</i> ^c
Cuffdiff
<i>DEGseq</i> ^c
<i>DESeq</i> ^c
<i>DESeq2</i> ^c
<i>EBSeq</i> ^c
<i>edgeR</i> ^c
<i>Limma</i> ^c
<i>NOISeq</i> ^c
<i>PoissonSeq</i> ^c
<i>SAMSeq</i> ^c

Statistical properties of edgeR (exact) as a function of log₂(FC) threshold, T , and the number of replicates, n_r

Differential expression

Common tools for differential expression analysis

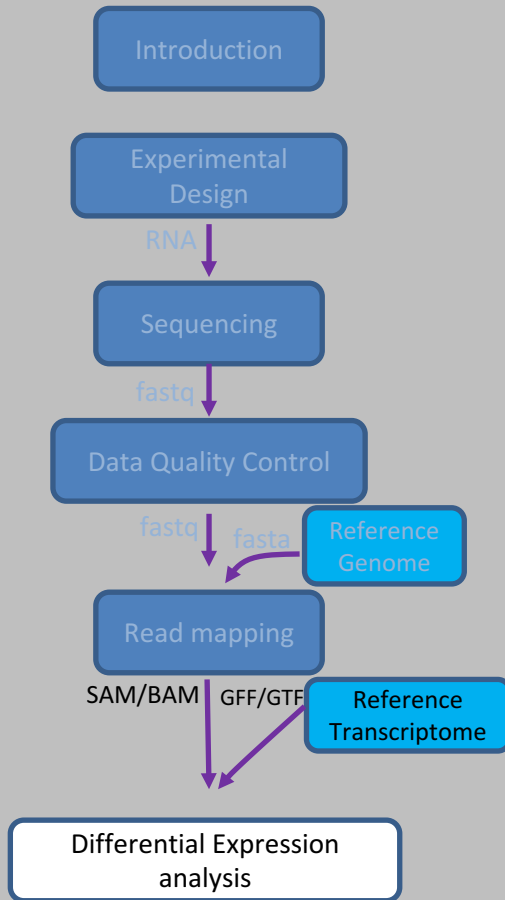
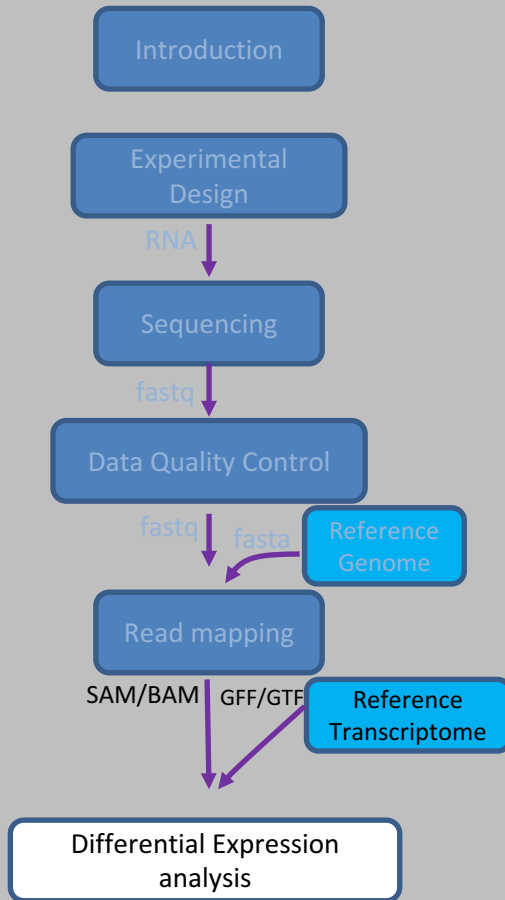


TABLE 8.1 List of (some) Software Tools for Differential Expression Analysis

Software Tool	Type of Software	Analysis Approach	Comment
DESeq	R/Bioconductor package	Count-based (negative binomial)	Considered conservative (low false-positive rate)
edgeR	R/Bioconductor package	Count-based (negative binomial)	Similar to DESeq in philosophy
tweeDESeq	R/Bioconductor package	Count-based (Tweedie distribution family)	More general than DESeq/edgeR, but new and not widely tested
Limma	R/Bioconductor package	Linear models on continuous data	Originally developed for microarray analysis, very thoroughly tested. Need to preprocess counts to continuous values
SAMSeq (samr)	R package	Nonparametric test	Adapted from the SAM microarray DE analysis approach. Works better with more replicates
NOISeq	R/Bioconductor package	Nonparametric test	
CuffDiff	Linux command line tool	Isoform deconvolution + count-based tests	Can give differentially expressed isoforms as well as genes (also differential usage of TSS, splice sites)
BitSeq	Linux command line tool and R package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Also calculates (gene and isoform) expression estimates
ebSeq	R/BioConductor package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Can be used in a pipeline preceded by RSEM expression estimation

Decision tree for software selection



Differentially expressed **exons** => *DEXSeq*

Differentially expressed **isoforms** => *BitSeq, Cuffdiff or ebSeq*

Differentially expressed genes => **Select type of experimental design**

Complex design (more than one varying factor) => *DESeq, edgeR, limma*

Simple comparison of groups => **How many biological replicates?**

More than about 5 biological replicates per group => *SAMSeq*

Less than 5 biological replicates per group => *DESeq, edgeR, limma*

Sources:

Conesa et al, *Genome Biology* 2016 **17**:13

Schurch et al, <https://arxiv.org/abs/1505.02017>

Zeng And Mortazavi *Nat Immunol* 13(9), 802-7

Mortazavi et al *Nature Methods* **volume5**, pages621–628 (2008)