

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Dealing with NGS data: Differential Expression

Club Moderators: Elton Vasconcelos, Euan McDonell, Chew Cheng, and Dapeng Wang

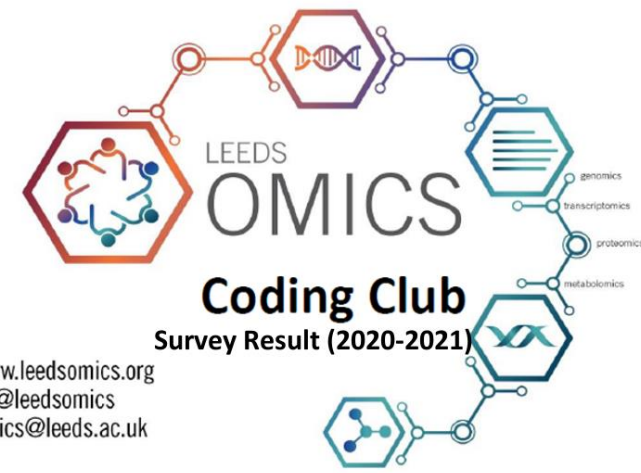
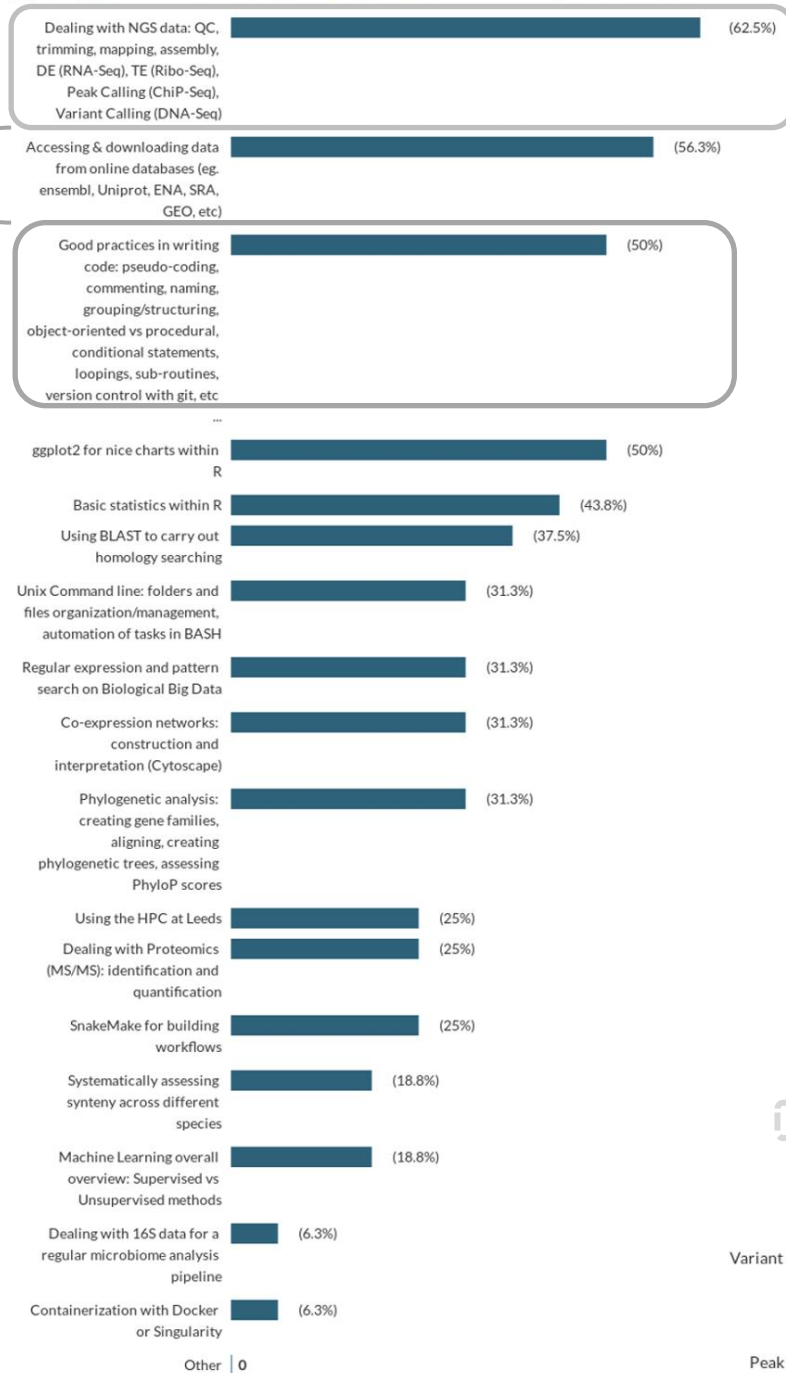
Topics to be addressed on the 2020-21 season - Survey Result

1st, 4th, **6th**, 8th, ... sessions

2nd session

3rd, 5th, 7th, 9th, ... sessions

1 Which of the following topics would you like to attend in our Coding Club sessions?



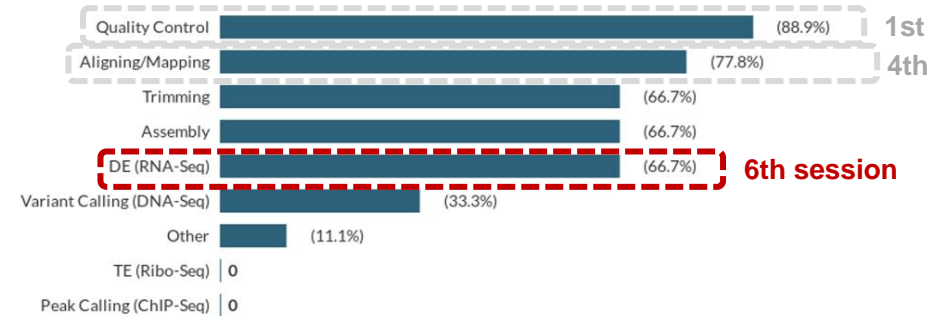
1.b Do you think we should address "Good practices in writing code" topic in more sessions?



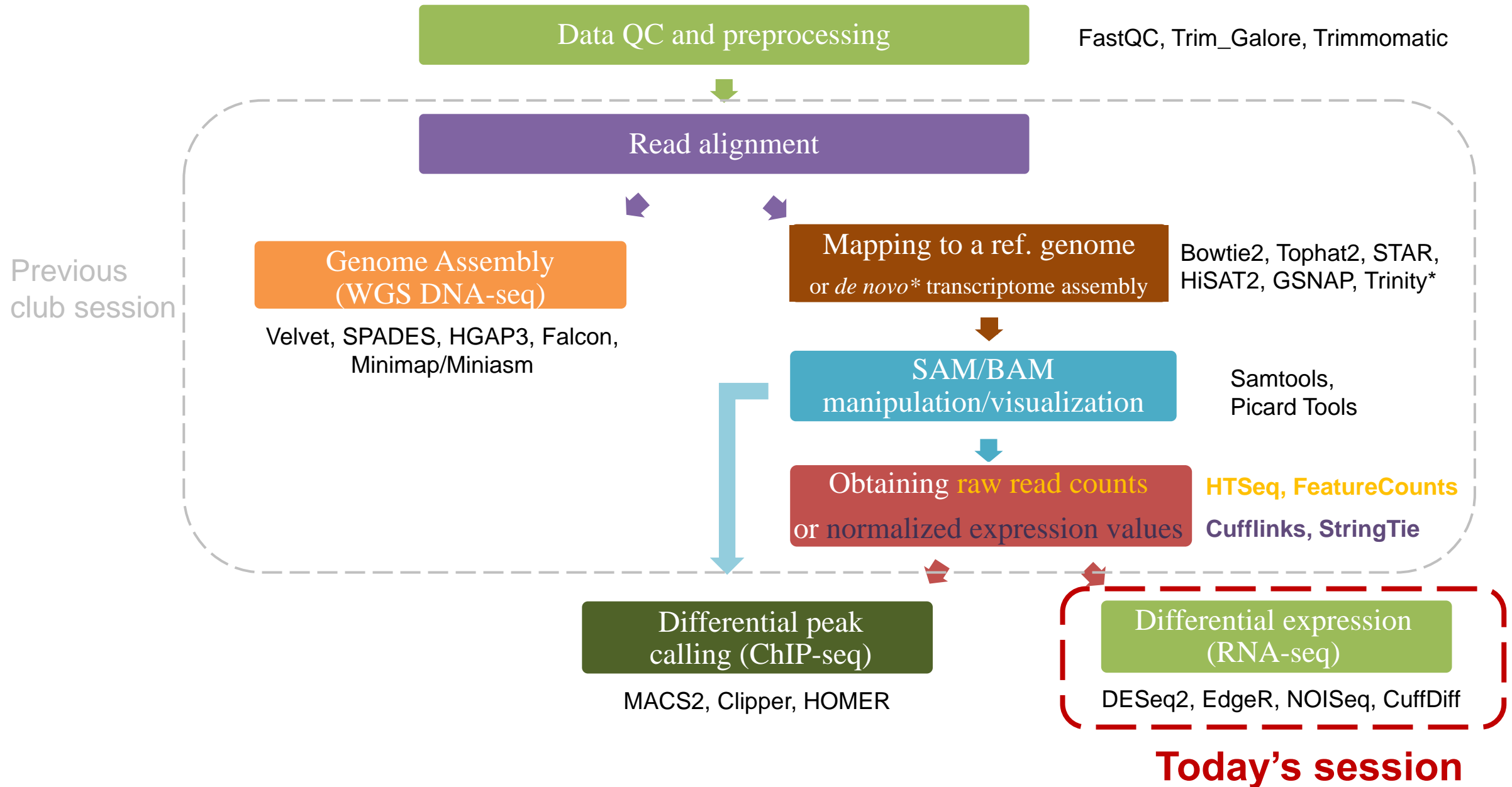
1.c Do you think we should address "Dealing with NGS data" topic in more sessions?

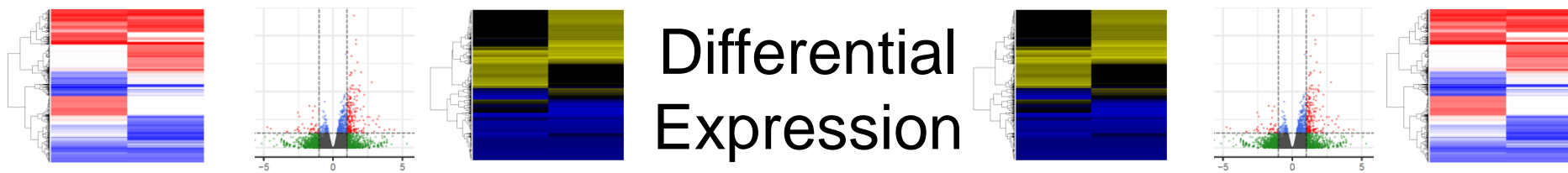


1.c.i Which sub-topics would be of most interest to you?



Important steps on NGS data analysis workflow





Software packages for DE analysis

Method	Version	Reference	Normalization ^a	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	<u>TMM</u> /Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (<u>quantile</u> /TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	<u>RPKM</u> /TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff2 (Cufflinks)	2.0.2-beta	[10]	<u>Geometric</u> (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	<i>t</i> -test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

^a In case of availability of several normalization methods, the default one is underlined.

→ **Important Output Metrics:** $\log_2(\text{FC})$, p-value and FDR provided in most output files

The DESeq2 model

- Perform a median-based normalization to correct for library size and RNA composition bias. (For each gene, calculate the ratio of read counts for each sample to the geometric mean of all read counts for that gene across all samples. The median of these ratios for a sample, called the “size factor”, is used to scale that sample.)
- Use shrinkage estimation for dispersions and fold changes because small numbers of replicates make it impossible to estimate within-group variance reliably.
- Fit negative binomial generalized linear models for each gene and uses the Wald test for significance testing.
- Detect automatically removes these genes with low counts from a analysis.

Prepare the data for DESeq2 analysis

countData: a matrix of non-negative integers

	normal.rep1	normal.rep2	normal.rep3	tumor.rep1	tumor.rep2	tumor.rep3
ENSG00000283047	0	0	0	1	1	0
ENSG00000283023	1	1	1	0	0	3
ENSG00000280341	0	0	1	0	1	1
ENSG00000279442	0	2	0	0	0	0
ENSG00000237299	0	0	0	3	0	3
ENSG00000233408	0	0	0	0	0	1
ENSG00000215268	1	0	1	0	0	0
ENSG00000230471	0	0	0	0	0	1
ENSG00000231565	0	0	0	2	1	2

colData: a DataFrame with at least a single column. Rows of colData correspond to columns of countData.

	condition
normal.rep1	normal
normal.rep2	normal
normal.rep3	normal
tumor.rep1	tumor
tumor.rep2	tumor
tumor.rep3	tumor

design: a formula expressing the variables which will be used in modelling.

The main three steps of running DESeq2

Create a DESeqDataSet object from input. Please note that the colnames of countData must be identical to the rownames of colData.

```
dds <- DESeqDataSetFromMatrix(countData = count.data.keep,  
colData = coldata, design = ~ condition)
```

Perform the differential expression analysis.

```
dds <- DESeq(dds, fitType = "local")
```

Extract a results table.

```
res <- results(dds)
```

Bring your issues on!