Effects of subsampling large RNA-seq data sets before differential gene expression analysis

Tobias Frick, Hugi Ásgeirsson, Sailendra Pradhananga and WongTsz Ching (Amy)

Introduction

Project Aim

When doing RNA-seq for multiple samples and technical replicates, sequencing depth can vary widely between samples. In this study, we try to investigate how subsampling of datasets unevenly distributed coverage affects differential expression analysis of genes between untreated blood samples and samples treated with LPS to simulate inflammation.

Dataset consists of Illunmina HiSeq paired end reads, 100 bp long. Blood samples from three individuals have been subjected to RNA-Seq. Two samples have been collected per individual, and one of each was treated with lipopolysaccharides (LPS) in order to stimulate an inflammatory response.

Methods

Data

Reads from two flowcells were combined, pooling the reads from the same lanes in the two flowcells. Adapters were trimmed off with Cutadapt and reads were quality controlled with FastQC, using the wrapper program TrimGalore. Reads were mapped to the Hg38 reference genome using the STAR mapper.

Reads mapped reads to ENSEMBL GTF featues were counted with HTSeq tools, using intesection-strict mode. We observe that 33% of mapped reads were annotated for features, meaning that two thirds of the reads were not used in downstream analysis.

Methods

Subsampling

We have developed a workflow which outputs a subsampling scheme based on a directory of BAM/SAM files and then generates subsamples with Samtools. Average read count is calculated for each sample group. Samples are ordered by average read count.

Step 1: A cut variable is defined in millions of reads. In Figure 2, 10 million was used.
Step 2: Wherever the difference between two average read counts is larger than the cut, the highest read count in the lowest of the two sample groups is defined as the maximum number of reads for that subsample.
Step 3: A smallest subsample cut can be defined by cutting so that no sample is fully included, if Figure 2 it is 50% of the reads in sample SN10 LPS L1.

Results

Differential expression analysis

Differential expression analysis was carried out comparing treated and untreated samples at FDR < 0.05% and at a fold change higher than 1 for all subsampled datasets. Genes meeting these requirements are thought to be differentially expressed.

Between subsamples 3 and 4 there is no significant change in differentially expressed genes. There is a small increase in DE genes between SS2 and SS3, and a substantial increase between SS1 and SS4.

Overall, SS1 stands out the most, which is to be expecteded as coverage is quite low.

Figure 3: Differential expression

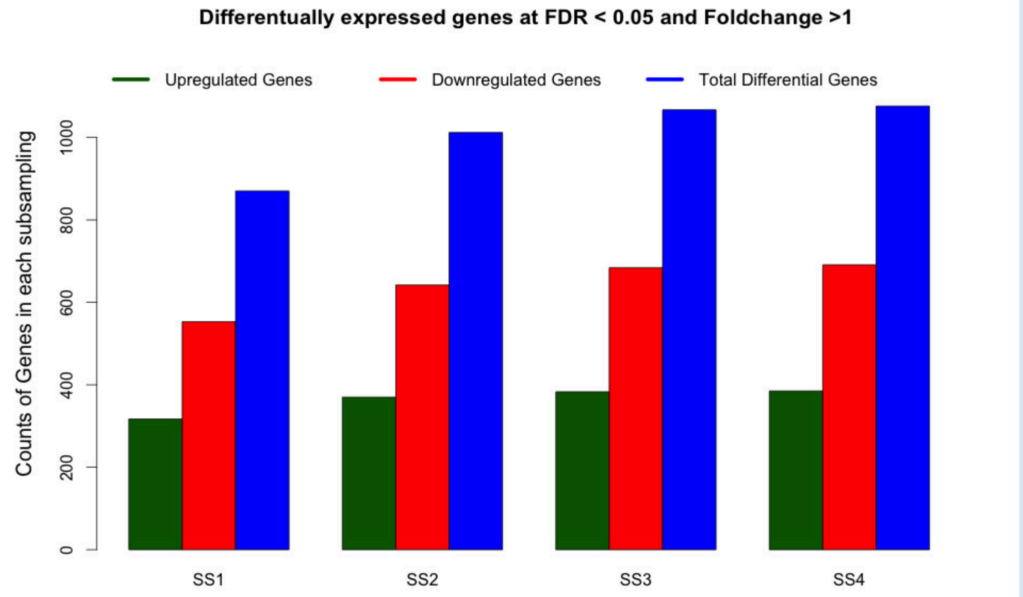


Figure 4: DE genes by subsample

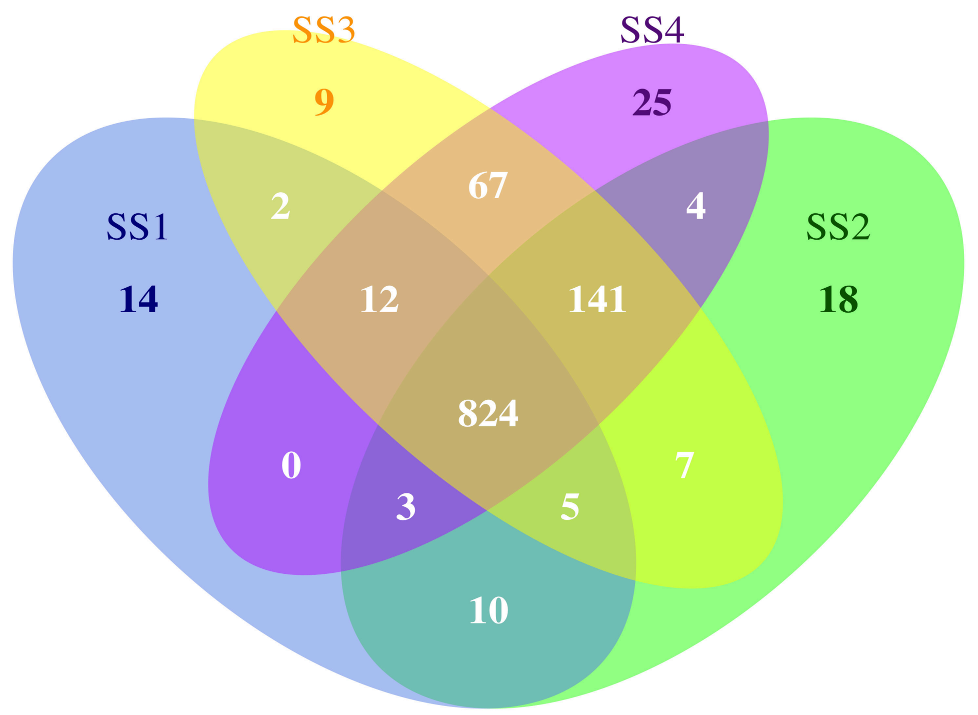


Figure 1: Workflow Overview

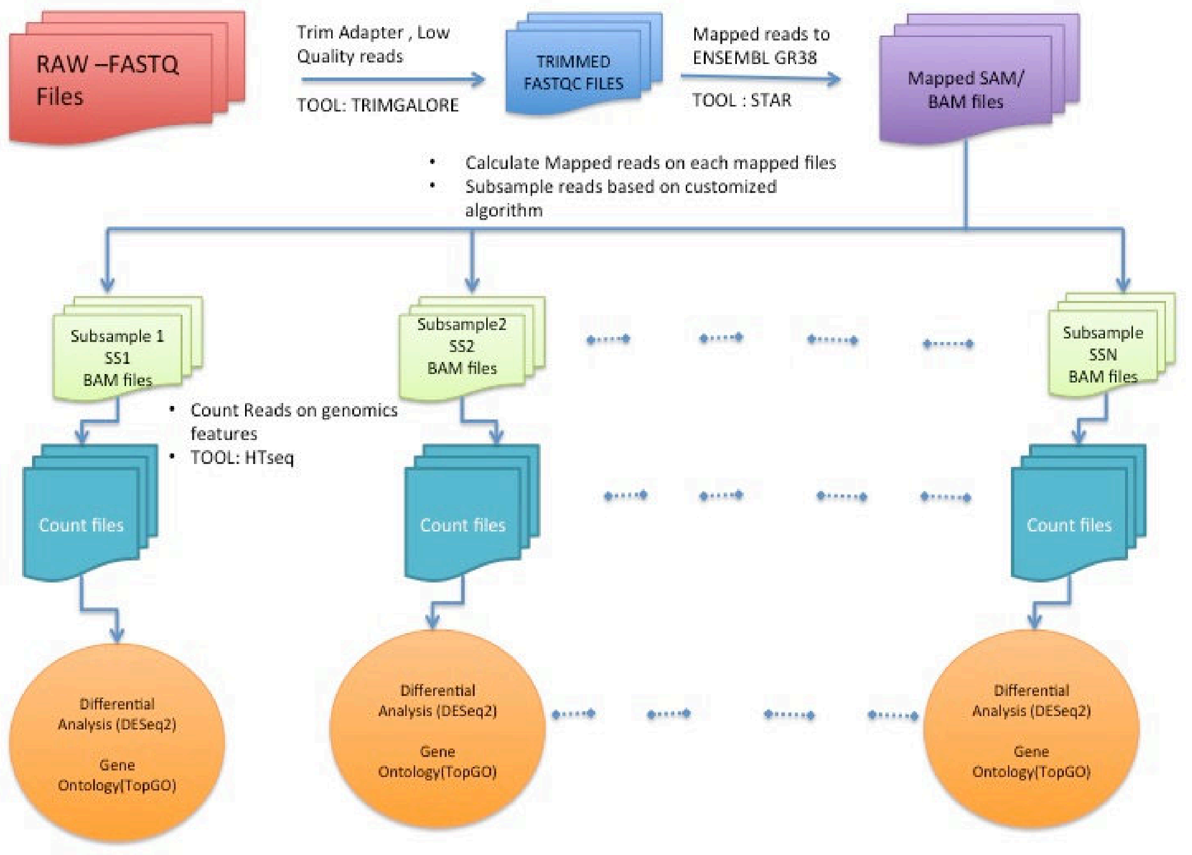
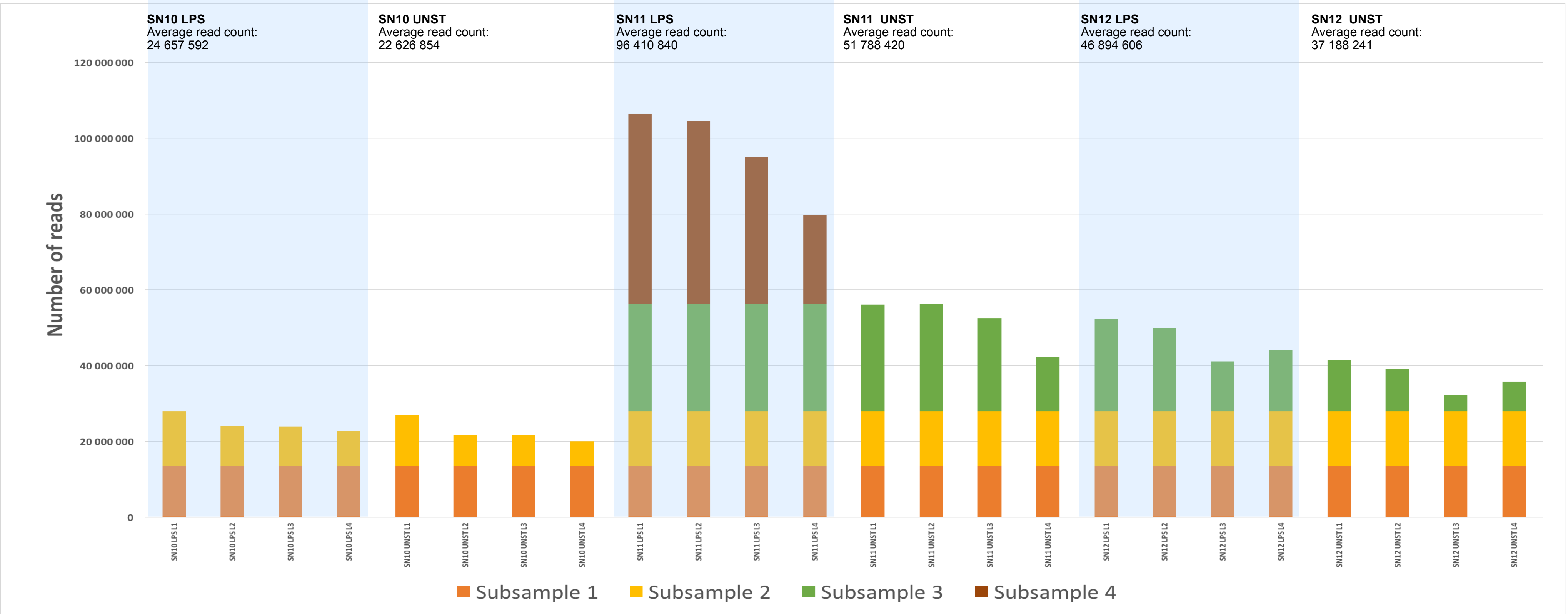

RAW –FASTQ Files
Trim Adapter , Low Quality reads
TOOL: TRIMGALORE
TRIMMED FASTQC FILES
Mapped reads to ENSEMBL GR38
TOOL : STAR
Mapped SAM/ BAM files
Calculate Mapped reads on each mapped files
Subsample reads based on customized algorithm
Subsample 1 SS1 BAM files
Subsample2 SS2 BAM files
Subsample SSN BAM files
Count Reads on genomics features
TOOL: HTseq
Count files
Count files
Count files
Differential Analysis (DESeq2) Gene Ontology(TopGO)
Differential Analysis (DESeq2) Gene Ontology(TopGO)
Differential Analysis (DESeq2) Gene Ontology(TopGO)

Figure 2: Subsampling of mapped reads


SN10 LPS
Average read count:
SN10 UNST
Average read count:
SN11 LPS
Average read count:
SN11 UNST
Average read count:
SN12 LPS
Average read count:
SN12 UNST
Average read count:
Number of reads
SN10 LPS L1
SN10 LPS L2
SN10 LPS L3
SN10 LPS L4
SN10 UNST L1
SN10 UNST L2
SN10 UNST L3
SN10 UNST L4
SN11 LPS L1
SN11 LPS L2
SN11 LPS L3
SN11 LPS L4
SN11 UNST L1
SN11 UNST L2
SN11 UNST L3
SN11 UNST L4
SN12 LPS L1
SN12 LPS L2
SN12 LPS L3
SN12 LPS L4
SN12 UNST L1
SN12 UNST L2
SN12 UNST L3
SN12 UNST L4
Subsample 1
Subsample 2
Subsample 3
Subsample 4