RNA-seq Experiment Report

Overview

The goal of this experiment was to compared the transcriptomics of myofibroblasts and senescent fibroblasts. To achieve this goal, RNA-seq experiment was carried out on human fetal foreskin fibroblasts 2 (HFFF2) treated with 2ng/ml TGF-beta-1 to induce myofibroblast differentiation or 10Gy gamma irradiation to induce senescence. RNA was isolated 7 days upon this treatments.

This report goes through some of the RNA-Seq analysis in R experiment starting with a read count matrix (processed using salmon). This report includes:

- Exploring count data after importing them into R and
- Normalizing RNA-seq counts
- Conducting quality assessment of counts

Data and Metadata

We begin with our counts table and our samples metadata:

	X1_CTR_BC_2 <int>	X2_TGF_BC_4 <int>	X3_IR_BC_5 <int>	X4_CTR_BC_6 <int>	X5_TGF_BC_7 <int>	X6_IR_BC_12 <int>
ENSG00000000003	1579	1547	1342	1704	1395	1264
ENSG00000000005	0	0	0	0	0	0
ENSG00000000419	1774	1775	1866	1809	1921	1776
ENSG00000000457	698	617	601	733	662	537
ENSG00000000460	246	309	116	224	255	108
ENSG00000000938	10	5	11	6	1	2
ENSG00000000971	6370	5238	5008	6486	5102	4832
ENSG00000001036	4009	3335	3855	4224	2864	3688
ENSG00000001084	1037	940	928	970	968	857
ENSG00000001167	1498	1622	1042	1406	1535	911

1-10 of 10 rows | 1-7 of 9 columns

Run <chr>	condition <chr>	Name <chr>	Replicate <dbl>	Treated <chr>
1_CTR_BC_2	CTR	CTR_1	1	N
2_TGF_BC_4	TGF	TGF_1	1	Y
3_IR_BC_5	IR	IR_1	1	Y
4_CTR_BC_6	CTR	CTR_2	2	N
5_TGF_BC_7	TGF	TGF_2	2	Y
6_IR_BC_12	IR	IR_2	2	Y
7_CTR_BC_13	ctr	CTR_3	3	N
8_TGF_BC_14	tgf	TGF_3	3	Y
9_IR_BC_15	ir	IR_3	3	Y

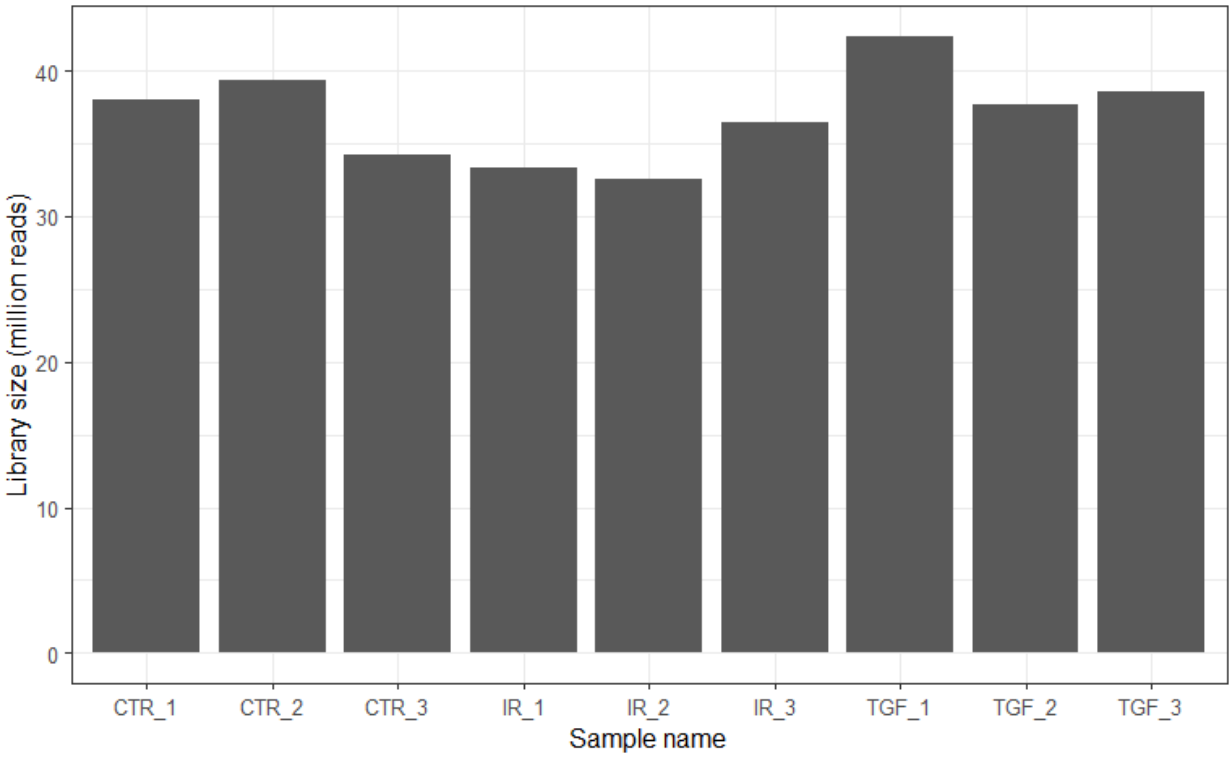
9 rows

Quality control of the imported counts

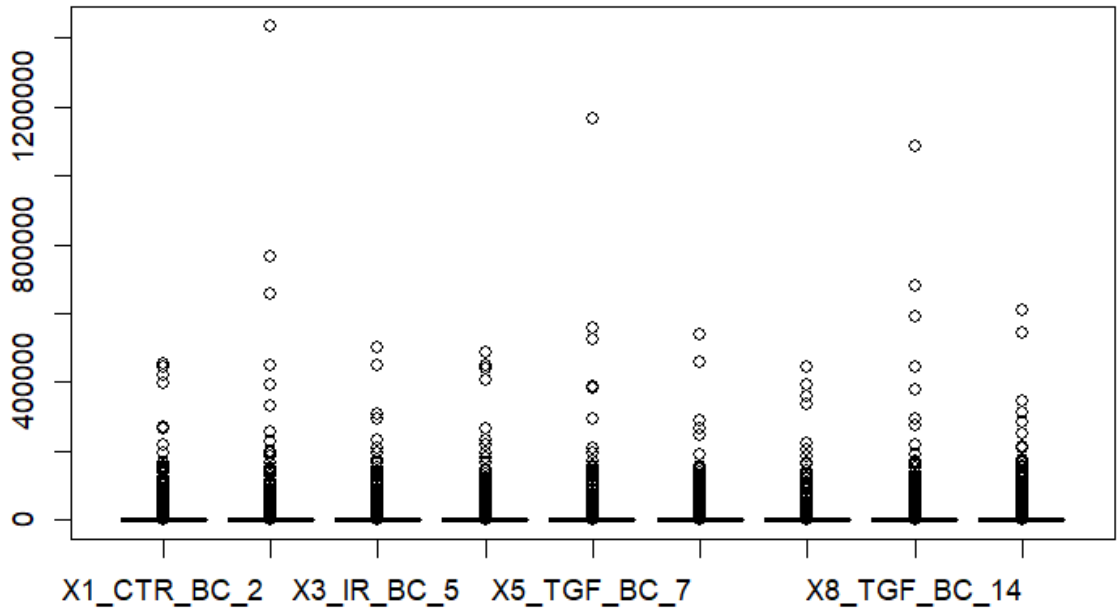
We can look at a few different plots to check that the data is good quality, and that the samples are behaving as we would expect. First, we can check how many reads we have for each sample.

Library Sizes

The library sizes generally look good for each of our samples, but are not completely uniform:



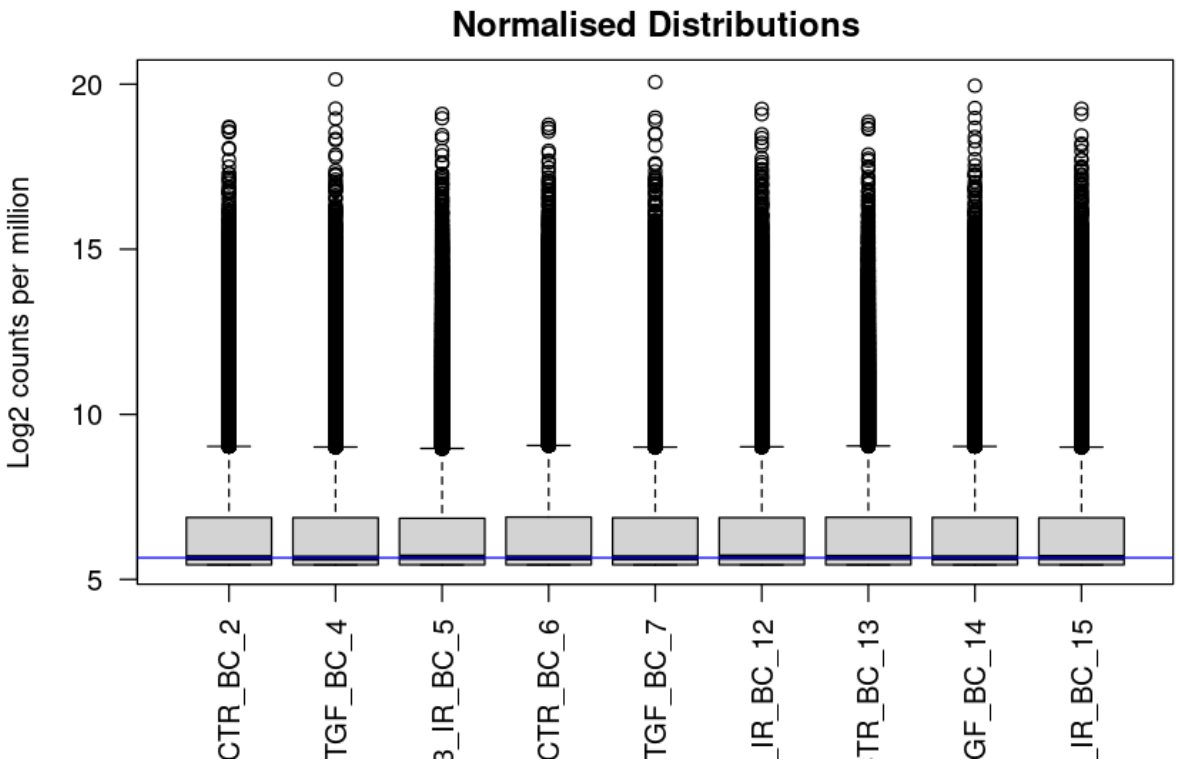
Thus, there are some outliers when we look at the distribution of counts:



Normalization

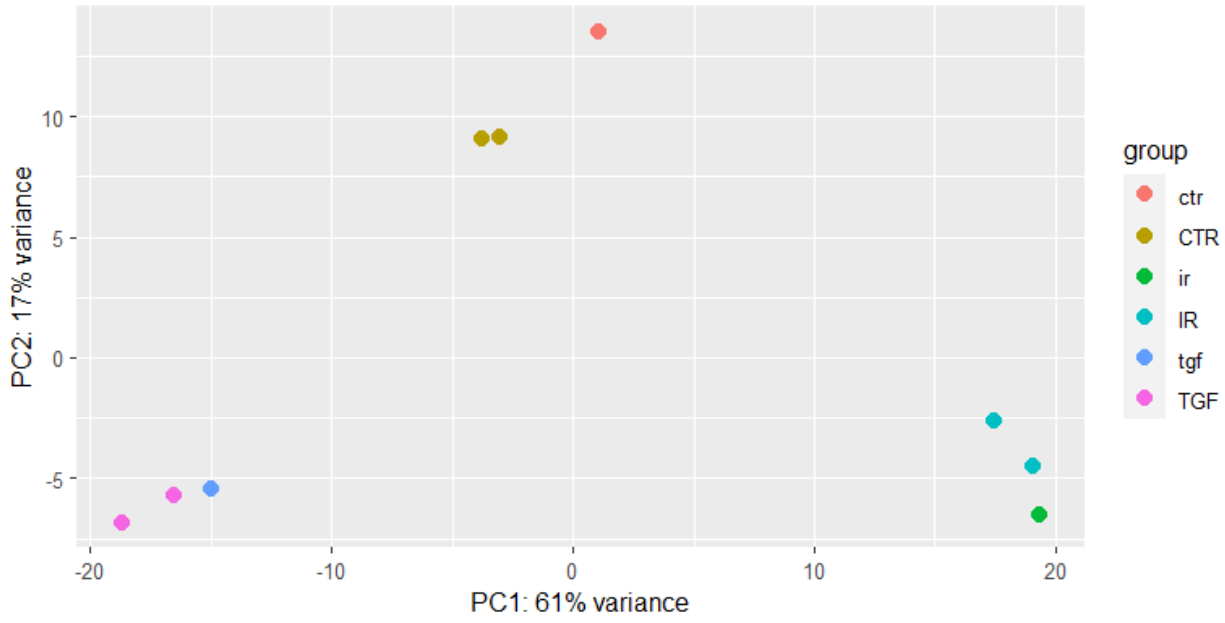
In order to test for differential expression, we operate on raw counts. However for other downstream analyses – e.g. for visualization or clustering – it is useful to work with transformed versions of the count data.

We used the variance stabilizing transformation (DESeq2::vst()) function to compensate for different library sizes. This transformation performs an approximately $\log_2$ transformation of the count data, as well as aims to make it homoscedastic; meaning that the variance does not depend on the mean. This is a common transformation as many common statistical methods for exploratory analysis of multidimensional data, especially methods for clustering and ordination (e.g., principal-component analysis and the like), work best for (at least approximately) homoscedastic. DESeq2's transformation calculates variance by mean in a condition-blind manner.

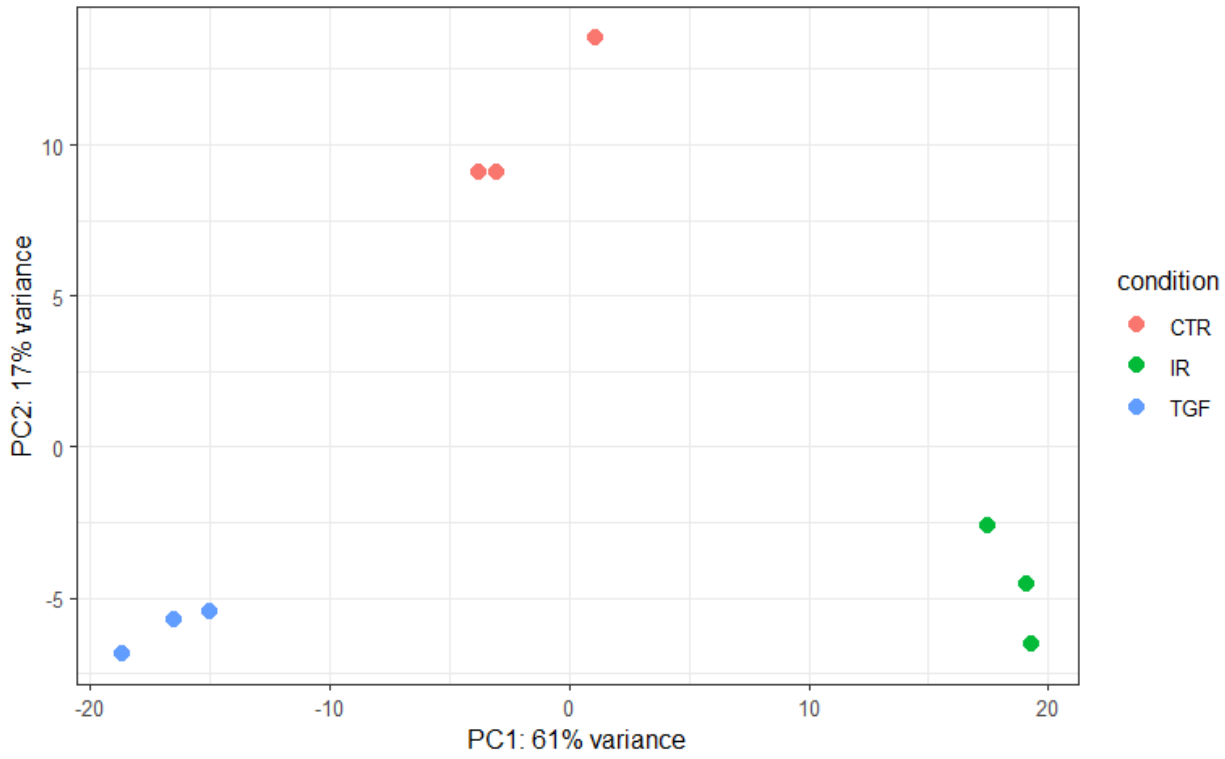

Normalised Distributions
Log2 counts per million
20
15
10
5
CTR_BC_2
TGF_BC_4
3_IR_BC_5
CTR_BC_6
TGF_BC_7
_IR_BC_12
TR_BC_13
GF_BC_14
_IR_BC_15

PCA

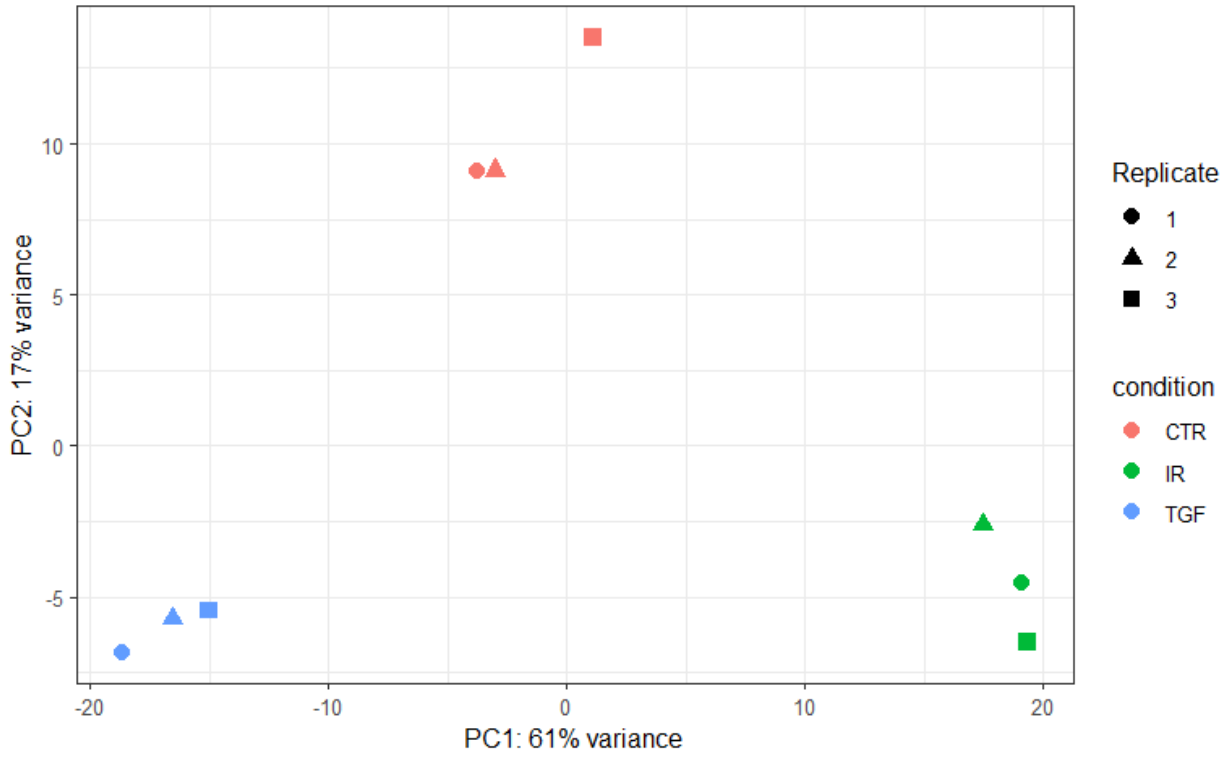
PCA of normalized samples colored by group (using plotPCA()):



Our groups have different coloring due to their labels having different capitalization. After correction, we can see an updated plot:



There appears to be little to no batch effect:



We also used prcomp() to try only using top genes for PCA:

Table of PCA results using prcomp() and the top 500 ranked genes only:

Importance of components:								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard d	15.469	8.1208	5.2935	4.0091	3.7323	3.2092	2.99442	2.76288
Proportion	0.6133	0.169	0.07182	0.0412	0.0357	0.0264	0.02298	0.01957
Cumulative	0.6133	0.7823	0.85416	0.8953	0.9311	0.9575	0.98043	1

Reference:

Based on Introduction or RNA-Seq analysis in R workshop offered by Sheffield Bioinformatics Core.

Original study: Mellone M, Hanley CJ, Thirdborough S, Mellows T, Garcia E, Woo J, Tod J, Frampton S, Jenei V, Moutasim KA, Kabir TD, Brennan PA, Venturi G, Ford K, Herranz N, Lim KP, Clarke J, Lambert DW, Prime SS, Underwood TJ, Vijayanand P, Eliceiri KW, Woelk C, King EV, Gil J, Ottensmeier CH, Thomas GJ. Induction of fibroblast senescence generates a non-fibrogenic myofibroblast phenotype that differentially impacts on cancer prognosis. Aging (Albany NY). 2016 Dec 15;9(1):114-132. doi: 10.18632/aging.101127. PMID: 27992856; PMCID: PMC5310659.