Concepts and Main Aspects of RNA-Seq

Wellcome Trust Centre for Human Genetics

25th, 28th and 29th April 2016

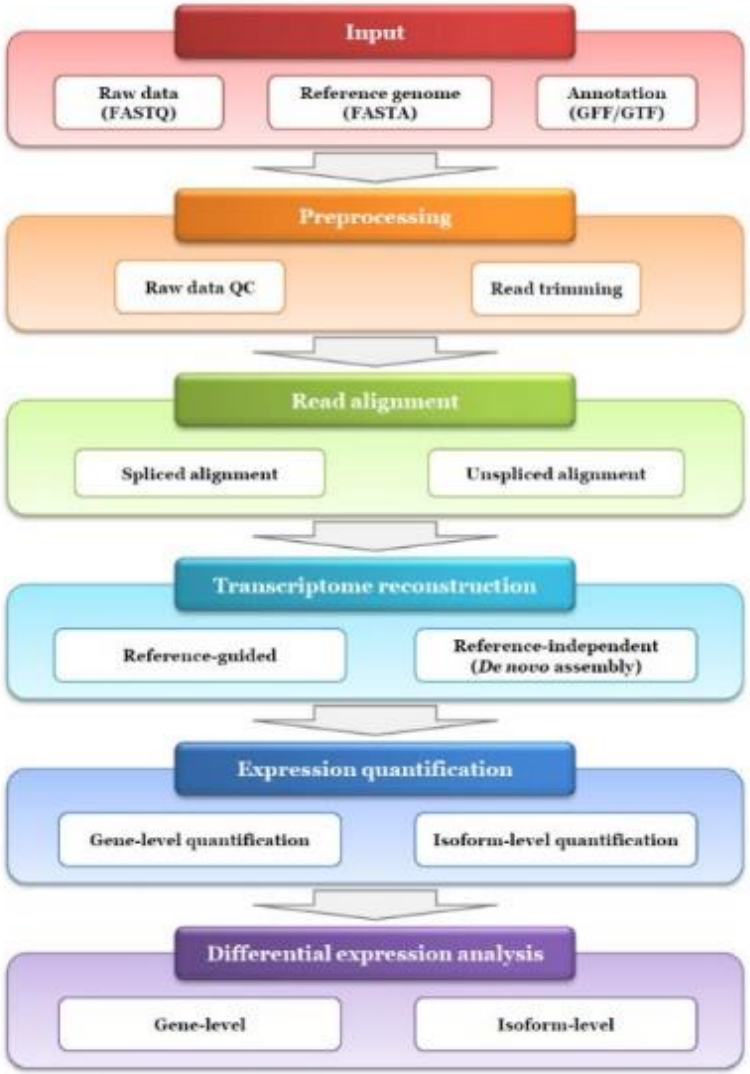
Organised by Helen Lockstone and Irina Pulyakhina

Expression Quantification

Helen Lockstone

28th April 2016

Main Steps in RNA-Seq



- Various tools developed for each step
- Also suites of tools, such as tuxedo, for each step of the process (Bowtie/TopHat, Cufflinks, CuffDiff etc)
- Distinction between isoform reconstruction and gene/transcript abundance

Figure from Yang and Kim. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. Genomics Inform. 2015 Dec; 13(4):119-125

Selecting tools to use

- Many tools available, all claiming to be the best performer – which to choose?
- Review articles can be helpful
- Sensible pipelines should give reasonable results – make sure suitable for your data/purpose, key parameters are set correctly
- Tools can have differing input requirements and output formats – make sure using the expected format for a given tool

Review articles

Kanitz *et al. Genome Biology* (2015) 16:150
DOI 10.1186/s13059-015-0702-5

Genome Biology

RESEARCH Open Access

Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data

Alexander Kanitz[†], Foivos Gy

Teng *et al. Genome Biology* (2016) 17:74
DOI 10.1186/s13059-016-0940-1

Genome Biology

METHOD Open Access

A benchmark for RNA-seq quantification pipelines

Mingxiang Teng[1,2,8], Michael I. Love[1,2], Carrie A. Davis[3], Sarah Djebali[4], Alexander Dobin[3], Brenton R. Graveley[5], Sheng Li[6], Christopher E. Mason[6], Sara Olson[5], Dmitri Pervouchine[4], Cricket A. Sloan[7], Xintao Wei[5], Lijun Zhan[5] and Rafael A. Irizarry[1,2*]

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał W[12] and Ali Mortazavi[16,17*]

Kanitz *et al. Genome Biology* (2015) 16:150
DOI 10.1186/s13059-015-0702-5

Genome Biology

RESEARCH Open Access

Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data

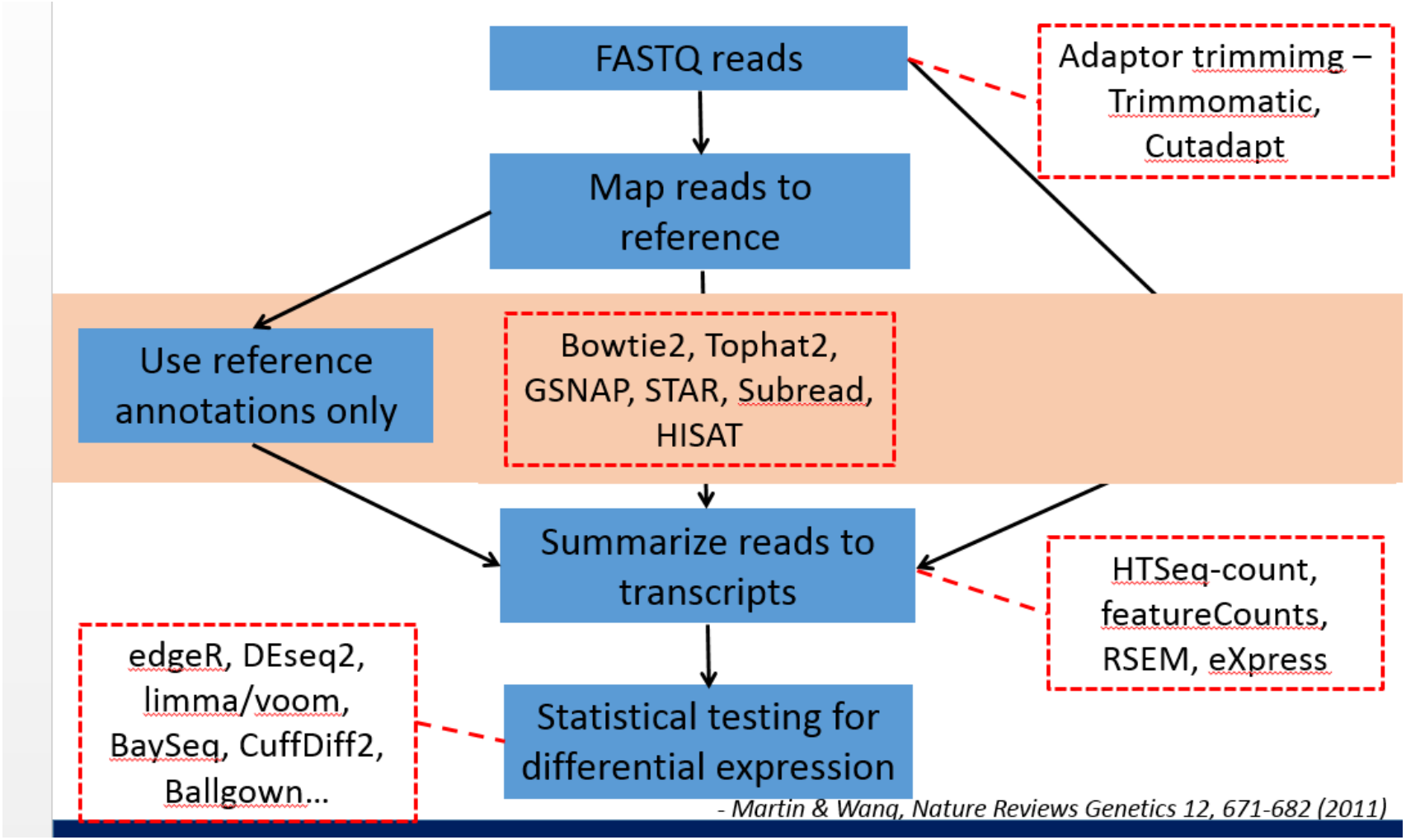
Alexander Kanitz[†], Foivos Gypas[†], Andreas J. Gruber, Andreas R. Gruber, Georges Martin and Mihaela Zavolan[*]

An important point

“During the course of this study we discovered a number of assumptions that the programs tacitly made and that affected the interpretation of the results. Therefore, a specific recommendation that we can make to developers is to ensure that sufficiently detailed information on input requirements, potential pitfalls and the implication of specific options (ideally including usage examples) is provided.”

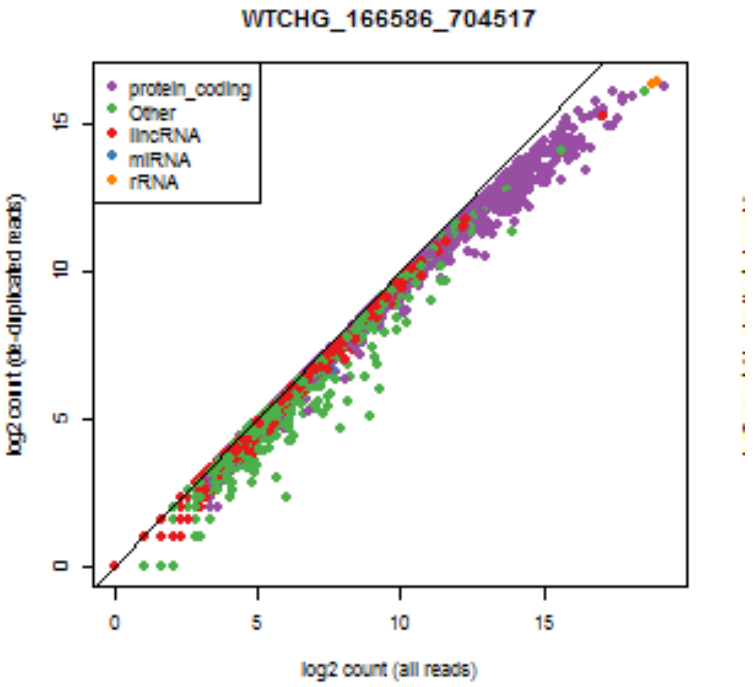
Kanitz et al. Genome Biology (2015) 16:150

Overview of RNA-Seq tools



General mapping issues

- Multi-mapped reads (map to more than one location)
  - Gene families, pseudogenes, repeat regions, MHC region
- Potential PCR duplicates (multiple reads mapping to identical position)
  - May exclude some genuine duplicate fragments, especially for high expressed genes

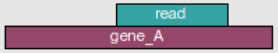
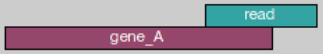
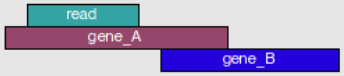
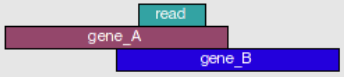


Expression quantification

- Process of using aligned reads to quantify gene/transcript abundance
  - Count-based (HTSeq, featureCounts)
    - Find overlap of aligned reads and known annotated features
    - output is raw counts
  - Model-based (e.g. Cufflinks, RSEM)
    - Try to find the optimal set of isoforms and their relative abundances from the observed data (aligned reads)
    - Short reads tend to map to multiple isoforms, hence the need for probabilistic models
    - Output usually RPKM/FPKM or TPM

HTSeq

- Python based suite of scripts/tools including ‘htseq-count’
  - Annotations from Ensembl/RefSeq etc
  - Reads overlapping gene features counted
  - Careful setting of parameters e.g. strandedness option in particular
  - Usage: htseq-count [options] <alignment_file (SAM)> <gff_file>

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Raw count table

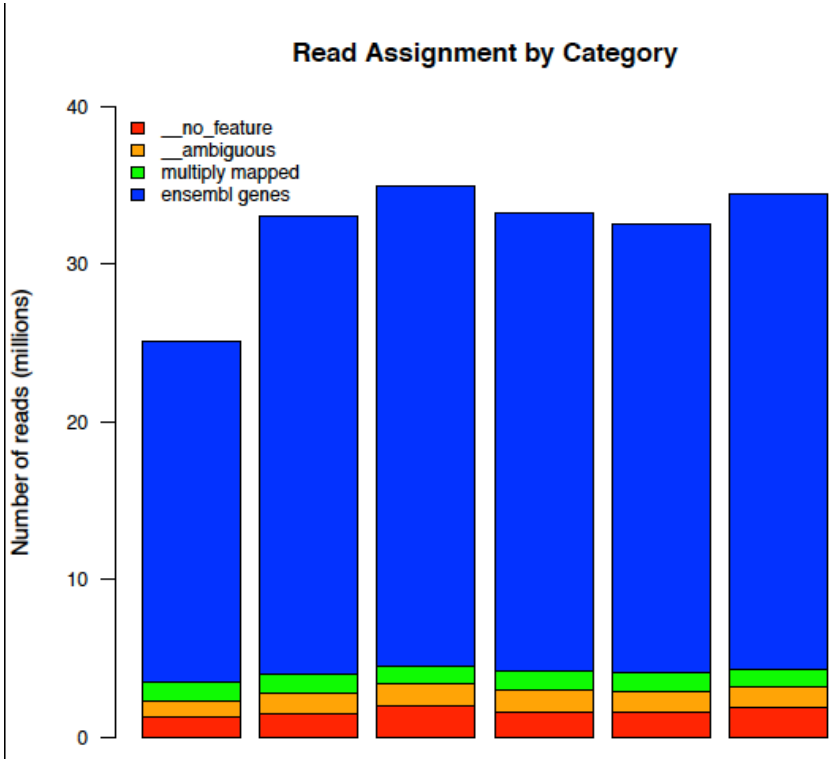
```
> head(counts)
                 Hypoxia1 Hypoxia2 Hypoxia3 Normoxia1 Normoxia2 Normoxia3
ENSG00000000003     1451     1770     1711      1036      1294      1411
ENSG00000000005        0        1        0         0         0         0
ENSG00000000419     2213     2433     2728      4074      4249      4278
ENSG00000000457      866      863      756       775       936       861
ENSG00000000460      841      952      977      1235      1309      1296
ENSG00000000938        0        3        0         1         3         3
> tail(counts)
                   Hypoxia1 Hypoxia2 Hypoxia3 Normoxia1 Normoxia2 Normoxia3
ENSG00000259772           1        6        5         0         0         0
ENSG00000259773           1        0        0         0         0         0
ENSG00000259774           8       14       11         5        11        11
alignment_not_unique  3509317  4004009  3405804   3668337   4568935   4499171
ambiguous             2776001  3375340  2832980   3388991   3821237   3828028
no_feature            6939028  7857177  6202574   4505149   6517785   5738978
>
```

RPKM (FPKM) – read(pairs) per kilobase of gene model per million reads. An early metric to generate ‘comparable’ values

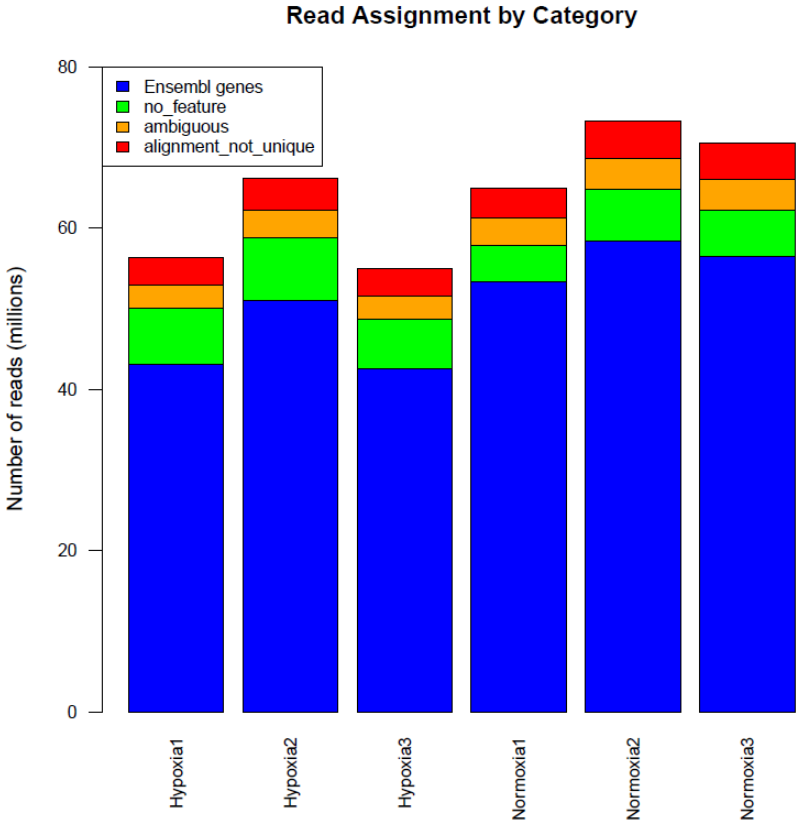
TPM – transcripts per kilobase million. Normalises for gene length first, then depth. Same TPM value in 2 different samples means the same proportion of reads map to that gene in both samples. Not the case with RPKM/FPKM

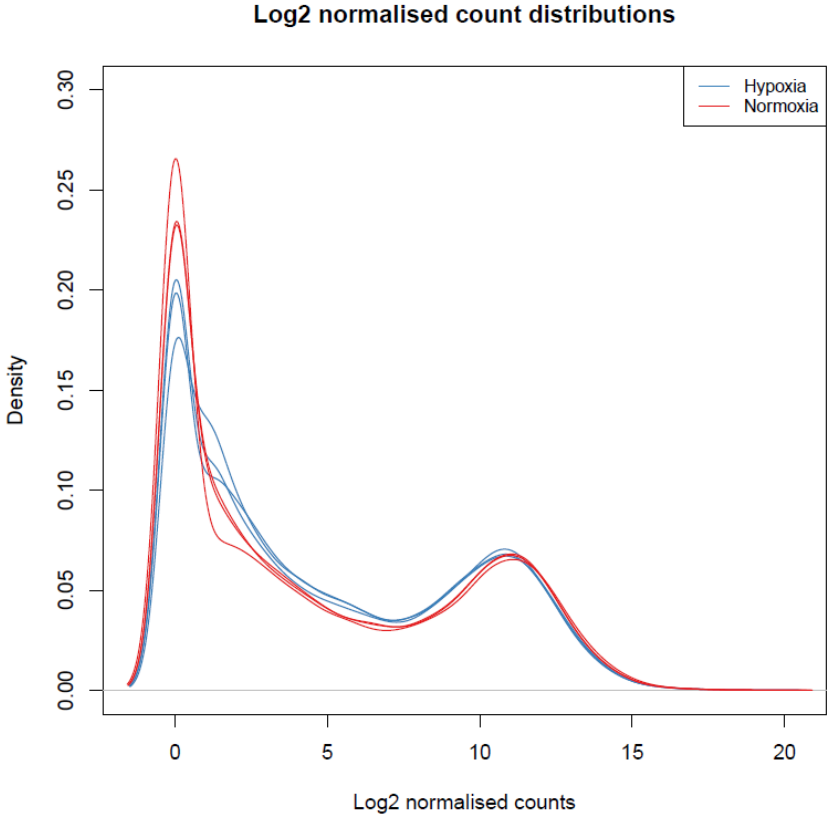
Normalisation

- Sequencing depth
- RNA composition effect – e.g. high expressed genes in one group, library complexity – affects way genes are sampled
- Counts obtained are influenced by what is sequenced

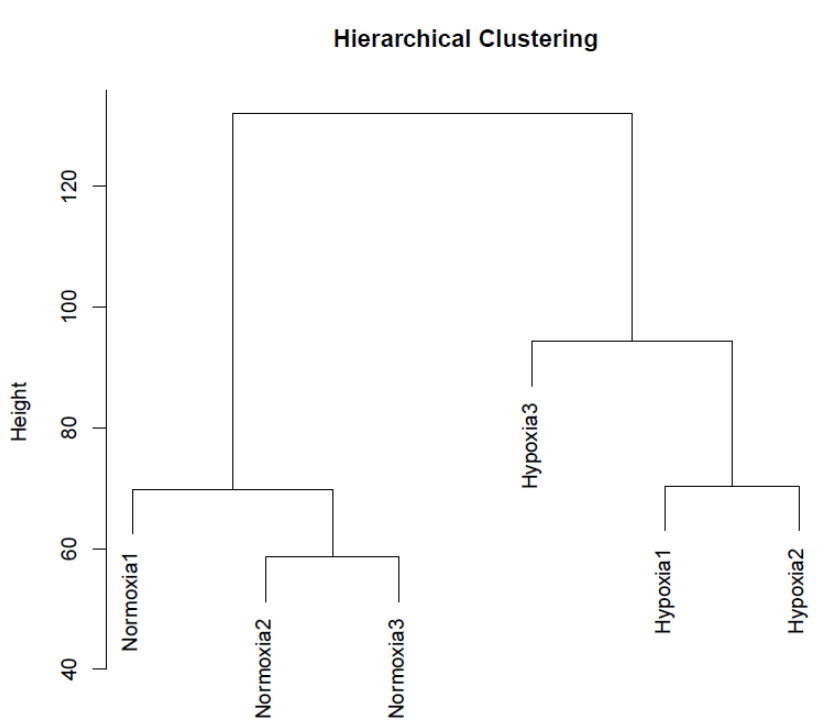


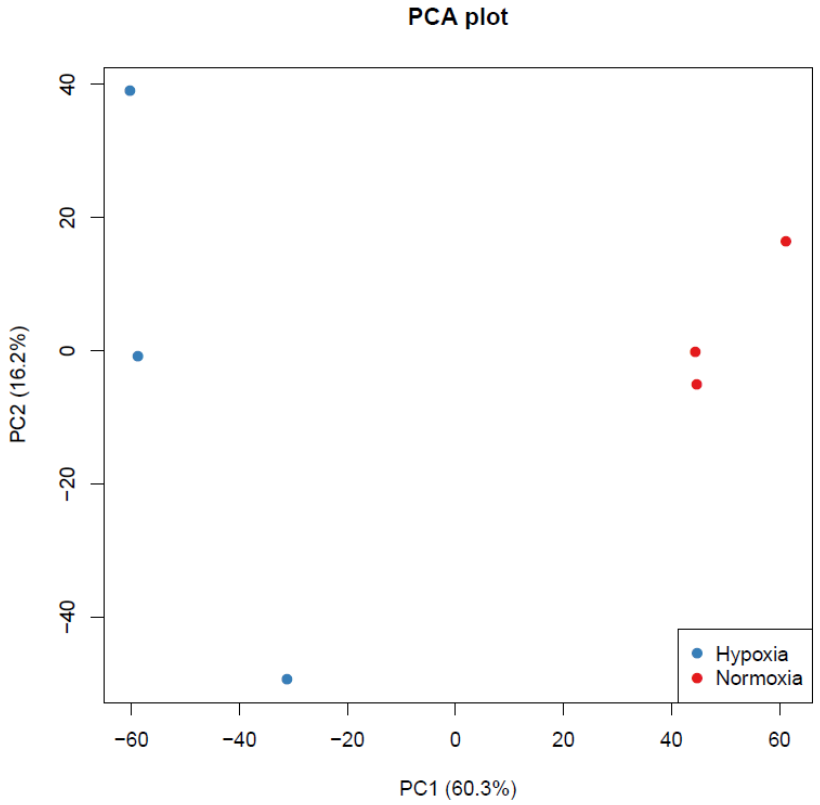
QC plots


Read Assignment by Category
Ensembl genes
no_feature
ambiguous
alignment_not_unique
Number of reads (millions)
0
20
40
60
80
Hypoxia1
Hypoxia2
Hypoxia3
Normoxia1
Normoxia2
Normoxia3


Log2 normalised count distributions
Hypoxia
Normoxia
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
5
10
15
20
Log2 normalised counts

Exploratory plots of data





Input raw counts were transformed in DESeq package to make more suitable for input to clustering algorithms and PCA

Concluding Remarks

- Each step has a variety of issues/complications, sometimes subtle
- Many choices made in processing data – choice of tools, parameters etc - an alternative pipeline may give different results

- If resulting genelist or findings are followed up with further work and validated, the limitations/biases various steps in the process to get that point become less of a concern
- The end of the RNA-Seq analysis is only the beginning….