RNA-Seq Analysis

Simon Andrews, Laura Biggins
simon.andrews@babraham.ac.uk
@simon_andrews
v2018-10
RNA-Seq Libraries

rRNA depleted mRNA
Fragment
Random prime + RT
2\textsuperscript{nd} strand synthesis (+ U)
A-tailing
Adapter Ligation
(U strand degradation)
Sequencing
Reference based RNA-Seq Analysis

- QC
- Trimming
- Mapping
- Statistical Analysis
- Quantitation
- Mapped QC
Sequence Data Processing
Raw Sequence Quality Control
FastQC
QC: Raw Data

- Sequence call quality
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

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

- Duplication level
Adapters and Trimming
Library Structure

Adapter  Insert  Adapter

Primer  Read 1

Adapter  Insert  Adapter

Primer  Read 1
Trimming Adapters
Trimming Quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Position in read (bp)

Babraham Bioinformatics
Trimming Adapters
Mapping to a reference
Mapping

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 convincingly with novel junction? → Report

Yes → Report

No → Add

Add → Maps convincingly with novel junction? → Report

Yes → Report

No → Discard
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
Mapping Statistics
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

Babraham Bioinformatics
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

Graph showing the data for different mutants (Mut1, Mut2, Mut3, Mut4, WT1, WT2, WT3, WT4) and all probes.
Exploratory Analyses

• Do *lots* of plotting
  – Clustering of samples (PCA / tSNE etc)
  – Do my groups separate as expected?
  – Pairwise comparisons of groups
  – Are expected effects present?
  – Can I see obvious changes?
  – Can I validate other aspects of the same (eg sex)
Differential Expression
Differential Expression

Mapped data

- Normalised expression (+ confidence)
  - Continuous stats
- Raw counts
  - Binomial stats
DE-Seq 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 NB 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
Visualising Differential Expression Results

5x5 Replicates

5,000 out of 22,000 genes (23%) identified as DE using DESeq ($p<0.05$)

Need further filtering!
Magnitude of effect filtering

Intensity difference test

• Fold change biases to low expression

• Need a test which
  – Has a statistical basis
  – Doesn’t bias by expression level
  – Returns sensible numbers of hits
Assumptions

• Noise is related to observation level
  – Similar to DESeq

• Differences between conditions are either
  – A direct response to stimulus
  – Noise, either technical or biological

• Find genes whose differences aren’t explained by general disruption
Method
Results
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
    • 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/](http://scotty.genetics.utah.edu/)
- RnaSeqSampleSize (BioConductor)
- Proper (BioConductor)

- All require an estimate of count vs variance
  - Pilot data (if only!)
  - “Similar” studies
Exercises

• Look at raw QC
• Mapping with HiSat2
  – Small test data
• Quantitation and visualisation
  – Use SeqMonk graphical program
  – Larger replicated data
• Differential expression
  – DESeq2
  – Intensity Difference
• Review in SeqMonk

• Larger Data
  – Multiple conditions
  – Many samples
• Explore
• Quantitate
• Analyse
• Cluster
Useful links

• FastQC  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
• HiSat2  https://ccb.jhu.edu/software/hisat2/
• SeqMonk  http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/
• Cufflinks  http://cufflinks.cbcb.umd.edu/
• DESeq2  https://bioconductor.org/packages/release/bioc/html/DESeq2.html
• Bioconductor  http://www.bioconductor.org/
• DupRadar  http://sourceforge.net/projects/dupradar/