Comparison of PE and SE for RNA Seq

Introduction

RNA samples are typically sequenced with either 2x125bp paired end (PE) or 1x50bp single end (SE) sequencing at the NGI. Whilst 1x50bp sequencing is typically sufficient for standard differential expression analysis experiments, users must either fill up an entire flow cell or face a potentially long wait for other users. As such, users often opt for the more expensive 2x125bp PE sequencing.

In this tech note, we look into the difference between the two types of sequencing, and how the cost of library preparations and sequencing affect project planning.

Ordering high throughput single end does however require the project to occupy all eight lanes of a flow cell. To properly evaluate this not only the cost is needed but also the expected yield of actually aligned reads against the target genome.

What many users currently do is that they have few biological replicates and thus are not able to fill an entire flow cell by themselves, thus it becomes cheaper to do paired end sequencing even though paired end High throughput is significantly more expensive than single end high throughput sequencing. This technote aims to discern what is actually the method that gives most value for money for different scenarios.

Comparing single- and paired-end data

To study how comparable single-end data is to paired-end, we took 2x125bp data and trimmed it to 1x50bp to simulate a single end dataset. We processed these two datasets in parallel using the new RNA-Seq Best Practice 2.0 pipeline (https://github.com/SciLifeLab/NGI-RNAseq).

Results

We find that the single-end data has a slightly lower alignment rate, as expected (Fig. 1). Despite this, we found the normalised gene counts (FPKM) to be highly similar, with an R^2 score of 0.9792 (Fig. 2), with only lowly expressed genes showing significant variation.

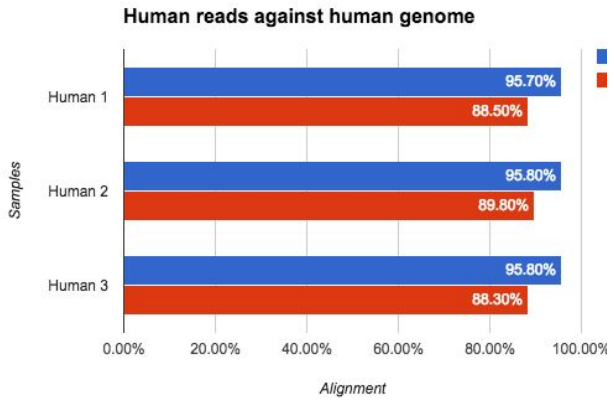

Figure 1: Shows alignment rates for SE (red) and PE (blue) data. Reads aligned against Human (GRCh37)

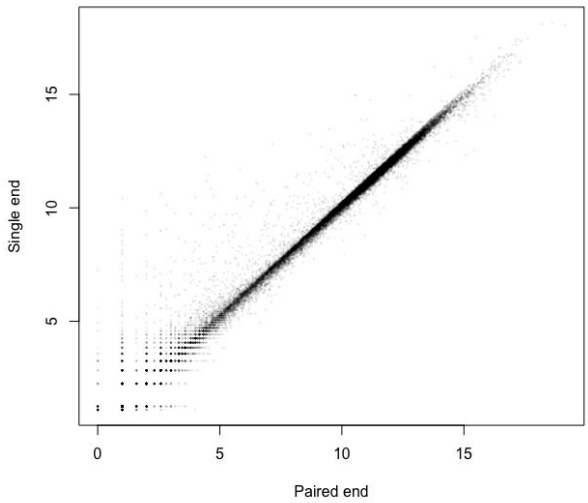

Figure 2: Shows a scatterplot of the FPKM values of one sample with paired end on the x-axis and single end on the y-axis. R^2: 0.9792

Sequencing cost calculations

There are currently three sequencing options for RNA-seq data at the NGI: High throughput 2x125bp, High throughput 1x50bp and Rapid mode 1x50bp. High throughput 1x50bp sequencing is significantly cheaper than the alternatives (Table 1), though requires all 8 lanes in a flowcell to be used.

Sequencing mode	Cost / lane	Reads / lane
High throughput PE	18000 kr	188 M
High throughput SE	7000 kr	188 M
Rapid mode SE	15000 kr	220 M

Table 1: Approximate cost as of June 2016, note that these numbers are the minimum guaranteed reads/ lane. Most of the time it is substantially higher (around 250 of HT and 300M for RM).

To accurately compare the costs of sequencing, we calculated the cost per aligned read to take into consideration the lower alignment rate of SE data. We find that one flowcell (8 lanes) of SE sequencing costs approximately the same as 3 lanes of PE sequencing.

When ordering Illumina Hiseq High Output mode SR 1x50bp it is required that one orders an entire flow cell (8 lanes). The cost for this is 56 000. If one compares that number to how many reads are expected and correct for the actual alignment rate which we found in this investigation then that corresponds to 3 lanes of Illumina Hiseq High Output mode PE 2x125 bp

Caveats

It should be noted that the above discussion is based on projects interested only in measuring

Replicates, replicates, replicates

It is well documented that biological variation is the key source of noise in RNA-seq experiments and that the number of biological replicates is critical for achieving the statistical power needed for accurate differential gene expression[1].

Most sequencing centres recommend six biological replicates per sample.Three replicates are typically required for any analysis and there is limited gain with over 12. For gene expression studies, 10 million reads per replicate is usually sufficient.

We were interested if the increased capacity of a full SE flowcell could be used for additional replicates. To fairly assess this, the cost of RNA-seq library preparations also need to be considered.

As an example, consider a project with 11 samples. If the project is run with three biological replicates on 2x125bp sequencing, the total cost will be ~ 145K SEK. If the same 11 samples are run with six replicates using SE sequencing the total cost is ~203K SEK. The cheaper sequencing goes some way towards negating the additional cost of the library prep kits. As a result, the project receives double the data with significantly more statistical power for a relatively modest increase in price.

Numerous recent publications have shown biological replicates to be crucial for accurate and sensitive detection of differential expression signals in RNA-seq experiments [2, 3]. A common recommendation is that all RNA-seq experiments should have at least 6-12 biological replicates [2].

However each library prep also add its own cost (1100 SEK), and when accounting for that the cost of running more samples increases. Library prep is thus a substantial part of the total price. Consider the following user case. We have 20 samples to analyse and choose to go for six replicates. In this case SE would clearly be the cheaper option, even when adjusting for the observed alignment rates (Fig 3). WIth a full flow cell we would be able to fit 22 SE or 24 PE samples after correcting for the expected alignment rate.

differential gene expression. Some RNA-seq analysis types require PE data, such as detection of fusion transcripts and analysis of repetitive regions. Always consider the downstream analysis requirements before deciding on a sequencing technology.

List of applications where PE is needed

- Fusion transcripts - FusionCatcher
- Cancer - fusion gene detection
- Repetitive transcriptomes

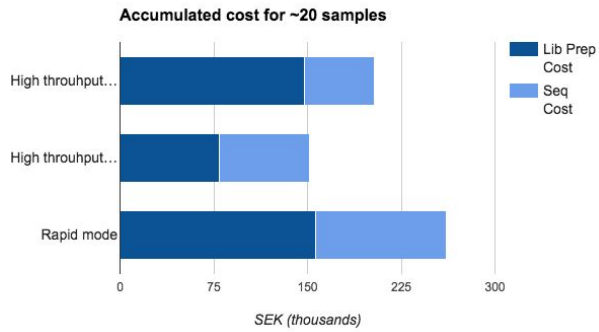

Figure 3 shows a cost calculation with 1 full flow cell for HT PE/SE and 7 lanes RM SE 22,24 and 24 samples respectively

Conclusions

Single end sequencing is cheaper to do than paired end sequencing, even when adjusting for the fact that paired end reads give slightly higher alignment rates. We advocate the use of as many biological replicates as possible and recommend the consideration of SE sequencing as a way mitigating the additional cost of this. More replicates means more preps to sequence, making it easier to fill a complete SE flowcell.

The NGI has greatly improved its ability to use low input amounts for RNA-seq libraries and we are hoping to develop a low-cost library prep kit soon. We hope that these developments encourage users to run as many biological samples with us as possible and that this document gives confidence in using single-end sequencing technology.

References:

[1] Hansen KD, et al. Sequencing technology does not eliminate biological variability. Nat. Biotech

[2] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.

[3] Auer PL, Doerge RW. Statistical Design and Analysis of RNA Sequencing Data.Genetics. 2010;185(2):405-416. doi:10.1534/genetics.110.114983.