National Aeronautics and Space Administration

RNA Sequencing and Data Processing

Biological and Physical Sciences

Amanda M. Saravia-Butler, Ph.D.
GeneLab Data Processing Lead
NASA Ames Research Center

What Came Before RNA sequencing?

Polymerase Chain Reaction (PCR)

Awesome YouTube video of PCR: https://www.youtube.com/watch?v=iQsu3Kz9NYo

Reverse Transcription Quantitative PCR (RTqPCR)

https://www.youtube.com/watch?v=YhXj5Yy4ksQ

FEMS Microbiology Ecology, Volume: 67, Issue: 1, Pages: 6-20, First published: 09 December 2008, DOI: (10.1111/j.1574-6941.2008.00629.x)

Reverse Transcription Quantitative PCR (RTqPCR)

FEMS Microbiology Ecology, Volume: 67, Issue: 1, Pages: 6-20, First published: 09 December 2008, DOI: (10.1111/j.1574-6941.2008.00629.x)

Limitations of RTqPCR

- Only annotated genes can be evaluated
- Potential for non-specific binding (SYBR green)
- Requires primer optimization (SYBR green)
- 3 conserved and unique regions are required for TaqMan (probe, F and R primers)
- Requires housekeeping genes (genes with consistent expression irrespective of treatment) for normalization
- 'Absolute' quantification of the target gene requires a standard curve, increasing the likelihood for issues if the standard curve is not properly generated
- Limited number of samples and number of genes that can be evaluated simultaneously
- All samples should be run in technical triplicates further limiting the 'n' number
- Others?

Microarray Overview

Modified from: https://microbenotes.com/dna-microarray/

Limitations of Microarrays

- Only annotated genes can be evaluated
- Requires hybridization and hence a risk of non-specific binding
  - Cross-hybridization of sequences with high identity
- Difficult to detect splice isoforms
- Dynamic range limited by the scanner (difficult to detect both low and high expressing genes simultaneously; saturation)
- Low expressed genes are highly variable
- Measures relative abundance – absolute quantitation is difficult
- Unequal labeling efficiency of fluorescent dyes
- Chip to chip variation
- Others?

RNA Sequencing

Sequencing Through the Ages

1977

1990

Now

∞

1st Generation
Sanger Sequencing

Assembly required

↓

2nd Generation
"NextGen"
Pyrosequencing

Assembly required

↓

3rd Generation
Single Molecule
Nanopores

Not much assembly required

Pyrosequencing aka Sequencing by Synthesis

https://www.youtube.com/watch?v=wY8to-_zAEo

https://www.dissertations.se/dissertation/e7acf1968c/

RNA Sequencing Overview

https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

RNA

mRNA makes up only ~2-5% of a total RNA sample

https://www.neb-online.de/wp-content/uploads/2015/04/NEB_selectivedepletionofabundantrnas_d.pdf

Illumina Sequencing Workflow

- Illumina is a very successful biotech company specializing in next generation technology that uses the pyrosequencing method
- ~90% of all sequencing worldwide is performed on an Illumina instrument (including GeneLab)
- The Illumina sequencing workflow has the following 4 steps:

  - Library Construction
  - Cluster Formation
  - Sequencing
  - Data Processing and Analysis

Library Preparation

Figure 1 Ribo-Zero Depleting and Fragmenting RNA

Figure 2 Synthesizing First Strand cDNA

Figure 3 Synthesizing Second Strand cDNA

Figure 4 Adenylating 3' Ends

Illumina TruSeq Stranded Total RNA Reference Guide

Library Preparation

Figure 5 Ligating Adapters

Illumina TruSeq Stranded Total RNA Reference Guide

https://www.idtdna.com/pages/products/next-generation-sequencing/adapters

Library Preparation

Figure 6 Enriching DNA Fragments

Figure 7 LS Final Library

Illumina TruSeq Stranded Total RNA Reference Guide

Library QC: Electrophoresis and Electropherogram

Electrophoresis

Electropherogram

RNA library

https://www.agilent.com/cs/library/applications/5989-1165EN.pdf

Library QC

- Libraries are evaluated using a bioanalyzer or a tape station to create an electropherogram to assess quality

When assessing library quality look for the following:

- NO adapter dimers! – Why?
- Library size is consistent with the number of desired sequencing cycles
  - ➢ If you're sequencing at PE 100, what is a good library size? (hint: library size = insert length + adapter length)
- What is the size range of library 1? Library 2?

  ~175-250bp; ~150-275bp

Modified from UC Davis Bioinformatics RNAseq Training

Library QC

- Assess library quality (bioanalyzer, TapeStation)
  - Adapter dimers
  - Fragment size
- Determine library quantity (Qubit, qPCR)

% AD	% PF	% AD Reads
Control	69.54	0.24
10%	10.87	84.25
5%	21.39	60.44
1%	51.88	6.46

https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseq-stranded-total-rna-workflow/truseq-stranded-total-rna-workflow-reference-1000000040499-00.pdf

Cluster Formation

***Figure 4**. Bridge PCR - a PCR method used to amplify samples for sequencing.*

https://www.goldbio.com/articles/article/Types-of-PCR-used-for-Genetic-Research

Illumina TruSeq Stranded Total RNA Reference Guide

Illumina's High Throughput Sequencing by Synthesis

Figure 8 Read 1 Sequencing

- For paired-end (PE) sequencing, after read 1 is sequenced, forward strand reagents are washed
- Sequencing of the reverse read (read 2) is then initiated

Figure 9 Read 2 Sequencing

Illumina TruSeq Stranded Total RNA Reference Guide

Sequencing Parameters

- **Read length**
  - Longer reads increase gene ID confidence
  - GL standard is 150bp
- **Sequencing depth**
  - Greater depth increases the likelihood of sequencing low-abundant transcripts, detecting novel transcripts, and quantifying isoforms
  - Greater depth is necessary for ribo-depleted samples (vs. poly-A enriched samples)
    - GL standard for mammals prepared with ripo-depletion is 40-60M reads/sample
  - More replicates is usually preferred over greater depth
- **Paired-end (PE) or Single-end (SE)**
  - PE is preferred

Illumina's High Throughput Sequencing by Synthesis

https://www.youtube.com/watch?v=fCd6B5HRaZ8

RNAseq Data Processing

RNAseq Workflow

RNAseq Workflow: FastQC

RNAseq Raw Data

- Raw data generated from the sequencer are stored in fastq files or base calls, which are then converted to fastq files
- Fastq files contain multiple reads and each read is formatted as follows:

```
@K00188:264:HG3WJBBXX:1:1101:6289:1595 1:N:0:TAGCTT
GGACTGCCTTTCAGCCCGTCGCAGAGGGAATGGGAGCCTCTGGAGCGGGTGCAGAGGCTCAGCAG
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJ
```

Line 1 - Begins with @, followed by information about the sequencing run such as the sequencing platform, run number, flow cell ID and cluster location, read number (forward/reverse), and/or the sample index.

Line 2 - Contains the sequence, written as base calls (A, T, C, or G). Note that the sequence length is equal to the number of cycles in the sequencing run.

Line 3 - A separator line, which begins with a plus (+) sign.

Line 4 - Quality scores of each base call that are Phred +33 encoded and use ASCII characters to represent the quality of the bases.

RNAseq Raw Data Quality Scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Raw Data Quality Control with FastQC

- The FastQC program was created to assess the quality of sequence data and provide a QC report for quick evaluation

Basic Statistics

Measure	Value
Filename	Mmus_C57-6J_LVR_RR1_FLT_noERCC_Rep1_M25_R2_raw.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	100600211
Sequences flagged as poor quality	0
Sequence length	150
%GC	52

- Are these data for forward or reverse reads? **Reverse**
- What is the read depth for this sample? **100,600,211**
- What is the sample read length? **150**

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Raw FastQC: Per base sequence quality

Per base sequence quality

- Do you notice anything about the quality as you move towards the end of the read?

  Quality decreases

- These data are for the reverse reads, do you think the quality would be different for the forward reads?

  You may not think so, but the quality of the forward reads is often better than the reverse – We're not entirely sure why…

Raw FastQC: Per sequence quality scores

Per sequence quality scores

- Would you consider these data to have good quality overall?

  Yes

Raw FastQC: Per base sequence content

Per base sequence content

- These values should be fairly consistent across the read length in an unbiased library
- If the library is biased, you will see spikes in the data
- The nature of RNAseq primers tend to cause some wobble at the beginning of the reads

Raw FastQC: Per sequence GC content

Per sequence GC content

- We expect the GC content to be normally distributed (blue line)
- Is the GC content normally distributed across all reads (red line)?

  Kind of?
- What do you think the peaks outside of the normal distribution indicate?

  Contamination

Raw FastQC: Per base N content

Per base N content

- If the sequencer cannot distinguish the base call, an `N` is recorded
- How often was a base unable to be called in these reads?

  Not very often at all, only rarely at the beginning of the read

Raw FastQC: Sequence Length Distribution

Sequence Length Distribution

- The length of all raw reads should be the same
- Are all reads the same length?
  Yes
- What is the read length for these data?
  150bp

Raw FastQC: Adapter Content

- If your sequence of interest is shorter than the number of sequencing cycles, you will sequence into your adapter.
- Does this example contain adapter sequences?
  Yes
- Where are the adapters relative to the read position?
  At the end of the read
- Do you think this will be a problem?
  Probably

Raw FastQC: Sequence Duplication Levels

Sequence Duplication Levels *Note: These data are based on a sampling of the total number of reads and only the first 50bp are considered

- Are duplicates present in this sample? Should we be concerned? **Yes and yes**
- Roughly what percent of reads are duplicates? **~73.5%**

- Used to determine how unique the reads are
- The blue line indicates the percent of duplicated reads at a specific duplication level (red shows what these data would be post deduplication)
- Read duplication can be either biological or technical in origin – How?
- If you have a very diverse library where all reads are unique, 100% of reads would have a duplication level of 1
- It's common, specifically when working with enriched libraries, to have some level of duplication – Why?
- Lower duplication levels (2-9 copies) are likely reads derived from 'interesting' genomic regions (i.e. mRNA)
- Moderate duplication levels (10-100 copies) are likely reads derived from rRNA and/or highly repetitive genomic regions
- High duplication levels (>100 copies) are indicative of a library issue (i.e. adapter dimers; too much PhiX) or contamination

RNAseq Raw Data – Reasons for Duplications

A Review of Sequencing Duplicate Types

- A single cluster that has falsely been called as two by RTA
- Third party tools may report patterned flow cell clustering duplicates as optical duplicates

Not on Patterned Flow Cells

Optical

Clustering

- Duplicates in nearby wells on Patterned Flow Cells
- During cluster generation a library occupies two adjacent wells

Unique to Patterned Flow Cells

- Duplicate molecules that arise from over amplification during sample prep

PCR

Sister

Complement strands of same library form independent clusters

- Treated as duplicates by some informatic pipelines

Present on all systems

illumina