RNA-Seq Analysis

Simon Andrews, Laura Biggins
simon.andrews@babraham.ac.uk
@simon_andrews
v2020-01-02
RNA-Seq Libraries

1. rRNA depleted mRNA
2. Fragment
3. Random prime + RT
4. 2\textsuperscript{nd} strand synthesis (+ U)
5. A-tailing
6. Adapter Ligation
7. (U strand degradation)
8. Sequencing
Reference based RNA-Seq Analysis

QC → Trimming → Mapping

Statistical Analysis ↔ Quantitation ↔ Mapped QC
Sequence Data Processing
Raw Sequence Quality Control
FastQC
FastQ Format Data

@HWUSI-EAS611:34:6669YAXXX:1:1:5069:1159 1:N:0:
TCGATAATACCCTTTTTTTTTCCGTTTGATGTTGATACCATT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

@HWUSI-EAS611:34:6669YAXXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGBD@GGGD4@CA3CGG>DDD:D,B

@HWUSI-EAS611:34:6669YAXXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTGGGGCCT
+
:GBGGGGGGGGGGGGGGGDDGDGGGGGDDHHDGHDGHDGHBGG:GG
QC: Raw Data

- Sequence call quality
QC: Raw Data

• Sequence bias

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen\textsuperscript{1,*}, Steven E. Brenner\textsuperscript{2} and Sandrine Dudoit\textsuperscript{1,3}

\textsuperscript{1}Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358,
\textsuperscript{2}Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and
\textsuperscript{3}Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Published online 14 April 2010
Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010
QC: Raw Data

• Duplication level
Adapters and Trimming
Trimming Adapters

![Graph showing trimming adapters]

- Illumina Universal Adapter
- Illumina Small RNA Adapter
- NkOera Transposase Sequence
Trimming Adapters
Mapping to a reference
Mapping

- Exon 1
- Exon 2
- Exon 3

Genome

Simple mapping within exons

Mapping between exons

Spliced mapping
RNA-Seq Mapping Software

• HiSat2 (https://ccb.jhu.edu/software/hisat2/)

• Star (http://code.google.com/p/rna-star/)

• Tophat (http://tophat.cbcb.umd.edu/)
HiSat2 pipeline

Reference FastA files → Indexed Genome

Reference GTF Models → Pool of known splice junctions

Reads (fastq) → Maps with known junctions

Maps with known junctions → Report

Maps convincingly with novel junction? → Add

Maps convincingly with novel junction? → Yes → Report

Maps convincingly with novel junction? → No → Discard

Yes → Report

Add → Pool of known splice junctions

No → Discard
Mapped Data QC
Mapping Statistics

Time loading forward index: 00:01:10
Time loading reference: 00:00:05
Multiseed full-index search: 00:20:47
24548251 reads; of these:
  24548251 (100.00%) were paired; of these:
    1472534 (6.00%) aligned concordantly 0 times
    21491188 (87.55%) aligned concordantly exactly 1 time
    1584529 (6.45%) aligned concordantly >1 times
94.00% overall alignment rate
Time searching: 00:20:52
Overall time: 00:22:02
Exercise: RNA-Seq QC and Data Processing
Running programs in Linux

• Open a shell (text based OS interface)

• Type the name of the program you want to run
  – Add on any options the program needs
  – Press return - the program will run
  – When the program ends control will return to the shell

• Run the next program!
Running programs

babraham@babraham-VirtualBox:~$ ls
Desktop  Documents  Downloads  examples.desktop
Music  Pictures  Public  Templates  Videos

babraham@babraham-VirtualBox:~$

- Command prompt - you can't enter a command unless you can see this
- The command we're going to run (ls in this case, to list files)
- The output of the command - just text in this case
The structure of a unix command

```
ls  -ltd --reverse  Downloads/  Desktop/  Documents/
```

Program name  Switches  Data (normally files)

Each option or section is separated by spaces. Options or files with spaces in must be put in quotes.
Command line switches

• Change the behaviour of the program
• Come in two flavours (each option usually has both types available)
  – Minus plus single letter (eg \(-x\) \(-c\) \(-z\))
    • Can be combined (eg \(-xcz\))
  – Two minuses plus a word (eg \(--extract\) \(--gzip\))
    • Can't be combined
• Some take an additional value, this can be an additional option, or use an = to separate (it's up to the program)
  – \(--f somfile.txt\) (specify a filename)
  – \(--width=30\) (specify a value)
Specifying file paths

• Relative paths from whichever directory you are currently in
  – If I'm in /home/simon/Documents/
  – Data/big_data.fq.gz
    • is the same as /home/simon/Documents/Data/big_data.fq.gz

• Move to the directory with the data and just use file names
  – cd Data
  – big_data.fq.gz
Command line completion

• Most errors in commands are typing errors in either program names or file paths

• Shells (ie BASH) can help with this by offering to complete path names for you

• Command line completion is achieved by typing a partial path and then pressing the TAB key (to the left of Q)
Command line completion

Actual files in a folder:

Desktop
Documents
Downloads
examples.desktop
Music
Pictures
Public
Templates
Videos

If I type the following and press tab:

- De [TAB] will complete to Desktop as it is the only option
- T [TAB] will complete to Templates as it is the only option
- Do [TAB] will no nothing (just beep) as it is ambiguous
- Do [TAB] [TAB] will show Documents and Downloads since those are the only options
- Do [TAB] [TAB] c [TAB] will complete to Documents

You should ALWAYS use TAB completion to fill in paths for locations which exist so you can't make typing mistakes
(it obviously won't work for output files though)
Debugging Tips

• If anything (except the splice site extraction) completes almost immediately then it didn't work!

• Look for errors before asking for help. They will either be
  – The last piece of text before the program exited
  – The first piece of text produced after it started (followed by the help file)

• To see if a program is running go to another shell and look at the last file produced to see if it's growing

• Programs which are stuck can be cancelled with Control+C
Some useful commands

`cd mydir` Change directory to `mydir`

`ls -l trh` List files in the current directory, show details and put the newest files at the bottom

`less x.txt` View the `x.txt` text file
   - Return = down one line
   - Space = down one page
   - q = quit
Data Visualisation and Exploration
Viewing Mapped Data

- Reads over exons
- Reads over introns
- Reads in intergenic
- Strand specificity
SeqMonk RNA-Seq QC (good)
SeqMonk RNA-Seq QC (bad)
SeqMonk RNA-Seq QC (bad)
Look at poor QC samples
Duplication (again)
Duplication (good)
Duplication (bad)
Fixing Duplication?

• If duplication is biased (some genes more than others)
  – Can’t be ‘fixed’ – can still analyse but be cautious

• If it’s unbiased (everything is duplicated)
  – Doesn’t affect quantitation
  – Will affect statistics

  – Can estimate global level and correct raw counts
Quantitation

Exon 1

Exon 2

Exon 3

Splice form 1

Splice form 2

Definitely splice form 1

Definitely splice form 2

Ambiguous
Simple Quantitation - Forget splicing

• Count read overlaps with exons of each gene
  – Consider library directionality
  – Simple
  – Gene level quantitation
  – Many programs
    • Seqmonk (graphical)
    • Feature Counts (subread)
    • BEDTools
    • HTSeq
Analysing Splicing

• Estimate transcript level expression
  – Cufflinks, RSEM, bitSeq etc.

• Analyse junction usage directly
  – Counts + DESeq/EdgeR

• Compare splicing decision ratios
  – rMATS, logistic regression
RPKM / FPKM / TPM

• **RPKM** (Reads per kilobase of transcript per million reads of library)
  – Corrects for total library coverage
  – Corrects for gene length
  – Comparable between different genes within the same dataset

• **FPKM** (Fragments per kilobase of transcript per million fragments of library)
  – Only relevant for paired end libraries
  – Pairs are not independent observation
  – Effectively halves raw counts

• **TPM** (transcripts per million)
  – Normalises to transcript copies instead of reads
  – Corrects for cases where the average transcript length differs between samples
Visualising Expression and Normalisation

Linear

Log2
Visualising Normalisation
Visualising Normalisation
Size Factor Normalisation

• RPKM/FPKM/TPM all use total number of reads as the fixed value between samples

• This can be a terrible measure, as it might not be a good representation for sampling depth

• Might need a more robust reference point
Size Factor Normalisation

• Make an ‘average’ sample from the mean of expression for each gene across all samples

• For each sample calculate the distribution of differences between the data in that sample and the equivalent in the ‘average’ sample

• Use the median of the difference distribution to normalise the data
Normalisation – Coverage Outliers
Normalisation – DNA Contamination
Exploratory Analyses

- Time to *understand* your data
  - Behaviour of raw data and annotation
  - Clustering of samples (PCA / tSNE etc)
  - Pairwise comparisons of samples and groups
  - Are expected effects present (e.g., KO)?
  - Can I see obvious changes?
  - Are the changes convincing?
  - Can I validate other aspects of the samples (e.g., sex)
Differential Expression
Differential Expression

Mapped data

- Normalised expression (+ confidence)
  - Continuous stats
- Raw counts
  - Binomial stats
DE-Seq2 binomial Stats

• Are the counts we see for gene X in condition 1 consistent with those for gene X in condition 2?

• Size factors
  – Estimator of library sampling depth
  – More stable measure than total coverage
  – Based on median ratio between conditions

• Variance – required for Negative Binomial distribution
  – Insufficient observations to allow direct measure
  – Custom variance distribution fitted to real data
  – Smooth distribution assumed to allow fitting
Dispersion shrinkage

- Plot observed per gene dispersion
- Calculate average dispersion for genes with similar observation
- Individual dispersions regressed towards the mean. Weighted by
  - Distance from mean
  - Number of observations
- Points more than 2SD above the mean are not regressed
5x5 Replicates

8,022 out of 18,570 genes (43%) identified as DE using DESeq ($p<0.05$)

Needs further filtering

Two options:
1. Decrease the p-value cutoff
2. Filter on magnitude of change

(both are a bit rubbish)
Visualising Differential Expression Results

Filter by p-value (fdr < 10^{-20})

Filter by fold change (abs log2 change > 1.5)
Fold Change Shrinkage

• Aims to make the log2 Fold change a more useful value
• Tries to remove systematic biases
• Two types:
  1. Fold Change Shrinkage – removes bias from both expression level and variance, produces a modified fold change
  2. Intensity difference – removes bias from just expression level, produces a p-value
Fold Change Shrinkage

No Shrinkage

Log2 Fold Change

Expression
Always check your shrunken fold changes

'Normal' Shrinkage

'Apeglm' Shrinkage
Result Validation
(can I believe the hits?)
Validation

2900097C17Rik  RIKEN cDNA 2900097C17 gene  
Hbb-b1  hemoglobin, beta adult major chain  
Rps27a-ps2  ribosomal protein S27A, pseudogene 2  
C230073G13Rik  RIKEN cDNA C230073G13 gene  
mt-Atp8  mitochondrially encoded ATP synthase 8  
mt-Nd4l  mitochondrially encoded NADH dehydrogenase  
AC151712.4  erythroid differentiation regulator 1  
Gm5641  predicted gene 5641
Experimental Design for RNA-Seq
Practical Experiment Design

• What type of library?
• What type of sequencing?
• How many reads?
• How many replicates?
What type of library?

• Directional libraries if possible
  – Easier to spot contamination
  – No mixed signals from antisense transcription
  – May be difficult for low input samples

• mRNA vs total vs depletion etc.
  – Down to experimental questions
  – Remember LINC RNA may not have polyA tail
  – Active transcription vs standing mRNA pool
What type of sequencing

• Depends on your interest
  – Expression quantitation of known genes
    • 50bp single end is fine
  – Expression plus splice junction usage
    • 100bp (or longer if possible) single end
  – Novel transcript discovery or per transcript expression
    • 100bp paired end
How many reads

• Typically aim for 20 million reads for human / mouse sized genome

• More reads:
  – De-novo discovery
  – Low expressed transcripts

• More replicates more useful than more reads
Replicates

- Compared to arrays, RNA-Seq is a very clean technical measure of expression
  - Generally don’t run **technical** replicates
  - Must run **biological** replicates

- For clean systems (eg cell lines) 3x3 or 4x4 is common

- Higher numbers required as the system gets more variable

- Always plan for at least one sample to fail

- Randomise across sample groups
Power Analysis

- Power Analysis is not simple for RNA-Seq data
  - Not a single test – one test per gene
  - Need to apply multiple testing correction
  - Each gene will have different power
    - Power correlates with observation level
    - Variations in variance per gene

- Several tools exist to automate power analysis
  - All require parameters which are difficult to estimate, and have dramatic effects on the outcome
Power analysis for RNA-Seq differential expression studies

Lianbo Yu*, Soledad Fernandez† and Guy Brock†
Tools available

• Scotty - http://scotty.genetics.utah.edu/
• RnaSeqSampleSize (BioConductor)
• Proper (BioConductor)

• All require an estimate of count vs variance
  – Pilot data (if only!)
  – “Similar” studies
Useful links

- HiSat2 [https://ccb.jhu.edu/software/hisat2/](https://ccb.jhu.edu/software/hisat2/)
- Cufflinks [http://cufflinks.cbcb.umd.edu/](http://cufflinks.cbcb.umd.edu/)