Expresión diferencial de genes con DSEq2

Análisis exploratorio de datos

Piero Palacios Bernuy

Contenido

1 Importación de los conteos de genes

Importamos las 100 muestras con su respectiva metadata. Esto lo haremos con el paquete `tximport` debido a que tiene una implementación para importar directamente datos provenientes de *RSEM*.

```
sample_table <- read_excel("D:/tesis cafe/SRaRunTablecoffeaarabica.xlsx")
rownames(sample_table)<-sample_table$Run

dir<-"D:/tesis cafe/DESeq2 coffea/Gene_counts"
list.files(dir)
```

```
 [1] "SRR11196520_.genes.results" "SRR11196521_.genes.results"
 [3] "SRR11196522_.genes.results" "SRR11196523_.genes.results"
 [5] "SRR11196524_.genes.results" "SRR11196525_.genes.results"
 [7] "SRR11196526_.genes.results" "SRR11196527_.genes.results"
 [9] "SRR11196528_.genes.results" "SRR11196529_.genes.results"
[11] "SRR11196530_.genes.results" "SRR11196531_.genes.results"
[13] "SRR11196532_.genes.results" "SRR11196533_.genes.results"
[15] "SRR11196534_.genes.results" "SRR11196535_.genes.results"
[17] "SRR11196536_.genes.results" "SRR11196537_.genes.results"
[19] "SRR11196538_.genes.results" "SRR11196539_.genes.results"
[21] "SRR11711678_.genes.results" "SRR11711679_.genes.results"
[23] "SRR11711680_.genes.results" "SRR11711681_.genes.results"
[25] "SRR11711682_.genes.results" "SRR11711683_.genes.results"
```

```
[27] "SRR11711684_.genes.results" "SRR11711685_.genes.results"
[29] "SRR11711686_.genes.results" "SRR11711687_.genes.results"
[31] "SRR11711688_.genes.results" "SRR11711689_.genes.results"
[33] "SRR11711690_.genes.results" "SRR11711691_.genes.results"
[35] "SRR11711692_.genes.results" "SRR11711693_.genes.results"
[37] "SRR11711695_.genes.results" "SRR11711706_.genes.results"
[39] "SRR11711717_.genes.results" "SRR11711728_.genes.results"
[41] "SRR11711739_.genes.results" "SRR11711760_.genes.results"
[43] "SRR11711771_.genes.results" "SRR11711782_.genes.results"
[45] "SRR11711793_.genes.results" "SRR11711804_.genes.results"
[47] "SRR11711805_.genes.results" "SRR11711806_.genes.results"
[49] "SRR11711807_.genes.results" "SRR11711808_.genes.results"
[51] "SRR11711809_.genes.results" "SRR11711810_.genes.results"
[53] "SRR11711811_.genes.results" "SRR11711812_.genes.results"
[55] "SRR11711813_.genes.results" "SRR11711814_.genes.results"
[57] "SRR11711815_.genes.results" "SRR11711816_.genes.results"
[59] "SRR11711817_.genes.results" "SRR11711818_.genes.results"
[61] "SRR11711819_.genes.results" "SRR11711820_.genes.results"
[63] "SRR11711821_.genes.results" "SRR11711822_.genes.results"
[65] "SRR11711823_.genes.results" "SRR11711824_.genes.results"
[67] "SRR11711825_.genes.results" "SRR11711826_.genes.results"
[69] "SRR11711827_.genes.results" "SRR11711828_.genes.results"
[71] "SRR11711833_.genes.results" "SRR11711844_.genes.results"
[73] "SRR11711855_.genes.results" "SRR11711866_.genes.results"
[75] "SRR11711877_.genes.results" "SRR11711888_.genes.results"
[77] "SRR11711899_.genes.results" "SRR11711907_.genes.results"
[79] "SRR11711908_.genes.results" "SRR11711909_.genes.results"
[81] "SRR11711910_.genes.results" "SRR11711911_.genes.results"
[83] "SRR11711912_.genes.results" "SRR11711913_.genes.results"
[85] "SRR11711914_.genes.results" "SRR11711915_.genes.results"
[87] "SRR11711916_.genes.results" "SRR11711927_.genes.results"
[89] "SRR11711938_.genes.results" "SRR11711949_.genes.results"
[91] "SRR11711950_.genes.results" "SRR11711951_.genes.results"
```

```
files<-list.files(file.path(dir),pattern = ".genes.results", full.names = TRUE)
files <- files[sapply(rownames(sample_table), function(x)grep(x, files, value=FALSE, fixed

names(files)<-rownames(sample_table)

txi.rsem<-tximport(files,type = "rsem",txIn = F,txOut = F)
head(txi.rsem$counts,6,6)
```

```
        SRR11711678 SRR11711679 SRR11711680 SRR11711681 SRR11711682
```

CoarCr001	2506.00	4131.00	4188.50	4212.50	4173.0
CoarCr002	9996.58	21086.07	17713.19	18550.78	18443.5
CoarCr003	18.00	21.00	22.00	24.00	24.5
CoarCr004	4.00	7.00	8.50	6.00	7.5
CoarCr005	4.00	7.00	8.50	6.00	7.5
CoarCr006	18.00	21.00	22.00	24.00	24.5
	SRR11711683	SRR11711684	SRR11711685	SRR11711686	SRR11711687
CoarCr001	4216.50	106460.0	4165.50	25.5	3613.50
CoarCr002	21566.75	221322.4	17425.13	113.0	16020.53
CoarCr003	17.00	1192.0	24.50	0.0	21.00
CoarCr004	4.00	317.0	7.00	0.0	4.50
CoarCr005	4.00	317.0	7.00	0.0	4.50
CoarCr006	17.00	1192.0	24.50	0.0	21.00
	SRR11711688	SRR11711689	SRR11711690	SRR11711691	SRR11711692
CoarCr001	3801.00	3718.00	3702.50	3701.00	3806.00
CoarCr002	15074.59	15338.64	12253.89	17175.86	15488.81
CoarCr003	14.00	19.50	18.50	24.00	16.00
CoarCr004	5.50	6.00	6.00	5.00	8.00
CoarCr005	5.50	6.00	6.00	5.00	8.00
CoarCr006	14.00	19.50	18.50	24.00	16.00
	SRR11711693	SRR11711695	SRR11711706	SRR11711717	SRR11711728
CoarCr001	3677.5	105988.0	106109.0	106615.5	4388.00
CoarCr002	12533.9	201294.7	219917.0	201384.8	12700.52
CoarCr003	21.5	1155.0	1149.0	1063.0	27.00
CoarCr004	6.5	273.0	292.5	271.5	7.00
CoarCr005	6.5	273.0	292.5	271.5	7.00
CoarCr006	21.5	1155.0	1149.0	1063.0	27.00
	SRR11711739	SRR11711760	SRR11711771	SRR11711782	SRR11711793
CoarCr001	29346.00	29361.00	29295.00	614.50	3636.00
CoarCr002	83955.17	86965.28	87741.62	1391.57	12066.16
CoarCr003	160.00	140.50	153.50	5.50	26.00
CoarCr004	63.00	58.50	61.50	2.00	10.50
CoarCr005	63.00	58.50	61.50	2.00	10.50
CoarCr006	160.00	140.50	153.50	5.50	26.00
	SRR11711804	SRR11711805	SRR11711806	SRR11711807	SRR11711808
CoarCr001	3514.50	4009.50	5287.50	5306.0	5257.50
CoarCr002	16224.23	14613.34	20870.63	15915.4	16185.21
CoarCr003	27.50	23.00	32.00	29.5	25.00
CoarCr004	10.50	2.50	5.50	7.5	10.00
CoarCr005	10.50	2.50	5.50	7.5	10.00
CoarCr006	27.50	23.00	32.00	29.5	25.00
	SRR11711809	SRR11711810	SRR11711811	SRR11711812	SRR11711813
CoarCr001	5209.50	5203.00	5314.00	5312.00	894.50
CoarCr002	18181.64	12381.07	21709.61	21792.54	791.01
CoarCr003	25.50	26.00	23.00	25.00	6.00

CoarCr004	8.00	4.50	6.00	9.50	1.00
CoarCr005	8.00	4.50	6.00	9.50	1.00
CoarCr006	25.50	26.00	23.00	25.00	6.00
	SRR11711814	SRR11711815	SRR11711816	SRR11711817	SRR11711818
CoarCr001	4079.50	3562.50	3978.00	4097.00	4049.00
CoarCr002	17754.36	16064.47	15481.25	15905.18	11849.59
CoarCr003	17.00	32.50	24.00	21.00	20.50
CoarCr004	6.00	5.50	4.50	7.00	5.50
CoarCr005	6.00	5.50	4.50	7.00	5.50
CoarCr006	17.00	32.50	24.00	21.00	20.50
	SRR11711819	SRR11711820	SRR11711821	SRR11711822	SRR11711823
CoarCr001	4001.00	4067.00	3990.50	4022.50	3989.00
CoarCr002	14373.16	11873.48	15331.26	17725.95	16280.08
CoarCr003	18.50	20.50	19.50	19.50	18.50
CoarCr004	5.00	7.50	4.50	0.00	4.50
CoarCr005	5.00	7.50	4.50	0.00	4.50
CoarCr006	18.50	20.50	19.50	19.50	18.50
	SRR11711824	SRR11711825	SRR11711826	SRR11711827	SRR11711828
CoarCr001	9124.50	10959.00	3612.00	11172.50	10938.0
CoarCr002	32383.59	38480.42	16779.05	36712.26	39774.3
CoarCr003	43.50	52.50	25.00	50.00	47.0
CoarCr004	11.50	10.50	8.00	14.50	12.0
CoarCr005	11.50	10.50	8.00	14.50	12.0
CoarCr006	43.50	52.50	25.00	50.00	47.0
	SRR11711833	SRR11711844	SRR11711855	SRR11711866	SRR11711877
CoarCr001	10994.50	11250.50	10824.00	10907.00	11155.50
CoarCr002	39785.32	40237.02	36332.75	38709.76	40890.28
CoarCr003	62.50	52.00	47.50	53.50	60.50
CoarCr004	14.00	18.50	14.00	13.00	15.50
CoarCr005	14.00	18.50	14.00	13.00	15.50
CoarCr006	62.50	52.00	47.50	53.50	60.50
	SRR11711888	SRR11711899	SRR11711907	SRR11711908	SRR11711909
CoarCr001	101384.5	106414.5	2684.50	4504.00	4643.00
CoarCr002	182608.5	207486.1	12252.43	23159.29	21085.25
CoarCr003	1122.5	1190.0	12.50	22.00	24.00
CoarCr004	277.5	303.5	3.00	7.00	5.00
CoarCr005	277.5	303.5	3.00	7.00	5.00
CoarCr006	1122.5	1190.0	12.50	22.00	24.00
	SRR11711910	SRR11711911	SRR11711912	SRR11711913	SRR11711914
CoarCr001	106265.0	3614.00	4529.50	4494.50	4558.00
CoarCr002	244885.7	14868.62	24555.91	22512.67	24694.78
CoarCr003	1118.5	26.00	26.00	22.50	31.50
CoarCr004	290.5	4.50	5.50	3.50	5.00
CoarCr005	290.5	4.50	5.50	3.50	5.00
CoarCr006	1118.5	26.00	26.00	22.50	31.50

	SRR11711915	SRR11711916	SRR11711927	SRR11711938	SRR11711949
CoarCr001	4711.50	29296.00	29075.00	29185.50	29139.00
CoarCr002	23858.33	82228.25	79439.67	94723.04	86972.71
CoarCr003	24.00	146.50	166.50	170.00	163.00
CoarCr004	7.50	55.00	54.00	52.50	51.50
CoarCr005	7.50	55.00	54.00	52.50	51.50
CoarCr006	24.00	146.50	166.50	170.00	163.00
	SRR11711950	SRR11711951	SRR11196520	SRR11196521	SRR11196522
CoarCr001	3537.50	3531.50	6016.50	8803.00	3435.50
CoarCr002	16120.86	18258.71	89.39	1158.75	1061.25
CoarCr003	27.50	27.50	0.00	0.00	0.00
CoarCr004	8.00	5.50	0.00	0.00	0.00
CoarCr005	8.00	5.50	0.00	0.00	0.00
CoarCr006	27.50	27.50	0.00	0.00	0.00
	SRR11196523	SRR11196524	SRR11196525	SRR11196526	SRR11196527
CoarCr001	3265.50	4206.50	2625.50	5376.00	4267.00
CoarCr002	248.69	436.98	1695.41	120.27	1959.33
CoarCr003	0.00	0.00	0.00	0.00	0.00
CoarCr004	0.00	0.00	0.00	0.00	0.00
CoarCr005	0.00	0.00	0.00	0.00	0.00
CoarCr006	0.00	0.00	0.00	0.00	0.00
	SRR11196528	SRR11196529	SRR11196530	SRR11196531	SRR11196532
CoarCr001	3022.00	7942.00	3182.00	67136.79	3326.50
CoarCr002	1507.18	856.27	146.99	5838.47	505.91
CoarCr003	0.00	0.00	0.00	0.00	0.00
CoarCr004	0.00	0.00	0.00	0.00	0.00
CoarCr005	0.00	0.00	0.00	0.00	0.00
CoarCr006	0.00	0.00	0.00	0.00	0.00
	SRR11196533	SRR11196534	SRR11196535	SRR11196536	SRR11196537
CoarCr001	5741.00	7323.00	3249.00	12193.00	7431.00
CoarCr002	1472.92	748.07	123.83	871.85	359.44
CoarCr003	0.00	0.00	0.00	0.00	0.00
CoarCr004	0.00	0.00	0.00	0.00	0.00
CoarCr005	0.00	0.00	0.00	0.00	0.00
CoarCr006	0.00	0.00	0.00	0.00	0.00
	SRR11196538	SRR11196539			
CoarCr001	6534.50	46723.50			
CoarCr002	3601.83	3758.95			
CoarCr003	0.00	0.00			
CoarCr004	0.00	0.00			
CoarCr005	0.00	0.00			
CoarCr006	0.00	0.00			

```
txi.rsem$length[txi.rsem$length == 0] <- 1
```

Una vez importado, formaremos un objeto del tipo *DESeq* con el siguiente diseño: $\sim Temperatura$. El análisis exploratorio nos dará indicios sobre si añadir los cultivares y los lugares de los laboratorios como parte del modelo.

```
dds<-DESeqDataSetFromTximport(txi.rsem,colData = sample_table,design =~temp)

dds_coll<-collapseReplicates(dds,groupby = dds$Replicate,run = dds$Run)

rownames(colData(dds_coll))<-dds_coll$Run

head(colData(dds_coll))
```

```
DataFrame with 6 rows and 40 columns
                   Run  Assay Type AvgSpotLen      Bases  BioProject
             <character> <character>  <numeric>  <numeric> <character>
SRR11196521 SRR11196521     RNA-Seq        250 2113436750 PRJNA609253
SRR11196522 SRR11196522     RNA-Seq        250 1487951750 PRJNA609253
SRR11196523 SRR11196523     RNA-Seq        250 2383968750 PRJNA609253
SRR11196524 SRR11196524     RNA-Seq        250 1614495500 PRJNA609253
SRR11196525 SRR11196525     RNA-Seq        250 2294075000 PRJNA609253
SRR11196526 SRR11196526     RNA-Seq        250 2973582250 PRJNA609253
               BioSample      Bytes            Center Name     Consent
              <character>  <numeric>            <character> <character>
SRR11196521 SAMN14239359  900200200 LABORATORIO DE FISIO..      public
SRR11196522 SAMN14239359  629243540 LABORATORIO DE FISIO..      public
SRR11196523 SAMN14239359 1035895087 LABORATORIO DE FISIO..      public
SRR11196524 SAMN14239359  683476291 LABORATORIO DE FISIO..      public
SRR11196525 SAMN14239358  973592764 LABORATORIO DE FISIO..      public
SRR11196526 SAMN14239358 1268556185 LABORATORIO DE FISIO..      public
             DATASTORE filetype DATASTORE provider       DATASTORE region
                    <character>        <character>            <character>
SRR11196521          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
SRR11196522          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
SRR11196523          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
SRR11196524          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
SRR11196525          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
SRR11196526          fastq,sra         gs,ncbi,s3 gs.US,ncbi.public,s3..
              Experiment          Instrument Library Name LibraryLayout
             <character>         <character>  <character>   <character>
SRR11196521  SRX7816484 Illumina HiSeq 2000     OpT_CA_3        PAIRED
SRR11196522  SRX7816483 Illumina HiSeq 2000     OpT_CA_2        PAIRED
SRR11196523  SRX7816485 Illumina HiSeq 2000     OpT_CA_4        PAIRED
SRR11196524  SRX7816482 Illumina HiSeq 2000     OpT_CA_1        PAIRED
SRR11196525  SRX7816481 Illumina HiSeq 2000     OpT_AC_5        PAIRED
```

```
SRR11196526  SRX7816479 Illumina HiSeq 2000     OpT_AC_3        PAIRED
            LibrarySelection  LibrarySource       Organism    Platform
                 <character>    <character>    <character> <character>
SRR11196521           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
SRR11196522           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
SRR11196523           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
SRR11196524           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
SRR11196525           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
SRR11196526           RANDOM TRANSCRIPTOMIC Coffea arabica    ILLUMINA
            ReleaseDate Sample Name   SRA Study   dev_stage BioSampleModel
              <POSIXct> <character> <character> <character>    <character>
SRR11196521  2020-02-28      OpT_CA   SRP251013   seedlings          Plant
SRR11196522  2020-02-28      OpT_CA   SRP251013   seedlings          Plant
SRR11196523  2020-02-28      OpT_CA   SRP251013   seedlings          Plant
SRR11196524  2020-02-28      OpT_CA   SRP251013   seedlings          Plant
SRR11196525  2020-02-28      OpT_AC   SRP251013   seedlings          Plant
SRR11196526  2020-02-28      OpT_AC   SRP251013   seedlings          Plant
            geo_loc_name_country geo_loc_name_country_continent
                     <character>                     <character>
SRR11196521               Brazil                   South America
SRR11196522               Brazil                   South America
SRR11196523               Brazil                   South America
SRR11196524               Brazil                   South America
SRR11196525               Brazil                   South America
SRR11196526               Brazil                   South America
                      geo_loc_name      tissue         AGE   Ecotype
                       <character> <character> <character> <logical>
SRR11196521 Brazil:Minas Gerais\\..      Leaves          NA        NA
SRR11196522 Brazil:Minas Gerais\\..      Leaves          NA        NA
SRR11196523 Brazil:Minas Gerais\\..      Leaves          NA        NA
SRR11196524 Brazil:Minas Gerais\\..      Leaves          NA        NA
SRR11196525 Brazil:Minas Gerais\\..      Leaves          NA        NA
SRR11196526 Brazil:Minas Gerais\\..      Leaves          NA        NA
            growth_protocol     temp Rango_temperatura   Replicate
                <character> <factor>         <logical> <character>
SRR11196521              NA  23_19_C                NA          10
SRR11196522              NA  23_19_C                NA          11
SRR11196523              NA  23_19_C                NA          12
SRR11196524              NA  23_19_C                NA          13
SRR11196525              NA  23_19_C                NA          14
SRR11196526              NA  23_19_C                NA          15
                  Cultivar common_name    Stranded                Article
               <character> <character> <character>            <character>
SRR11196521 Catuai IAC 144      coffee          No https://www.ncbi.nlm..
SRR11196522 Catuai IAC 144      coffee          No https://www.ncbi.nlm..
```

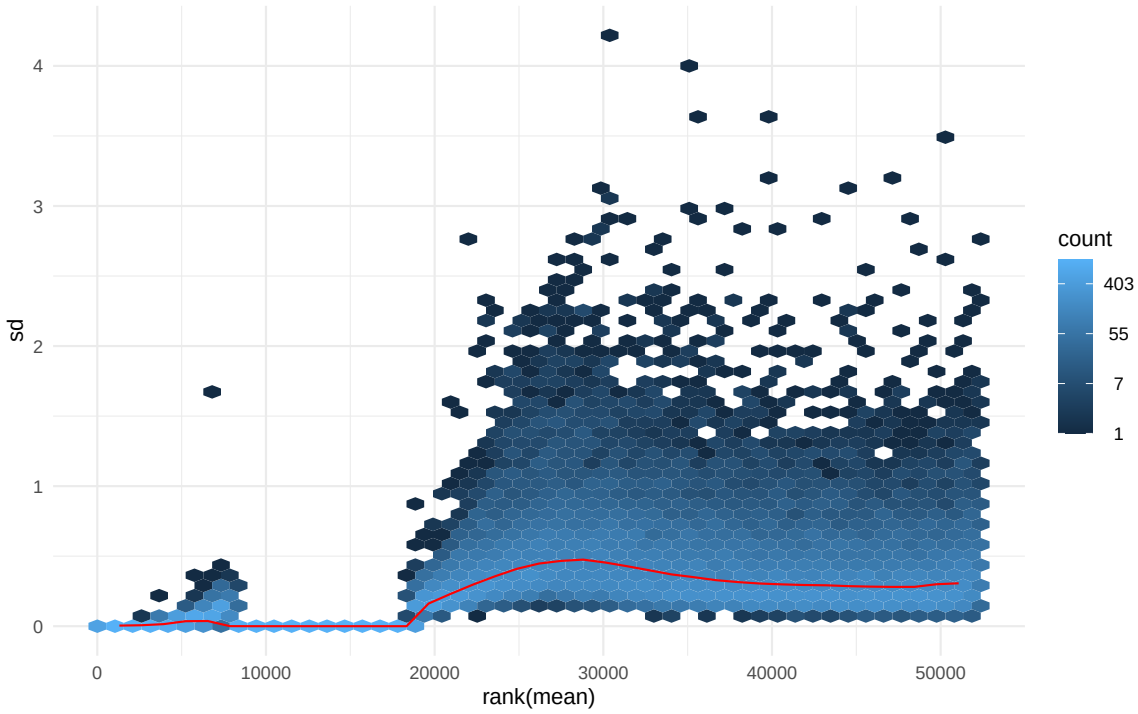
```
SRR11196523 Catuai IAC 144      coffee          No https://www.ncbi.nlm..
SRR11196524 Catuai IAC 144      coffee          No https://www.ncbi.nlm..
SRR11196525          Acaua      coffee          No https://www.ncbi.nlm..
SRR11196526          Acaua      coffee          No https://www.ncbi.nlm..
            runsCollapsed
              <character>
SRR11196521   SRR11196521
SRR11196522   SRR11196522
SRR11196523   SRR11196523
SRR11196524   SRR11196524
SRR11196525   SRR11196525
SRR11196526   SRR11196526
```

2 Análisis exploratorio de datos

Usaremos la transformación `rlog` del paquete DESeq2 para la exploración de los datos de conteo.

Se puede visualizar que rlog controla bien la varianza.

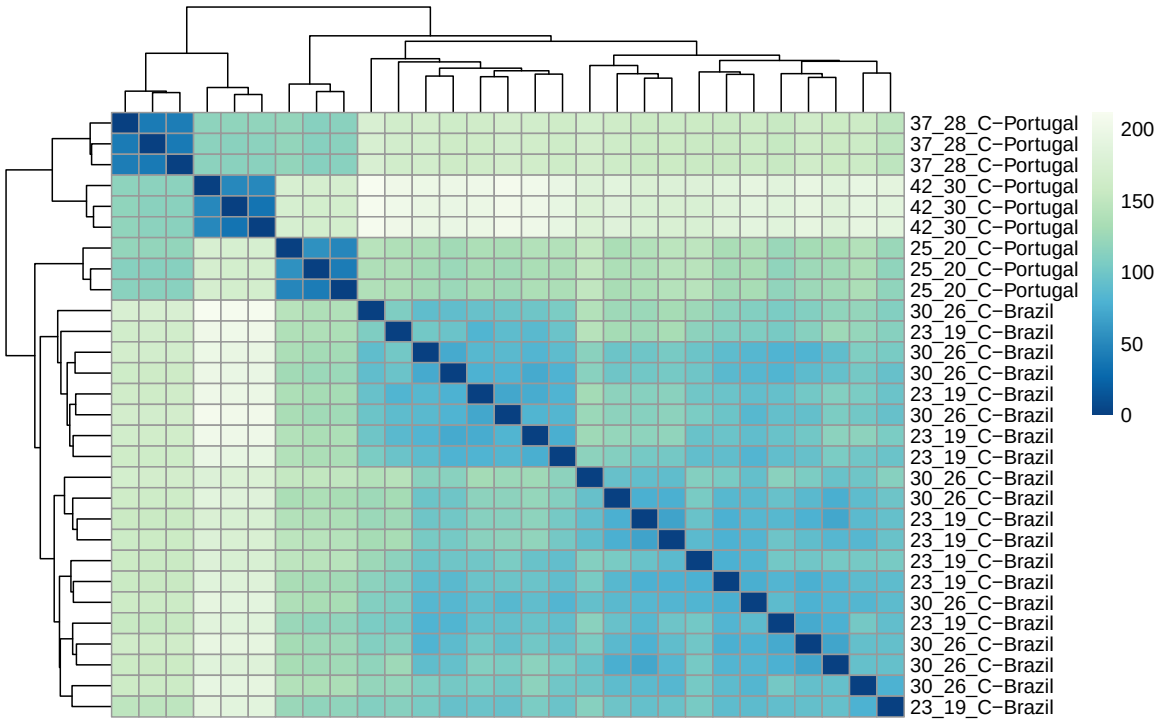
```
library(vsn)
library(pheatmap)

rld<-rlog(dds_coll)
meanSdPlot(assay(rld))
```



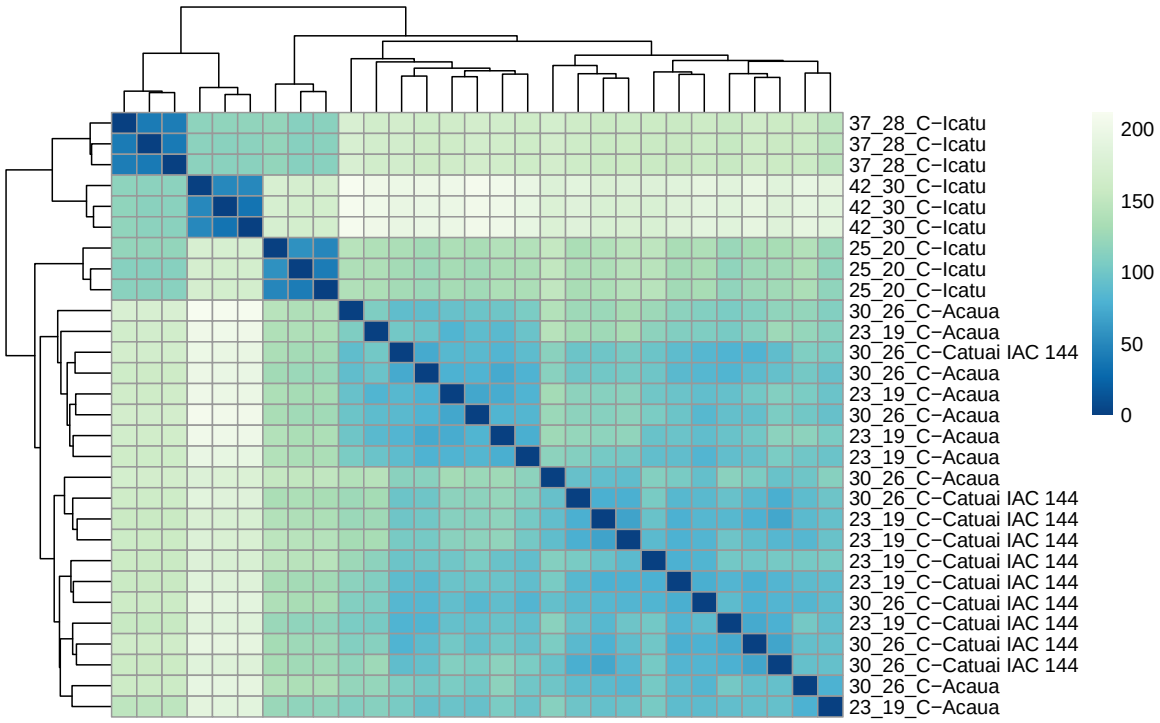

```
dds_coll<-estimateSizeFactors(dds_coll)
select <- order(rowMeans(counts(dds_coll,normalized=TRUE)),
                decreasing=TRUE)[1:30]
```

Tambien veamos la distancia entre muestras lo que nos puede dar indicios de los datos sin temperatura.

```
sampleDists <- dist(t(assay(rld)))

library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(rld$temp, rld$geo_loc_name_country, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "GnBu")) )(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors)
```

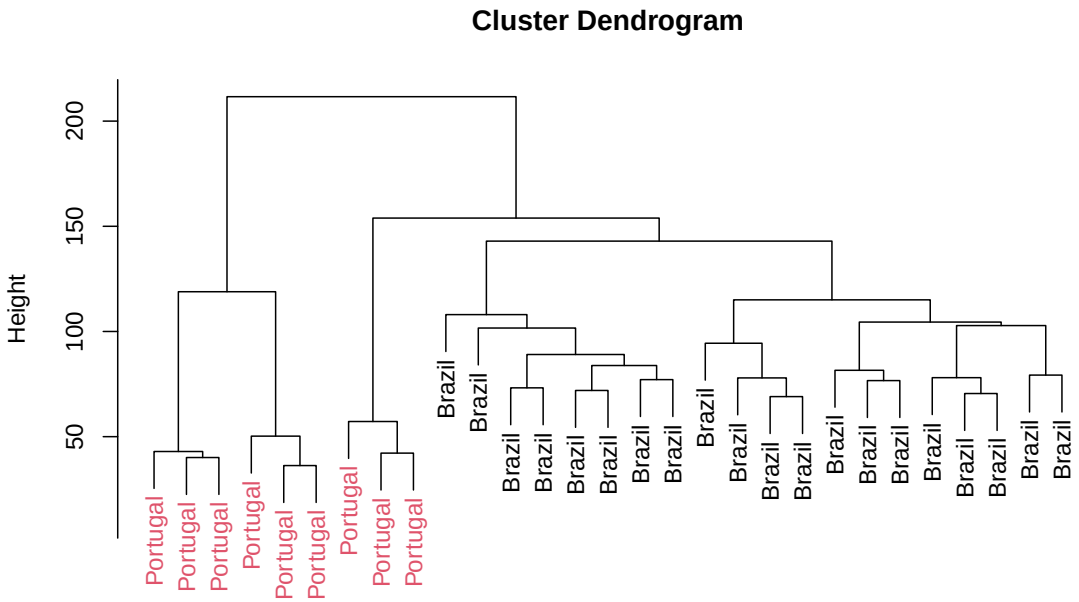



```
rownames(sampleDistMatrix) <- paste(rld$temp, rld$Cultivar, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "GnBu")) )(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors)
```

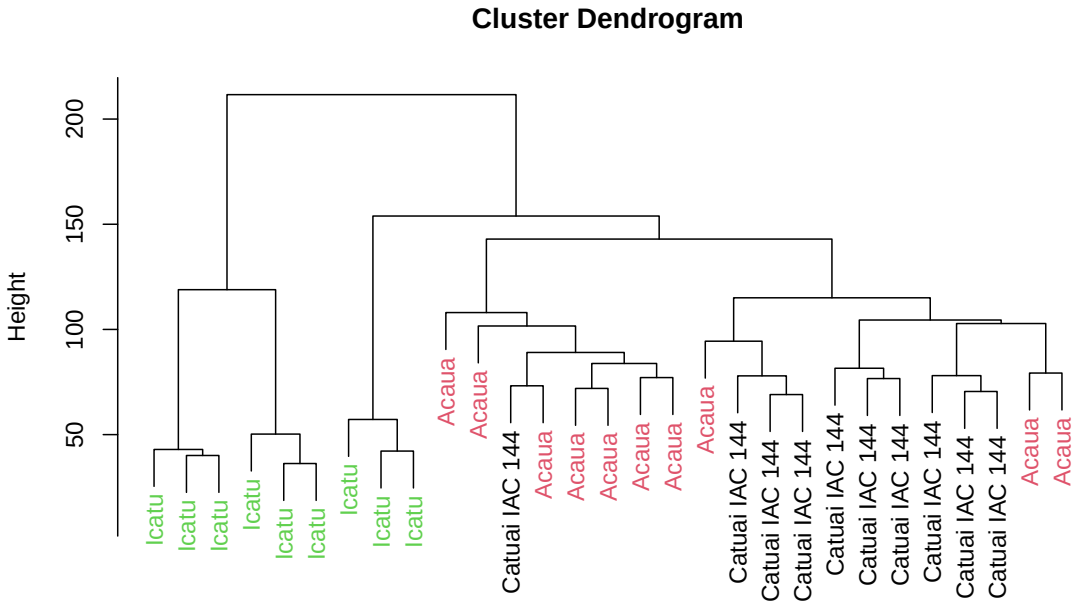


Otra manera de ver las distancia entre muestras es con métodos de agrupamiento o *clustering*. Para esto podemos usar los dendrogramas o k-means.

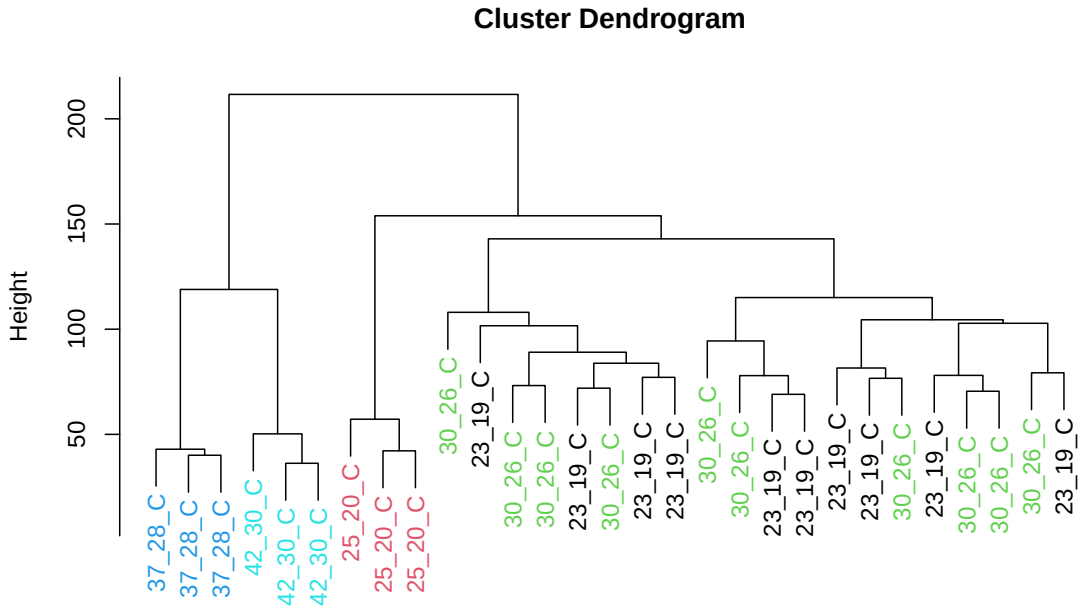
```
library(rafalib)

hc<-hclust(sampleDists)

myplclust(hc,labels = dds_coll$geo_loc_name_country,lab.col = as.fumeric(dds_coll$geo_loc_
```




```
myplclust(hc,labels = dds_coll$Cultivar,lab.col = as.fumeric(dds_coll$Cultivar))
```

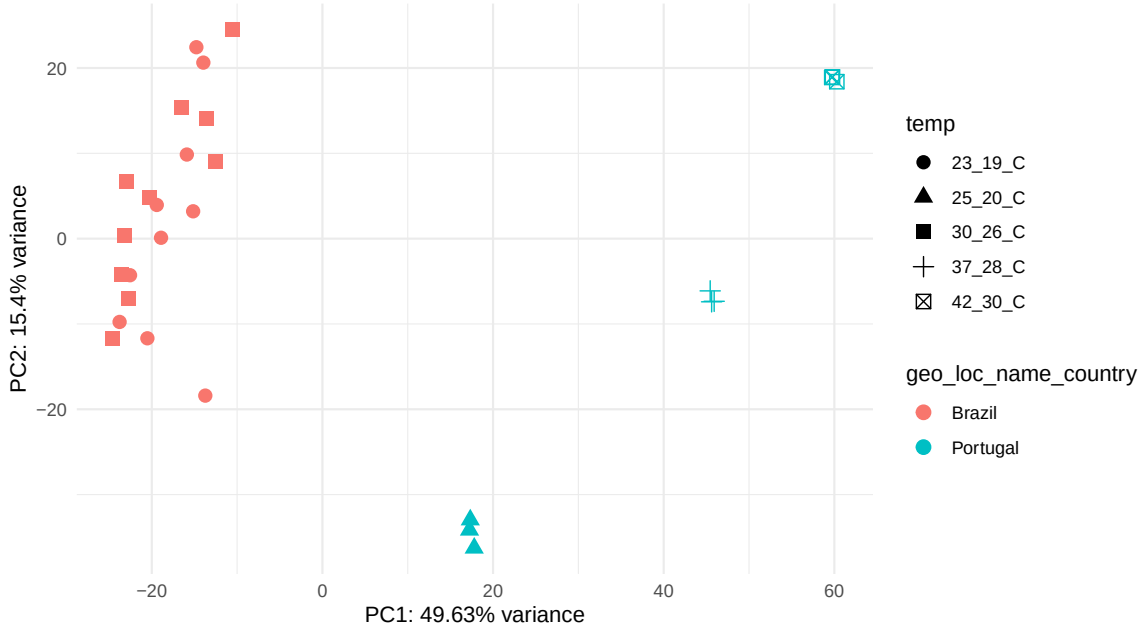


```
myplclust(hc,labels = as.character(dds_coll$temp),lab.col = as.numeric(dds_coll$temp))
```

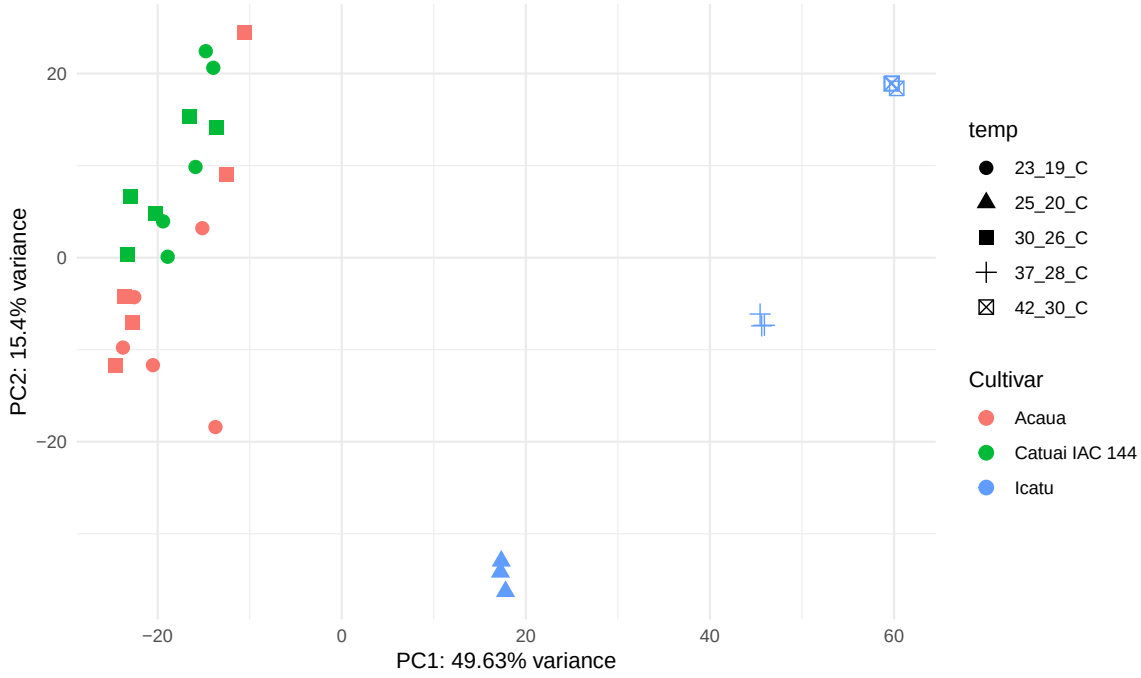


El pca tambien nos puede servir para identificar a esos datos sin temperatura. Además, nos ayudará a confirmar si las variables *Cultivar* y *Lugar* deben ir en el modelo de **DESeq2**.

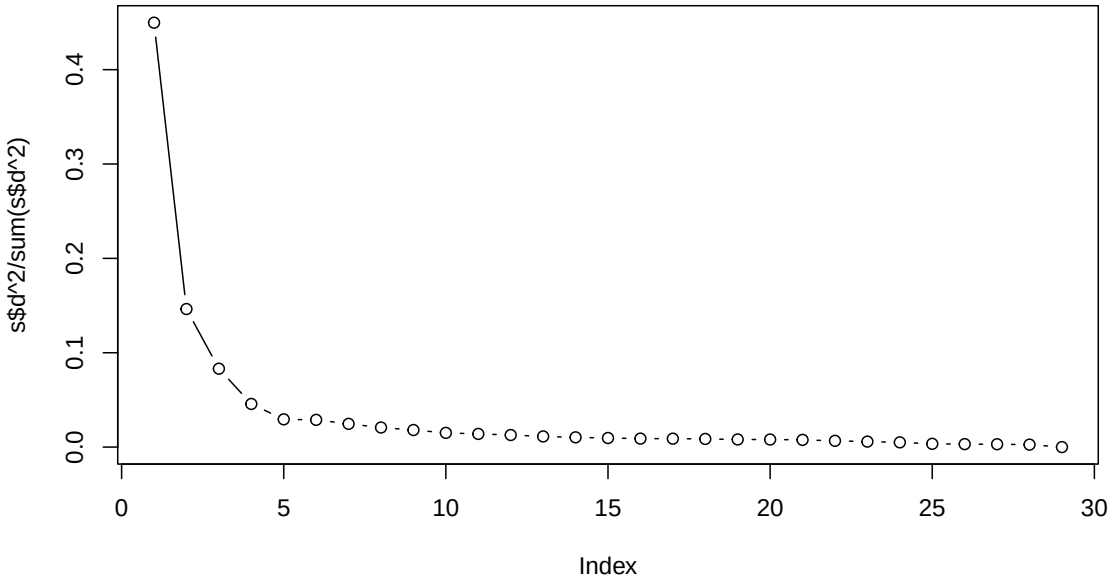
```
pcaData <- plotPCA(rld, intgroup=c("temp","geo_loc_name_country"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"),2)
ggplot(pcaData, aes(PC1, PC2, color=geo_loc_name_country, shape=temp)) +
  geom_point(size=3) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  coord_fixed()
```




```
pcaData2 <- plotPCA(rld, intgroup=c("temp","Cultivar"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"),digits = 2)
ggplot(pcaData2, aes(PC1, PC2, color=Cultivar, shape=temp)) +
  geom_point(size=3) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  coord_fixed()
```




```
y<-assay(rld)-rowMeans(assay(rld))
s<-svd(y)
plot(s$d^2/sum(s$d^2),type="b")
```


s$d^2/sum(s$d^2)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
30
Index