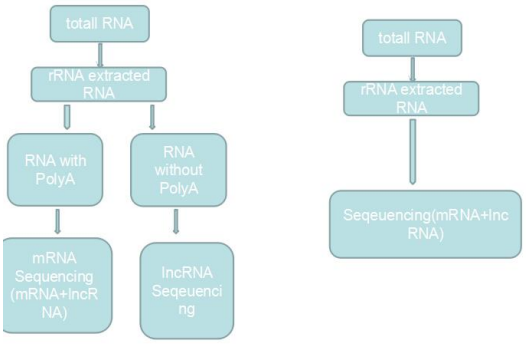
Your Article:

Article Title (3 to 12 words)	Transcriptomics sequencing and lncRNA Sequencing Analysis: Quality Evaluation and Genome Alignment
Article Summary (In short - What is your article about – Just 2 or 3 lines)	The data analysis of transcriptomics sequencing and lncRNA sequencing is the key element of successful interpretation of sequencing. And the quality evaluation and genome alignment are the bases for the data analysis.
Category	*Bioinformatics*

Your full article (between 500 to 5000 words) - - Do check for grammatical errors or spelling mistakes

The data analysis of transcriptomics sequencing and lncRNA sequencing is the key element of successful interpretation of sequencing. And the quality evaluation and genome alignment are the bases for the data analysis.

Basically there are two strategies for lncRNA Sequencing:



Based on the first strategies,we choose an example of the analysis of transcirptomics sequencing of the liver tissues of a patient with advanced liver cancer. The four tissues include adjacent tissue(N), primary site(P), metastasis(M), and portal vein thrombosis metastasis(V). It stands for the four stages of liver cancer. Firstly we extract all the RNA with PolyA and then do mRNA sequencing. We choose the Illumina Hiseq 2000; the size of the inserts in the library is 300bp; we take paired-end sequencing with a read-length of 100 base pairs; based on the D-UTP strand-specific library.

The method of transcriptomics analysis

The identification of RNA is to figure out the quantity of total RNA, lncRNA and message RNA. Different expression quantity of RNA may have different functions, so we need to confirm the specific quantity of all kinds of RNA. The variation is the analysis of changes on structures, expression quantity between RNA. Finally functional annotation is needed for figuring out the function and cellular pathways of RNA.

Quality evaluation

The data of paired-end sequencing will show like this:

N_R1.fastq

@HWI-EAS724_0001:8:32:374:374#0/1

GAGCTGTATATGAATAATAGTTCGTTTTTCATTATCCAAGATGGATCGGTATAAAGTCTGCTAAAATAAAGGTACAACG

+HWI-EAS724_0001:8:32:374:374#0/1

fcfcfggdfggggfggggcgggggggggfgggggcgggfWgggggggggfgcggdgcgcggggfacbbb][bgcgggggd

N_R2.fastq

@HWI-EAS724_0001:8:32:374:374#0/2

TACCGTTAATAGCAGTAATATCATAATAGTAATAGCATCATAACGGTAGTCCCATAAAAGTGTGTCAGTAGTAGTAGTA

+HWI-EAS724_0001:8:32:374:374#0/2

ggggfgggggd_adcggggeggfggeggegf`geececdeggggggfegcfegggegggfgac[aced`bd___c[[Yb

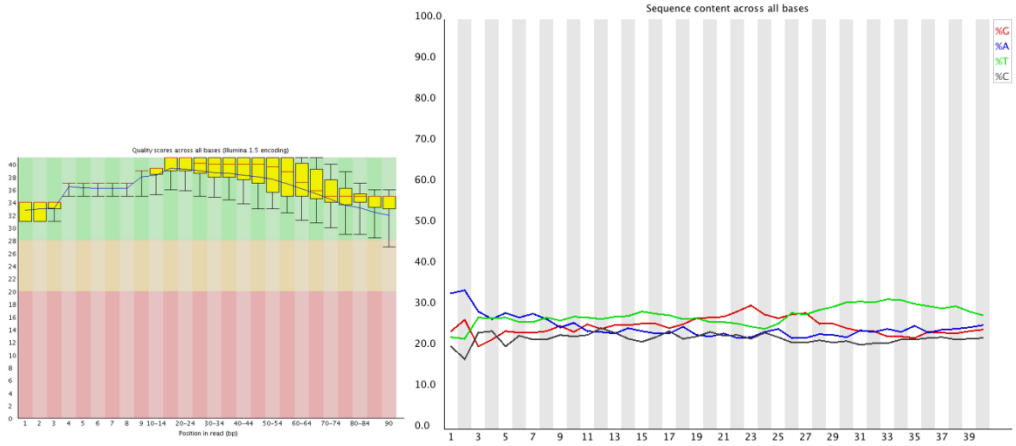
We will have four pairs of data like this. The first line and the third line stand for the name of the read. The second line is the specific information of bases. The last line is the quality of every base expressed by characters. Every single character has a corresponding ASCII code. For Illumina 1.3 and above, subtract 64 from ASCII code and get the quality score of bases. For Sanger, subtract 33. With the quality score we can calculate the error rate of sequencing. The formula is indicated below:

Q_{sanger}= -10 $\log_{10}$P

Qsolexa-prior to v1.3= -10 $\log_{10}\frac{P}{1-P}$

With this formula we can calculate the error rate, P, which can directly indicate the quality of sequencing. However, with the help of some tools we can also get the result with a quick and convenient way.

I highly recommend fastQC. You can easily see the result in a graph like this:



The Q1 is the quality distribution of bases represented by different colors. The red stands for a hundredth and even more error rate. If twenty percent and above of your sequences are in this part, you may get a low-quality sequencing result. The Q2 is the sequence content of every base. Generally four lines with different colors are very close to each other, if part of your bases are not well-distributed, you should consider the reason or cut this part to continue.

Genomic alignment

We choose Tophat here for genomic alignment. This tool is powerful for mapping and discovering splicing junction with RNA-Seq. With a dual alignment strategy, the genome mapped with Tophat and coding genes mapped with Bowtie, we can know the fitness of read map. We choose a dual strategy here because sometimes the high alignment does not mean a good result.

genomic alignment	transcriptomic aligment	analysis
>80%	>50%	Good
>80%	<10%	DNA pollution
<40%	<30%	low quality of sequencing
<10%	<10%	Reads joint, barcode, and uncleaned PolyA

As shown in Q3, the high genomic alignment with a low rate of transcriptomic alignment often means a DNA pollution(since we need an alignment of RNA). If you got a low quality of both, you should consider that every step of your experiment such as PolyA extraction.

Action command

Quality evaluation

fastqc –o QC_outdir_N N_R1.fastq N_R2.fastq

fastqc –o QC_outdir_P P_R1.fastq P_R2.fastq

fastqc –o QC_outdir_M M_R1.fastq M_R2.fastq

fastqc –o QC_outdir_V V_R1.fastq V_R2.fastq

Genomic alignment(hg19 is the index file of genome in Bowtie2)

tophat –o tophat_outdir_N --library-type fr-firststrand --fusion-search hg19 N_R1.fastq N_R2.fastq

tophat –o tophat outdir _P --library-type fr-firststrand --fusion-search hg19 P_R1.fastq P_R2.fastq

tophat –o tophat_outdir_M --library-type fr-firststrand --fusion-search hg19 M_R1.fastq M_R2.fastq

tophat –o tophat_outdir_V --library-type fr-firststrand --fusion-search hg19 V_R1.fastq V_R2.fastq

Transcriptomic alignment

```
bowtie –o  bwt_outdir_N  refgene -1 N_R1.fastq  -2 N_R2.fastq  -S  N.sam

bowtie –o  bwt_outdir _P  refgene  -1  P_R1.fastq  -2 P_R2.fastq -S  P.sam

bowtie –o  bwt_outdir_M  refgene  -1  M_R1.fastq  -2 M_R2.fastq -S  M.sam

bowtie –o  bwt_outdir_V   refgene  -1  V_R1.fastq  -2  V_R2.fastq -V  P.sam
```

References (if any)

1.
2.

About Author:

Your Full Name (published)	Sherry Green
A few lines about you: (published)	Sherry Green, from CD Genomics, www.cd-genomics.com, a biotechnology company which provides sequencing, genotyping, microarray service for global researchers.

Terms - Do not remove or change this section (It should be emailed back to us as is)

- This form is for genuine submissions related to biotechnology topics only.
- You should be the legal owner and author of this article and all its contents.
- If we find that your article is already present online or even containing sections of copied content then we treat as duplicate content - such submissions are quietly rejected.
- If your article is not published within 3-4 days of emailing, then we have not accepted your submission. Our decision is final therefore do not email us enquiring why your article was not published. We will not reply. We reserve all rights on this website.
- Do not violate copyright of others, you will be solely responsible if anyone raises a dispute regarding it.
- Similar to paper based magazines, we do not allow editing of articles once they are published. Therefore please revise and re-revise your article before sending it to us.
- Too short and too long articles are not accepted. Your article must be between 500 and 5000 words.
- We do not charge or pay for any submissions. We do not publish marketing only articles or inappropriate submissions.
- Full submission guidelines are located here: http://www.biotecharticles.com/submitguide.php
- Full Website terms of service are located here: http://www.biotecharticles.com/privacy.php

As I send my article to be published on BiotechArticles.com, I fully agree to all these terms and conditions.
