# Fibroblast data normalization

2024-09-04

## Preparation / QC

Important note: the following (commented) code is for the local data exported from kallisto as I ran the entire analysis from scratch. To use the data from the GEO datasets, use code below (GEO data preparation).

```r
rm(list = ls())
data_dir <- here::here("Data", "Fibroblasts", "ProcessedData")
library(limma)
library(Glimma)
library(edgeR)
library(tidyverse)
library(tximport)
library(DESeq2)
library(RColorBrewer)
# genewizDir <- "/Users/annamonzel/Library/CloudStorage/GoogleDrive-asm2285@cumc.columbia.edu/.shortcut
#
# # Get metadata and samples included in analysis
# samples <- read_csv(here::here("Data", "Fibroblasts", "OriginalData", "RNAseq_meta.csv")) %>%
#   mutate(tmp = "Sample_") %>%
#   unite(RNAseq_sampleID, tmp, RNAseq_sampleID, sep = "") %>%
#   as.data.frame() %>%
#   dplyr::select(RNAseq_sampleID, Cell_line_group, Clinical_condition, Treatments, Percent_oxygen, Uni
#   separate(RNAseq_sampleID, into = c("Sample", "ID"), remove = F) %>%
#   mutate(ID = as.numeric(ID))
#
# samples_all <- read_csv(here::here("Data", "Fibroblasts", "OriginalData", "RNAseq_meta.csv"))
# length(unique(samples_all$Cell_line_inhouse))
# max(na.omit(samples_all$Passage))
```

### Tximport

```r
# samples_to_read <- sort(as.numeric(samples$ID))
# files <- file.path(genewizDir, 'kallisto_output_newv3', samples_to_read, "abundance.tsv")
# names(files) <- paste0("Sample_", samples_to_read)
# x = read.csv(here::here(genewizDir, 'kallisto_output_newv3/1/abundance.tsv'),sep='\t',header=T)
#
# ## summarize only across coding RNAs
# mRNA <- x$target_id[which(grepl("mRNA", x$target_id)==TRUE)]
# mtRNA <- x$target_id[which(grepl("CDS", x$target_id)==TRUE)]
# mRNA_genes = sapply(lapply(strsplit(as.character(mRNA), '\\;'), 'rev'), '[', 2)
# mtRNA_genes = sapply(lapply(strsplit(as.character(mtRNA), '\\;'), 'rev'), '[', 2)
# tx2gene = data.frame(target_id = as.character(c(mRNA, mtRNA)), gene = c(mRNA_genes, mtRNA_genes))
#
# ## Important: order samples in counts dataset in the same order they are in the metadata (samples)
```

```
# txi <- tximport(files, type = "kallisto",tx2gene=tx2gene, countsFromAbundance = "no")
# gene_counts <- txi$counts[,samples$RNAseq_sampleID]
# gene_abundances <- txi$abundance[,samples$RNAseq_sampleID]
# gene_lengths <- txi$length[,samples$RNAseq_sampleID]
#
# txi.tx <- tximport(files, type = "kallisto",tx2gene=tx2gene, txOut = TRUE) #countsFromAbundance = "le:
# transcript_counts <- txi.tx$counts[,samples$RNAseq_sampleID]
# transcript_abundances <- txi.tx$abundance[,samples$RNAseq_sampleID]
# transcript_lengths <- txi.tx$length[,samples$RNAseq_sampleID]
#
# dim(gene_counts)
# dim(transcript_counts)
```

## QC

```
# for_viz <- as.data.frame(transcript_counts) %>%
#   rownames_to_column("Genes") %>%
#   separate(Genes, into = c("REFSEQ1", "REFSEQ2", "SYMBOL", "TYPE"), sep = ";")  %>%
#   pivot_longer(cols = -c("REFSEQ1", "REFSEQ2", "SYMBOL", "TYPE"), names_to = "RNAseq_sampleID", value:
#   full_join(samples, by = "RNAseq_sampleID") %>%
#   group_by(TYPE, RNAseq_sampleID) %>%
#   mutate(mean = mean(exprs)) %>%
#   ungroup() %>%
#   dplyr::select(TYPE, RNAseq_sampleID, mean) %>%
#   unique()
#
# for_viz %>%
#   ggplot(aes(x = reorder(RNAseq_sampleID, mean), y = mean)) +
#   geom_col(mapping = NULL) +
#   facet_wrap(~TYPE, scales = "free") +
#   theme(legend.position = "none")
# rm(samples_to_read, files,  for_viz,  x, tx2gene)
```

### Visualize library size

```
# x <- DGEList(counts = gene_counts,
#              samples = samples,
#              group = samples$Cell_line_group
#              )
# plotly::ggplotly(x$samples %>%
#   rownames_to_column("sample") %>%
#   unite(Condition, Clinical_condition, Treatments) %>%
#   ggplot(aes(x = reorder(sample, lib.size), y = lib.size, fill = Condition, label = Cell_line_group))
#   geom_bar(stat = "identity") +
#   labs(y = "Library size (total number of mapped and quantified reads)",
#        x = "Sample", fill = "Group") +
#   coord_flip())
```

# nTPM normalization (for mitotyping)

https://www.nature.com/articles/s41467-023-41132-w

```
# rm(list = setdiff(ls(), c("samples", "genewizDir", "data_dir")))
# samples_to_read <- sort(as.numeric(samples$ID))
# files <- file.path(genewizDir, 'kallisto_output_newv3', samples_to_read, "abundance.tsv")
# names(files) <- paste0("Sample_", samples_to_read)
# x = read.csv(here::here(genewizDir, 'kallisto_output_newv3/1/abundance.tsv'),sep='\t',header=T)
# ## summarize only across coding RNAs
# mRNA <- x$target_id[which(grepl("mRNA", x$target_id)==TRUE)]
# mtRNA <- x$target_id[which(grepl("CDS", x$target_id)==TRUE)]
# mRNA_genes = sapply(lapply(strsplit(as.character(mRNA), '\\;'), 'rev'), '[', 2)
# mtRNA_genes = sapply(lapply(strsplit(as.character(mtRNA), '\\;'), 'rev'), '[', 2)
# tx2gene = data.frame(target_id = as.character(c(mRNA, mtRNA)), gene = c(mRNA_genes, mtRNA_genes))
#
# txi <- tximport(files, type = "kallisto",tx2gene=tx2gene, countsFromAbundance = "lengthScaledTPM")
# gene_counts <- txi$counts[,samples$RNAseq_sampleID]
# gene_abundances <- txi$abundance[,samples$RNAseq_sampleID]
# gene_lengths <- txi$length[,samples$RNAseq_sampleID]
#
#
# sapply(as.data.frame(gene_counts), sum)
# pTPM <- cpm(gene_counts)
# sapply(as.data.frame(pTPM), sum)
# #saveRDS(pTPM,  here::here(data_save_dir,"pTPM_Lifespan.rds"))
#
# ref.samples <- samples %>%
#   filter(Clinical_condition %in% "Normal") %>%
#   filter(Treatments %in% "Control") %>%
#   filter(Percent_oxygen %in% "21") %>%
#   pull(RNAseq_sampleID)
# ref.column <- data.frame(Median = rowMedians(pTPM[,ref.samples]))
# rownames(ref.column) = rownames(pTPM)
# pTPM <- bind_cols(ref.column, pTPM)
#
# nTPM <- NOISeq::tmm(pTPM, refColumn = "Median")
# sapply(as.data.frame(nTPM), sum)
# nTPM <- nTPM[,-1]
#
# saveRDS(nTPM,  here::here(data_dir,"nTPM_Lifespan.rds"))
# rm(list = setdiff(ls(), c("nTPM","keep.exprs", "samples", "genewizDir", "data_dir")))
```

# Mitotyping preparation

## Mitogenes found in data

Note: mtDNA genes in Mitocarta are "MT-XXX" and only "XXX" in kallisto. (changed here to match)

```
# genes_found <- rownames(nTPM)
#
# mitogenes <- readxl::read_xls(here::here("Data", "MitoCarta","OriginalData", "HumanMitoCarta3_0.xls")
#   pull(Symbol)
# mitogenes_mt <- mitogenes[grepl("MT-", mitogenes)]
# mitogenes_mt <- str_remove(mitogenes_mt, "MT-")
# mitogenes <- mitogenes[!grepl("MT-", mitogenes)]
# mitogenes <- c(mitogenes, mitogenes_mt)
```

```
#
# genes_found_mt <- unique(sort(genes_found[which(genes_found %in% mitogenes)]))
# length(genes_found_mt)
```

The following genes were not found

```
# mitogenes[which(!mitogenes%in% genes_found_mt)]
```

Note: MT-CO1, MT-CO2 and MT-CO3 are differently annotated in kallisto. Here are all mtDNA genes as annotated in Kallisto:

```
# samples_to_read <- sort(as.numeric(samples$ID))
# files <- file.path(genewizDir, 'kallisto_output_newv3', samples_to_read, "abundance.tsv")
# names(files) <- paste0("Sample_", samples_to_read)
# x = read.csv(here::here(genewizDir, 'kallisto_output_newv3/1/abundance.tsv'),sep='\t',header=T)
#
# ## summarize only across coding RNAs
# mRNA <- x$target_id[which(grepl("mRNA", x$target_id)==TRUE)]
# mtRNA <- x$target_id[which(grepl("CDS", x$target_id)==TRUE)]
# mRNA_genes = sapply(lapply(strsplit(as.character(mRNA), '\\;'), 'rev'), '[', 2)
# mtRNA_genes = sapply(lapply(strsplit(as.character(mtRNA), '\\;'), 'rev'), '[', 2)
# tx2gene = data.frame(target_id = as.character(c(mRNA, mtRNA)), gene = c(mRNA_genes, mtRNA_genes))
#
# ## Important: order samples in counts dataset in the same order they are in the metadata (samples)
# txi <- tximport(files, type = "kallisto",tx2gene=tx2gene, countsFromAbundance = "no")
# gene_counts <- txi$counts[,samples$RNAseq_sampleID]
# gene_abundances <- txi$abundance[,samples$RNAseq_sampleID]
# gene_lengths <- txi$length[,samples$RNAseq_sampleID]
#
# txi.tx <- tximport(files, type = "kallisto",tx2gene=tx2gene, txOut = TRUE)
# transcript_counts <- txi.tx$counts[,samples$RNAseq_sampleID]
# transcript_abundances <- txi.tx$abundance[,samples$RNAseq_sampleID]
# transcript_lengths <- txi.tx$length[,samples$RNAseq_sampleID]
#
#
# mtDNAgenes_in_data <- as.data.frame(transcript_counts) %>%
#   rownames_to_column("Genes") %>%
#   separate(Genes, into = c("REFSEQ1", "REFSEQ2", "SYMBOL", "TYPE"), sep = ";") %>%
#   filter(TYPE %in% "CDS") %>% unique()
# mtDNAgenes_in_data$SYMBOL
```

This has to be changed in downstream analyses, but kept for now. I just search by mitogenes + CDS annotation in kallisto. The other two genes not found are MYG1, also known as C12orf10, and RP11_469A15.2, which is a Long intergenic non-coding RNA (lincRNA). C12orf10 could be found, but RP11_469A15.2 was excluded (only mtDNA-RNA and mRNA kept)

```
# genes_found <- rownames(nTPM)
# genes_found_mt <- unique(sort(genes_found[which(genes_found %in% c(mitogenes, mtDNAgenes_in_data$SYMB
# length(genes_found_mt)
# rm(genes_found, genes_found_mt, mtDNAgenes_in_data, mitogenes, mitogenes_mt)
```

# GEO data preparation

Run this if you downloaded the data from GEO. Note that the data on GEO was prepared by Sturm et al. 2022, with different R software and package versions. It might not 100% match the data prepared using

the code above (kallisto, tximport). This does not affect the figures overall, but some statistics might differ slightly (e.g. the PCA variances). The raw data that was used above can be shared upon request.

```r
gene_counts <- read_csv(here::here("Data/Fibroblasts/OriginalData", "GSE179848_txi_gene_count_lengthScal
  pivot_longer(cols = -`...1`) %>%
  separate(name, into = c("Sample", "Type"), sep = "_") %>%
  select(-Type) %>%
  mutate(Sample2 = "Sample") %>%
  unite(Sample, Sample2, Sample, sep = "_") %>%
  pivot_wider(names_from = Sample, values_from = value) %>%
  column_to_rownames("...1")
```

```
## New names:
## Rows: 28633 Columns: 352
## -- Column specification
## -------------------------------------------------------------------------------------------- Delimiter:
## (1): ...1 dbl (351): 1_lenscaleTPM, 10_lenscaleTPM, 100_lenscaleTPM, 101_lenscaleTPM, 102_lenscaleTPM
## 103_lenscaleTPM...
## i Use `spec()` to retrieve the full column specification for this data. i Specify the column types o
## `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
samples <- read_csv(here::here("Data", "Fibroblasts", "OriginalData", "RNAseq_meta.csv")) %>%
  mutate(tmp = "Sample_") %>%
  unite(RNAseq_sampleID, tmp, RNAseq_sampleID, sep = "") %>%
  as.data.frame() %>%
  dplyr::select(RNAseq_sampleID, Cell_line_group, Clinical_condition, Treatments, Percent_oxygen, Unique
  separate(RNAseq_sampleID, into = c("Sample", "ID"), remove = F) %>%
  mutate(ID = as.numeric(ID))
```

```
## Rows: 345 Columns: 46
## -- Column specification ---------------------------------------------------------------------------
## Delimiter: ","
## chr  (12): Cell_line_group, Unique_variable_name, Cell_line, Cell_line_inhouse, Cell_type, Sex, Clin
## dbl  (33): Sample, Donor_age, Study_part, Replicate_line, Percent_oxygen, pre_designated_time_point_
## time  (1): Time_of_Passage
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sapply(as.data.frame(gene_counts), sum)
```

```
##    Sample_1  Sample_10 Sample_100 Sample_101 Sample_102 Sample_103 Sample_104 Sample_105 Sample_106 S
##    24121173   28068561   18073068   21862855   16161203   16240550   16164500   19652441   32521027
## Sample_108 Sample_109  Sample_11 Sample_110 Sample_111 Sample_112 Sample_113 Sample_114 Sample_115 S
##    28052932   26585455   25770289   17017719   15988597   15335512   15315648   32400378   29525720
## Sample_117 Sample_118 Sample_119  Sample_12 Sample_120 Sample_121 Sample_122 Sample_123 Sample_124 S
##    17148961   16725654   14893595   29913771   14716611   32018294   14696524   16343487   15696374
## Sample_126 Sample_127 Sample_128 Sample_129  Sample_13 Sample_130 Sample_131 Sample_132 Sample_133 S
##    24360293   15692107   16759403   16321255   27046126   26393324   32720498   16775203   16034219
## Sample_135 Sample_136 Sample_137 Sample_138 Sample_139  Sample_14 Sample_140 Sample_141 Sample_142 S
##    15769954   15189129   34322206   16464534   18286467   23395123   16004589   32263526   16966592
## Sample_144 Sample_145 Sample_146 Sample_147 Sample_148 Sample_149  Sample_15 Sample_150 Sample_151 S
##    16476002   15249094   16371617   31874629   15436643   31414557   22159800   17095270   15080323
## Sample_153 Sample_154 Sample_155 Sample_156 Sample_157 Sample_158 Sample_159  Sample_16 Sample_160 S
##    30673378   15455748   20348587   32376557   28415145   27872854   20984736   26778552   15379338
```

```
## Sample_162 Sample_163 Sample_164 Sample_165 Sample_166 Sample_167 Sample_168 Sample_169  Sample_17 Sa
##    43030609   17022274   39533116   17397731   30162065   25987017   17762789   32530009   21747913
## Sample_171 Sample_172 Sample_173 Sample_174 Sample_175 Sample_176 Sample_177 Sample_178 Sample_179   S
##    34079826   16643159   31602177   16411681   36023102   28067664   37811275   27886856   57407880
## Sample_180 Sample_181 Sample_182 Sample_183 Sample_184 Sample_185 Sample_186 Sample_187 Sample_188 Sa
##    27736172   27442933   27748290   23988606   27504197   20234189   26439951   25808298   20443532
##  Sample_19 Sample_190 Sample_191 Sample_192 Sample_193 Sample_194 Sample_195 Sample_196 Sample_197 Sa
##    23315592   18651843   20115710   20035136   19060742   18799928   20420732   21255619   20014220
## Sample_199   Sample_2  Sample_20 Sample_200 Sample_201 Sample_202 Sample_203 Sample_204 Sample_205 Sa
##    18440425   28127609   25414941   20159125   17452643   19100341   20129452   19854538   18151660
## Sample_207 Sample_208 Sample_209  Sample_21 Sample_210 Sample_211 Sample_212 Sample_213 Sample_214 Sa
##    16682096   20822215   18524637   25313017   20581205   18611663   20733889   17923352   17219943
## Sample_216 Sample_217 Sample_218 Sample_219  Sample_22 Sample_220 Sample_221 Sample_222 Sample_223 Sa
##    20427140   15210543   15731326   16702374   20890266   18127190   15638143   15530759   16665951
## Sample_225 Sample_226 Sample_227 Sample_228 Sample_229  Sample_23 Sample_230 Sample_231 Sample_232 Sa
##    14873291   15767471   18123730   17992631   15921825   20339367   15754577   16524941   17895006
## Sample_234 Sample_235 Sample_236 Sample_237 Sample_238 Sample_239  Sample_24 Sample_240 Sample_241 Sa
##    18338232   17576536   15503979   17452081   15825035   14384258   22948263   16632678   17219495
## Sample_243 Sample_244 Sample_245 Sample_246 Sample_247 Sample_248 Sample_249  Sample_25 Sample_250 Sa
##    17597840   17333325   18158663   16742446   14623763   16073443   15860384   24769372   16014772
## Sample_252 Sample_253 Sample_254 Sample_255 Sample_256 Sample_257 Sample_258 Sample_259  Sample_26 Sa
##    25851549   21856005   21977244   20085904   19691262   25089446   24522304   27441698   25072003
## Sample_261 Sample_262 Sample_263 Sample_264 Sample_268  Sample_27 Sample_272 Sample_273 Sample_275 Sa
##    26285465   24913369   26211792   24992848    4871840   24134928   28282487   27148858   29160364
## Sample_279  Sample_28 Sample_280 Sample_281 Sample_282 Sample_283 Sample_284 Sample_285 Sample_286 Sa
##    27881727   25872216   23042341   27777808   24764343   30935252   30034579   29826367   23934403
## Sample_288 Sample_289  Sample_29 Sample_290 Sample_291 Sample_292 Sample_293 Sample_294 Sample_295 Sa
##    22665784   23390959   23532515   25872662   23532542   25997477   24964728   23800680   24349863
## Sample_297 Sample_298 Sample_299   Sample_3  Sample_30 Sample_300 Sample_301 Sample_302 Sample_303 Sa
##    23519863   23931382   26800586   26481876   23624209   30092547   23293797   24948859   25560715
## Sample_305 Sample_306 Sample_307 Sample_308 Sample_309  Sample_31 Sample_310 Sample_311 Sample_312 Sa
##    23862478   25300091   25100668   24328037   23556377   22136186   22554461   25794488   22084665
## Sample_314 Sample_315 Sample_316 Sample_317 Sample_318 Sample_319  Sample_32 Sample_320 Sample_321 Sa
##    28485758   26987403   28942634   26898974   26294935   28388535   25335513   26489800   22273644
## Sample_323 Sample_324 Sample_325 Sample_326 Sample_327 Sample_328 Sample_329  Sample_33 Sample_330 Sa
##    23672528   21236621   20726401   20335675   19817550   21911365   28672474   22519490   29551499
## Sample_332 Sample_333 Sample_334 Sample_335 Sample_336 Sample_337 Sample_338 Sample_339  Sample_34 Sa
##    29177516   26751877   27382543   28412881   26956753   30351182   25897012   25661723   22773589
## Sample_341 Sample_342 Sample_343 Sample_344 Sample_345 Sample_346 Sample_347 Sample_348 Sample_349   S
##    22411852   23159050   25386451   22391816   25104267   24103925   24995733   25400273   23937203
## Sample_350 Sample_351 Sample_352 Sample_353 Sample_354 Sample_355 Sample_356 Sample_357 Sample_358 Sa
##    23644976   25677967   23471807   28039008   27191003   30974673   28955689   23375190   26204823
##  Sample_36 Sample_360  Sample_37  Sample_38  Sample_39   Sample_4  Sample_40  Sample_41  Sample_42   S
##    23605309   27538675   22547854   21808157   19702060   27902063   22238066   26365273   27592302
##  Sample_44  Sample_45  Sample_46  Sample_47  Sample_48  Sample_49   Sample_5  Sample_50  Sample_51   S
##    25173832   22791654   23881267   21650451   23150224   25356326   26685556   28085624   23795269
##  Sample_53  Sample_54  Sample_55  Sample_56  Sample_57  Sample_58  Sample_59   Sample_6  Sample_60   S
##    24411833   22972476   23903581   23152774   25015737   25633098   23223994   23656954   27207763
##  Sample_62  Sample_63  Sample_64  Sample_65  Sample_66  Sample_67  Sample_68  Sample_69   Sample_7   S
##    24617158   24381198   23086040   26229090   28633158   26742024   30064215   27098721   21905857
##  Sample_71  Sample_72  Sample_73  Sample_74  Sample_75  Sample_76  Sample_77  Sample_78  Sample_79
##    24866231   24133480   24114391   28226413   26792238   28455857   27123339   25820258   25282890
##  Sample_80  Sample_81  Sample_82  Sample_83  Sample_84  Sample_85  Sample_86  Sample_87  Sample_88   S
##    25246768   31065176   29904244   31663130   33096338   32257316   30259020   25439243   30127125
```

```
##   Sample_9  Sample_90  Sample_91  Sample_92  Sample_93  Sample_94  Sample_95  Sample_96  Sample_97  S
##   25147363   22166321   23633231   22934244   23246987   21281789   20454092   19436759   17289795
## Sample_99
##   22647903
```

```r
pTPM <- cpm(gene_counts)
sapply(as.data.frame(pTPM), sum)
```

```
##    Sample_1  Sample_10 Sample_100 Sample_101 Sample_102 Sample_103 Sample_104 Sample_105 Sample_106 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_108 Sample_109  Sample_11 Sample_110 Sample_111 Sample_112 Sample_113 Sample_114 Sample_115 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_117 Sample_118 Sample_119  Sample_12 Sample_120 Sample_121 Sample_122 Sample_123 Sample_124 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_126 Sample_127 Sample_128 Sample_129  Sample_13 Sample_130 Sample_131 Sample_132 Sample_133 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_135 Sample_136 Sample_137 Sample_138 Sample_139  Sample_14 Sample_140 Sample_141 Sample_142 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_144 Sample_145 Sample_146 Sample_147 Sample_148 Sample_149  Sample_15 Sample_150 Sample_151 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_153 Sample_154 Sample_155 Sample_156 Sample_157 Sample_158 Sample_159  Sample_16 Sample_160 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_162 Sample_163 Sample_164 Sample_165 Sample_166 Sample_167 Sample_168 Sample_169  Sample_17 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_171 Sample_172 Sample_173 Sample_174 Sample_175 Sample_176 Sample_177 Sample_178 Sample_179  S
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_180 Sample_181 Sample_182 Sample_183 Sample_184 Sample_185 Sample_186 Sample_187 Sample_188 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
##  Sample_19 Sample_190 Sample_191 Sample_192 Sample_193 Sample_194 Sample_195 Sample_196 Sample_197 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_199   Sample_2  Sample_20 Sample_200 Sample_201 Sample_202 Sample_203 Sample_204 Sample_205 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_207 Sample_208 Sample_209  Sample_21 Sample_210 Sample_211 Sample_212 Sample_213 Sample_214 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_216 Sample_217 Sample_218 Sample_219  Sample_22 Sample_220 Sample_221 Sample_222 Sample_223 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_225 Sample_226 Sample_227 Sample_228 Sample_229  Sample_23 Sample_230 Sample_231 Sample_232 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_234 Sample_235 Sample_236 Sample_237 Sample_238 Sample_239  Sample_24 Sample_240 Sample_241 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_243 Sample_244 Sample_245 Sample_246 Sample_247 Sample_248 Sample_249  Sample_25 Sample_250 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_252 Sample_253 Sample_254 Sample_255 Sample_256 Sample_257 Sample_258 Sample_259  Sample_26 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_261 Sample_262 Sample_263 Sample_264 Sample_268  Sample_27 Sample_272 Sample_273 Sample_275 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_279  Sample_28 Sample_280 Sample_281 Sample_282 Sample_283 Sample_284 Sample_285 Sample_286 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_288 Sample_289  Sample_29 Sample_290 Sample_291 Sample_292 Sample_293 Sample_294 Sample_295 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_297 Sample_298 Sample_299   Sample_3  Sample_30 Sample_300 Sample_301 Sample_302 Sample_303 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_305 Sample_306 Sample_307 Sample_308 Sample_309  Sample_31 Sample_310 Sample_311 Sample_312 Sa
##        1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06      1e+06
## Sample_314 Sample_315 Sample_316 Sample_317 Sample_318 Sample_319  Sample_32 Sample_320 Sample_321 Sa
```

```
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
## Sample_323 Sample_324 Sample_325 Sample_326 Sample_327 Sample_328 Sample_329  Sample_33 Sample_330 Sa
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
## Sample_332 Sample_333 Sample_334 Sample_335 Sample_336 Sample_337 Sample_338 Sample_339  Sample_34 Sa
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
## Sample_341 Sample_342 Sample_343 Sample_344 Sample_345 Sample_346 Sample_347 Sample_348 Sample_349
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
## Sample_350 Sample_351 Sample_352 Sample_353 Sample_354 Sample_355 Sample_356 Sample_357 Sample_358 Sa
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_36 Sample_360  Sample_37  Sample_38  Sample_39   Sample_4  Sample_40  Sample_41  Sample_42
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_44  Sample_45  Sample_46  Sample_47  Sample_48  Sample_49   Sample_5  Sample_50  Sample_51
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_53  Sample_54  Sample_55  Sample_56  Sample_57  Sample_58  Sample_59   Sample_6  Sample_60
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_62  Sample_63  Sample_64  Sample_65  Sample_66  Sample_67  Sample_68  Sample_69   Sample_7
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_71  Sample_72  Sample_73  Sample_74  Sample_75  Sample_76  Sample_77  Sample_78  Sample_79
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_80  Sample_81  Sample_82  Sample_83  Sample_84  Sample_85  Sample_86  Sample_87  Sample_88
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##   Sample_9  Sample_90  Sample_91  Sample_92  Sample_93  Sample_94  Sample_95  Sample_96  Sample_97  S
##      1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06       1e+06
##  Sample_99
##      1e+06
```

```r
#saveRDS(pTPM,  here::here(data_save_dir,"pTPM_Lifespan.rds"))

ref.samples <- samples %>%
  dplyr::filter(Clinical_condition %in% "Normal") %>%
  dplyr::filter(Treatments %in% "Control") %>%
  dplyr::filter(Percent_oxygen %in% "21") %>%
  pull(RNAseq_sampleID)

ref.column <- data.frame(Median = rowMedians(pTPM[,ref.samples]))
rownames(ref.column) = rownames(pTPM)
pTPM <- bind_cols(ref.column, pTPM)

nTPM <- NOISeq::tmm(pTPM, refColumn = "Median")
sapply(as.data.frame(nTPM), sum)
```

```
##     Median    Sample_1   Sample_10  Sample_100 Sample_101 Sample_102 Sample_103 Sample_104 Sample_105 Sa
##   991481.4   1050766.1   1021626.8   1059849.3    992136.4   1002143.8   1004410.5   1020797.6   1060170.2
## Sample_107 Sample_108 Sample_109   Sample_11 Sample_110 Sample_111 Sample_112 Sample_113 Sample_114 Sa
##   939446.7   1004216.6   1025859.6   1081069.0   1034176.9    923848.5    939013.6    933429.5    911295.5
## Sample_116 Sample_117 Sample_118 Sample_119   Sample_12 Sample_120 Sample_121 Sample_122 Sample_123 Sa
##  1147535.3   1186344.6   1199752.0    856746.2   1071178.9    904099.4    986486.2    955686.8    951149.7
## Sample_125 Sample_126 Sample_127 Sample_128 Sample_129   Sample_13 Sample_130 Sample_131 Sample_132 Sa
##   954388.2    966605.1    941739.6    964392.6    988886.9   1131013.8    983944.9   1013929.6   1037328.4
## Sample_134 Sample_135 Sample_136 Sample_137 Sample_138 Sample_139   Sample_14 Sample_140 Sample_141 Sa
##   966052.5    949961.0    936483.2    935497.4    966344.8    930953.7   1017601.4    980984.2   1008276.6
## Sample_143 Sample_144 Sample_145 Sample_146 Sample_147 Sample_148 Sample_149   Sample_15 Sample_150 Sa
##  1037602.7   1005085.2    922887.2   1017146.6   1012301.8    992302.3   1057783.6    975887.1   1045695.6
## Sample_152 Sample_153 Sample_154 Sample_155 Sample_156 Sample_157 Sample_158 Sample_159   Sample_16 Sa
##  1319097.4   1171766.9   1090194.4   1423246.5   1609946.3   1003877.4   1078730.9   1638355.5    943563.4
```

```
## Sample_161 Sample_162 Sample_163 Sample_164 Sample_165 Sample_166 Sample_167 Sample_168 Sample_169  
## 1033709.3 1029057.5  862784.9 1095373.5 1190964.6 1119450.3  892753.8  988951.9  921656.6
## Sample_170 Sample_171 Sample_172 Sample_173 Sample_174 Sample_175 Sample_176 Sample_177 Sample_178 Sa
##  920897.9  971562.2  954866.3 1335379.4  969426.2  932003.4 1281946.5  942553.9  930057.4
##  Sample_18 Sample_180 Sample_181 Sample_182 Sample_183 Sample_184 Sample_185 Sample_186 Sample_187 Sa
## 1026001.9  943665.3  962007.9  967698.7  951926.4  962286.0  985675.5 1055178.1 1021182.1  
## Sample_189  Sample_19 Sample_190 Sample_191 Sample_192 Sample_193 Sample_194 Sample_195 Sample_196 Sa
##  924228.0 1064841.1  968633.7 1015914.6 1009461.6  973096.4  915891.6  999157.6  975334.9  
## Sample_198 Sample_199   Sample_2  Sample_20 Sample_200 Sample_201 Sample_202 Sample_203 Sample_204 Sa
## 1014185.5  981763.1 1127072.7 1052924.0 1017077.1 1002929.4  946570.3  939863.3  997164.4  
## Sample_206 Sample_207 Sample_208 Sample_209  Sample_21 Sample_210 Sample_211 Sample_212 Sample_213 Sa
##  978580.9  995164.6 1048434.5  997109.2 1026260.6 1053927.0  935633.4  988063.5 1045840.2  
## Sample_215 Sample_216 Sample_217 Sample_218 Sample_219  Sample_22 Sample_220 Sample_221 Sample_222 Sa
## 1054976.6 1062966.4  946389.8  905750.9  935392.3  986706.6  934388.0  954049.4  897514.2  
## Sample_224 Sample_225 Sample_226 Sample_227 Sample_228 Sample_229  Sample_23 Sample_230 Sample_231 Sa
## 1053513.3  928321.3  952841.8  999801.4  978406.4  972211.2 1011741.1  921814.0  974164.8  
## Sample_233 Sample_234 Sample_235 Sample_236 Sample_237 Sample_238 Sample_239  Sample_24 Sample_240 Sa
##  919243.3  924189.0  926176.8 1015612.9  920679.4  917318.9  896150.5 1023412.3  896050.7
## Sample_242 Sample_243 Sample_244 Sample_245 Sample_246 Sample_247 Sample_248 Sample_249  Sample_25 Sa
##  898822.1  871954.1  876832.2 1025349.8  931935.6  900678.6  866820.0  881021.9 1059398.3  
## Sample_251 Sample_252 Sample_253 Sample_254 Sample_255 Sample_256 Sample_257 Sample_258 Sample_259  S
##  904950.0  979308.0  911075.8  903931.6  895790.7  832252.2  969839.4  988993.4 1023860.5  
## Sample_260 Sample_261 Sample_262 Sample_263 Sample_264 Sample_268  Sample_27 Sample_272 Sample_273 Sa
## 1005956.4 1013720.0 1132808.0 1080317.0 1041207.7  591009.2 1055125.3 1039794.2 1058765.5  
## Sample_276 Sample_279  Sample_28 Sample_280 Sample_281 Sample_282 Sample_283 Sample_284 Sample_285 Sa
##  934888.2 1058710.2 1067108.5 1030460.4 1065964.5 1057949.6 1035406.9 1107847.8 1141781.6  
## Sample_287 Sample_288 Sample_289  Sample_29 Sample_290 Sample_291 Sample_292 Sample_293 Sample_294 Sa
##  947230.8  957321.8  925780.0 1053836.6  907061.2  911593.7  902483.5  912028.6  891579.4  
## Sample_296 Sample_297 Sample_298 Sample_299   Sample_3  Sample_30 Sample_300 Sample_301 Sample_302 Sa
##  951616.3  907157.5  932840.3 1005548.0 1115683.9 1090416.1 1017416.0 1008504.7 1025205.6  
## Sample_304 Sample_305 Sample_306 Sample_307 Sample_308 Sample_309  Sample_31 Sample_310 Sample_311 Sa
## 1013868.8  950287.2  976579.7  997119.3 1015931.0  975437.1 1027479.8  957066.0 1017974.3  
## Sample_313 Sample_314 Sample_315 Sample_316 Sample_317 Sample_318 Sample_319  Sample_32 Sample_320 Sa
## 1009472.5  986582.9  951723.5  992946.1 1012313.1  994100.9 1056228.2  957883.7 1016472.8  
## Sample_322 Sample_323 Sample_324 Sample_325 Sample_326 Sample_327 Sample_328 Sample_329  Sample_33 Sa
## 1036409.1 1106093.3 1066129.6 1085573.8 1104581.4 1083780.1 1182702.2  983443.7  934435.5  
## Sample_331 Sample_332 Sample_333 Sample_334 Sample_335 Sample_336 Sample_337 Sample_338 Sample_339  S
##  963981.8  977825.5 1020098.5 1024791.0 1019203.8 1081725.2 1098705.4 1155634.5  930859.5  
## Sample_340 Sample_341 Sample_342 Sample_343 Sample_344 Sample_345 Sample_346 Sample_347 Sample_348 Sa
##  887252.0  911957.3  919930.6  912450.1  914866.1  919987.2  930069.6  932189.7  958607.6  
##  Sample_35 Sample_350 Sample_351 Sample_352 Sample_353 Sample_354 Sample_355 Sample_356 Sample_357 Sa
##  957835.6  968666.3 1004630.4  975995.7  901300.8  928833.4 1016147.8  972336.1  964149.6  
## Sample_359  Sample_36 Sample_360  Sample_37  Sample_38  Sample_39   Sample_4  Sample_40  Sample_41  S
## 1000615.7 1005549.6  989689.6 1041849.4 1085709.2 1031037.7 1022081.1 1064321.8 1102632.4  
##  Sample_43  Sample_44  Sample_45  Sample_46  Sample_47  Sample_48  Sample_49   Sample_5  Sample_50  S
##  881066.4  943020.5  927983.0  975316.4  959968.0  952483.4  914579.4 1112519.2 1051944.1  
##  Sample_52  Sample_53  Sample_54  Sample_55  Sample_56  Sample_57  Sample_58  Sample_59   Sample_6  S
## 1038861.1 1148191.4 1118094.4 1072673.8 1036009.8 1031868.7 1029098.4  986829.3  988950.2  
##  Sample_61  Sample_62  Sample_63  Sample_64  Sample_65  Sample_66  Sample_67  Sample_68  Sample_69  
## 1023985.8 1089695.9 1081765.7 1182696.4  960315.0 1046059.1 1104412.5 1108508.8 1181114.2  
##  Sample_70  Sample_71  Sample_72  Sample_73  Sample_74  Sample_75  Sample_76  Sample_77  Sample_78  S
## 1063963.5  958256.8 1006708.5  953222.5 1079888.5  984256.0 1005736.6 1054483.2 1030095.1  
##   Sample_8  Sample_80  Sample_81  Sample_82  Sample_83  Sample_84  Sample_85  Sample_86  Sample_87  S
## 1050176.5 1064412.7 1023303.7  915363.3  950365.2  956201.3  987311.8  981801.8  986229.9  
```

9

```
## Sample_89    Sample_9  Sample_90   Sample_91   Sample_92   Sample_93   Sample_94   Sample_95   Sample_96
## 1009442.5  1018277.3  1039161.9  1030798.4    985900.3  1083904.0    899046.8    965621.0    979753.6
##  Sample_98  Sample_99
## 1006696.0   963602.1
```
```r
nTPM <- nTPM[,-1]
```

## Filter mitogenes in nTPM data

```r
data_nTPM <- nTPM %>%
  as.data.frame() %>%
  rownames_to_column("Gene") %>%
  pivot_longer(cols = -Gene, names_to = "RNAseq_sampleID", values_to = "nTPM")

mitocarta <- readxl::read_xls(here::here("Data", "MitoCarta", "OriginalData",
                                         "HumanMitoCarta3_0.xls"), sheet = 4) %>%
  select(MitoPathway, Genes) %>%
  na.omit()
```
```
## New names:
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
```
```r
gene_to_pathway <- splitstackshape::cSplit(mitocarta, 'Genes', ',') %>%
  column_to_rownames("MitoPathway") %>%
  t() %>%
  as.data.frame()
gene_to_pathway <- gene_to_pathway %>%
  pivot_longer(cols = colnames(gene_to_pathway), names_to = "Pathway", values_to = "Gene") %>%
  na.omit()

## Mitogenes in data
mitocarta <- readxl::read_xls(here::here("Data", "MitoCarta", "OriginalData","HumanMitoCarta3_0.xls"), s
mitogenes <- mitocarta$Symbol
all_genes <- unique(data_nTPM$Gene)
mitogenes_in_data <- sort(mitogenes[which(mitogenes %in% all_genes)])
length(mitogenes_in_data)
```
```
## [1] 1121
```
```r
mitogenes_not_in_data <- mitogenes[which(!mitogenes %in% mitogenes_in_data)]
mitogenes_not_in_data
```
```
##  [1] "MT-ATP6"        "MT-CO2"        "MT-CO1"        "MT-ND2"        "MT-ND4"        "MT-ND5"
##  [7] "MT-CYB"         "MT-ATP8"       "MT-CO3"        "MT-ND3"        "MT-ND1"        "MT-ND4L"
## [13] "MYG1"           "MT-ND6"        "RP11_469A15.2"
```
```r
manual_mitogene_pull <- c("ND1", "ATP6", "COX2", "COX1", "ND2", "ND4", "ND5", "CYTB", "ATP8", "COX3", "
all_mitogenes_in_data <- c(mitogenes[which(!mitogenes %in% mitogenes_not_in_data)], manual_mitogene_pul
data_mito <- data_nTPM %>%
  filter(Gene %in% all_mitogenes_in_data ) %>%
  mutate(Gene = case_when(
    # renaming to match with mitocarta
    Gene == "C12orf10" ~ "MYG1",
```

```r
    Gene == "ND1" ~ "MT-ND1",
    Gene == "ND2" ~ "MT-ND2",
    Gene == "COX1" ~ "MT-CO1",
    Gene == "COX2" ~ "MT-CO2",
    Gene == "ATP8" ~ "MT-ATP8",
    Gene == "ATP6" ~ "MT-ATP6",
    Gene == "COX3" ~ "MT-CO3",
    Gene == "ND3" ~ "MT-ND3",
    Gene == "ND4L" ~ "MT-ND4L",
    Gene == "ND4" ~ "MT-ND4",
    Gene == "ND5" ~ "MT-ND5",
    Gene == "ND6" ~ "MT-ND6",
    Gene == "CYTB" ~ "MT-CYB",
    TRUE ~ Gene
  )) %>%
  mutate(Genome = case_when(
    grepl("MT-", Gene) ~ "mtDNA",
    TRUE~"NucDNA"
  ))

write_csv(data_mito, here::here("Data","Fibroblasts", "ProcessedData", "RNAseqData_MitoGenes.csv"))
```