Pipeline PEC1

Alba Moya Garcés

2 de mayo, 2020

Contents

1. Identificar que grupos hay y a qué grupo pertenece cada muestra.
2. Control de calidad de los datos crudos
3. Normalización
4. [Control de calidad de los datos normalizados] (opcional)
5. Filtraje no específico [opcional]
6. Identificación de genes diferencialmente expresados
7. Anotación de los resultados:
8. Comparación entre distintas comparaciones (si hay más de una comparación, ver que genes han sido seleccionados en más de una comparación)
9. Análisis de significación biológica ("Gene Enrichment Analysis")

Estudio elegido:

Gene expression in mitotic tissues of Drosophila larvae without centrosomes or too many centrosomes

La expresión genética de las líneas mutantes se compararon con ambas líneas salvajes de control.

- Tipo de microarrays que utilizan: se hibridaron con arrays *Affymetrix Drosophila Genome 2.0*
- Número de muestras y grupos que contiene el estudio:

Se diseccionaron tejidos mitóticos (cerebrales y de los discos imaginales) de 10 larvas de *Drosophila* en estadío 3:

ausencia de centrosomas | | exceso de centrosomas | tipo salvaje | |
DSas-4 | DSas-6 | SakOE | WT w67 | WT OregonR|
| | | | |

Se extrajo el ARN de **tres réplicas** por cada línea (15 muestras en total)

|GSM864362 | brains and imaginal discs from D-Sas4 mutant 3rd instar Drosophila larvae | biological replicate 1| |GSM864363 | brains and imaginal discs from D-Sas4 mutant 3rd instar Drosophila larvae | biological replicate 2|

|GSM864364 brains and imaginal discs from D-Sas4 mutant 3rd instar Drosophila larvae, biological replicate 3| |GSM864365 brains and imaginal discs from D-Ssas6 mutant 3rd instar Drosophila larvae, biological replicate 1 GSM864366 brains and imaginal discs from D-Ssas6 mutant 3rd instar Drosophila larvae, biological replicate 2 GSM864367 brains and imaginal discs from D-Ssas6 mutant 3rd instar Drosophila larvae, biological replicate 3 GSM864368 brains and imaginal discs from Sak overexpressing 3rd instar Drosophila larvae, biological replicate 1 GSM864369 brains and imaginal discs from Sak overexpressing 3rd instar Drosophila larvae, biological replicate 2 GSM864370 brains and imaginal discs from Sak overexpressing 3rd instar Drosophila larvae, biological replicate 3 GSM864371 brains and imaginal discs from w67 wild type 3rd instar Drosophila larvae, biological replicate 1 GSM864372 brains and imaginal discs from w67 wild type 3rd instar Drosophila larvae, biological replicate 2 GSM864373 brains and imaginal discs from w67 wild type 3rd instar Drosophila larvae, biological replicate 3 GSM864374 brains and imaginal discs from OregonR wild type 3rd instar Drosophila larvae, biological replicate 1 GSM864375 brains and imaginal discs from OregonR wild type 3rd instar Drosophila larvae, biological replicate 2 GSM864376 brains and imaginal discs from OregonR wild type 3rd instar Drosophila larvae, biological replicate 3

- Que pregunta principal persigue responder: Evaluar cómo la pérdida o amplificación de los centrosomas puede afectar la fisiología celular al perfilarse el transcriptoma global cerebral y de los discos imaginales de las larvas de *Drosophila*.
- Enlace a los datos y/o al artículo artículo: https://bio.biologists.org/content/1/10/983.short
- Comentarios adicionales

Methods

Preparación del área de trabajo:

Para llevar a cabo un análisis de microarrays, el analista debe gestionar una gran cantidad de archivos entre aquesllos que ocupan los datos originales y los generados durante su análisis. Es por ello que siempre se debería comenzar creando als carpetas necesarias para simplificar la ruta de trabajo. Se recomienda generar una **carpeta principal** con el nombre de nuestro proyecto en cuyo interior alojaremos una carpeta con los archivos de **datos** y otra con los **resultados** generados del análisis

Estas carpetas las genereamos rápidamente desde el explorador de archivos o la consola de cualquiera de los sistemas operativos usuales. Desde R también podemos generar estas subcarpetas mediante el siguiente código:

El código completo para desarrollar este análisis, o cualquier otro a partir de su adaptación, puede descargarse del siguiente repositorio de *GitHub*:

https://github.com/albamgarces/reanalisis_microarrays.git.

Instalación de paquetes en R

Se necesitarán paquetes adicionales a los incluidos en la intalación básica de R para poder llevar a cabo el análisis. Estos paquetes pueden descargarse tanto del repositorio CRAN para los paquetes típicos de R o directamente de Bioconductor para las funciones del mismo.

A continuación se muestran los paquetes necesarios fara este estudio que requieren instalación:

Descarga de los datos

En lugar de descargar los archivos .CEL y construir manualmente el archivo "targets", se utilizará el paquete `geoQuery` para generar de forma automática el objeto ExpressionSet necesario para el análisis. De esta forma se evitan posibles errores de transcripción o codificador por parte del analista.

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 18952 features, 15 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM864362 GSM864363 ... GSM864376 (15 total)
##   varLabels: title geo_accession ... tissue:ch1 (41 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 1616608_a_at 1622892_s_at ... AFFX-TrpnX-M_at (18952
##     total)
##   fvarLabels: ID CLONE_ID_LIST ... Gene Ontology Molecular Function (16
##     total)
##   fvarMetadata: Column Description labelDescription
## experimentData: use 'experimentData(object)'
##   pubMedIds: 23213376
## Annotation: GPL1322
```

```
## [1] "genotype: DSas-4 mutant"      "genotype: DSas-6 mutant"
## [3] "genotype: OregonR wild type"  "genotype: Sak overexpression"
## [5] "genotype: white wild type"
```

El objeto `expressionSet`combina las difetrentes fuentes de información del estudio en una única estructura. Además de incluir toda la información generada durante el desarrollo del experimento, podemos cambiar la visualización de los datos para que sea más sencillo su uso.

Control de calidad de los datos sin procesar

Se debe primero analizar si los datos tiene suficiente calidad para poder trabajar con ellos. Unos datos de mala calidad prodían producir demasiado ruido en el análisis que no será resuelto al realizar el proceso de normalización.

El primer para llevar esto a cabo será descargar el paquete `ArrayQualityMetrics` que nos permite desarrollar un estudio de calidad de los datos. Si resulta algún array fuera de los límites de calidad propuestos, aparecerá marcado con un asterisco y podrá ser detectado inmediatamente. Si el mismo array destaca tres veces, debería ser analizado y considerar su eliminación del estudio para mejorar cualitativamente el experimento.

```
arrayQualityMetrics(rawData, outdir="./results/rawdata_quality", force=TRUE)
```

Se realizará un análisis conjunto de la calidad de los datos y se creará una nueva carpeta con un informe llamado *index.html* en el que podremos acceder al resumen de los análisis desarrollados y que nos indicará las muestras de calidad dudosa mediante una marca en cada análisis y muestra en las que la calidad sea deficiente. En el objeto de nuestro estudio y como podemos ver en la figura @ref(fig:dataset) no tenemos ninguna muestra cuya calidad deba preocuparnos.

Desde este documento, se puede acceder a los gráficos de medición de la calidad de cada una de las muestras. A continuación desarrollaremos algunas de ellas de manera global.

Gráficos de densidad

Mediante un histograma de densidad de Kernel, podemos hacernos una idea de las distribuciones de los distintos arrays del conjunto de datos. En la figura @ref(fig:histrawData)Podemos apreciar que, probablemente debido al gran número de arrays, todos siguen el mismo patrón de distribución de la señal.

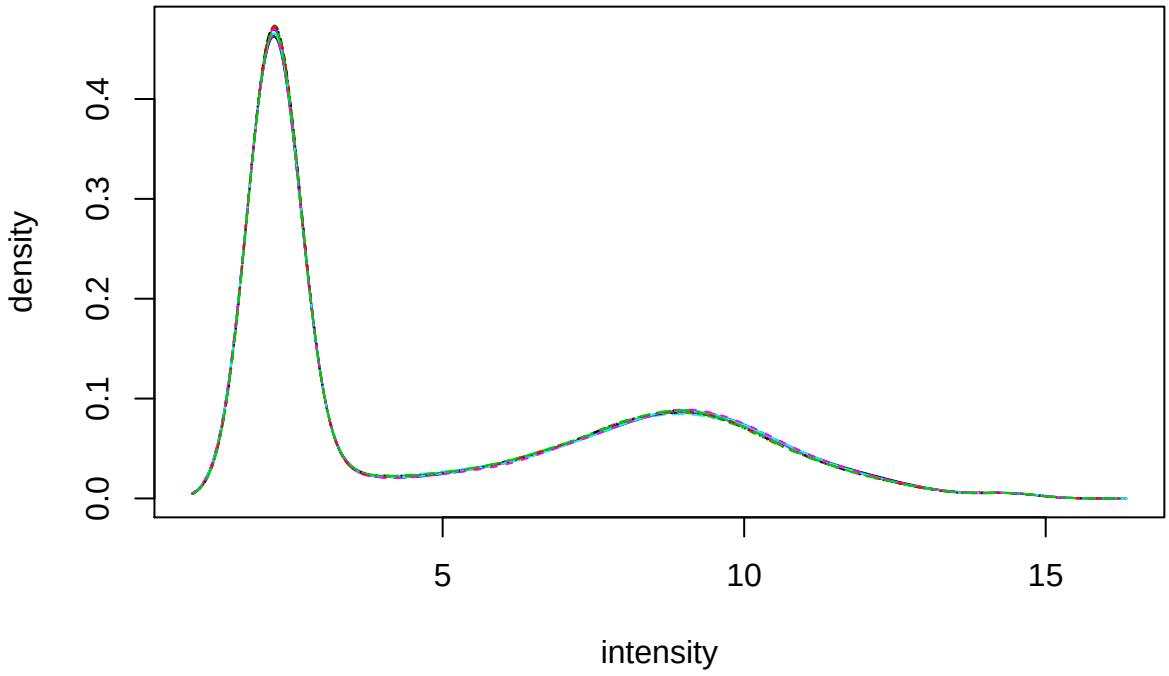

Figure 1: Histograma de los arrays del conjunto de datos.

Diagramas de cajas

El diagrama de cajas también no mostrará la distribución de las intensidades, en la figura @ref(fig:PCARaw) se pueden apreciar pequeñas variaciones esperables en los datos sin procesar.

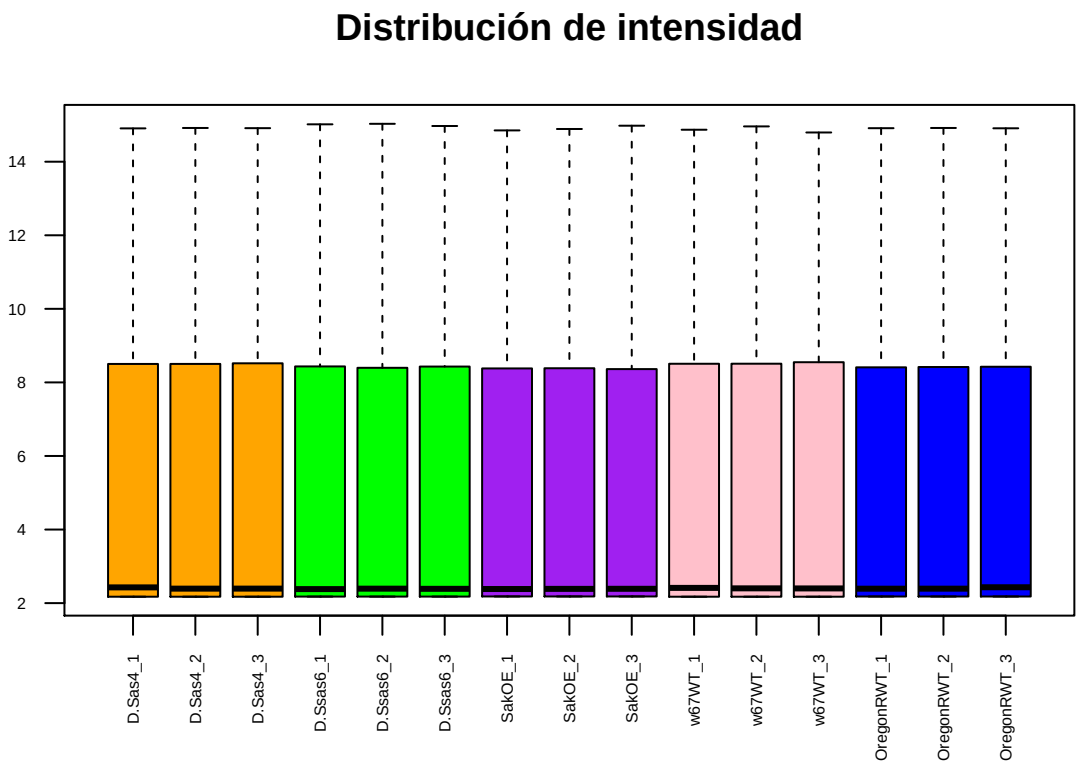

Figure 2: Diagramas de caja de la intensidad de los arrays para los datos sin procesar.

Análisis Componentes Principales

Mediante el análisis de componentes principales podemos detectar si las muestras se agrupan entre otras muestras del mismo grupo o si no hay una clara correspondencia entre ellas. Que las muestras no se agrupen por "familias" podría ser debido al efecto *batch* por defectos técnicos.

Podemos realizar el análisis de componentes principales (ACP) de los datos, observando en el gráfico de la figura @ref(fig:plot_rawData_ACP), la distribución de las dos primeras componentes de la expresión de cada gen (observaciones) sobre cada muestra (variables). Podemos ver como se distribuyen uniformemente a lo largo de la primera componente principal, salvo una perturbación que ocurre en valores altos de ambas componentes que habría que analizar.

```
## Importance of components:
##                             PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     13.5716 0.39882 0.34649 0.3322 0.22932 0.19722 0.18510
## Proportion of Variance  0.9967 0.00086 0.00065 0.0006 0.00028 0.00021 0.00019
## Cumulative Proportion   0.9967 0.99758 0.99823 0.9988 0.99911 0.99932 0.99950
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
```

```
## Standard deviation     0.14542 0.12419 0.11269 0.10449 0.09683 0.09171 0.08711
## Proportion of Variance 0.00011 0.00008 0.00007 0.00006 0.00005 0.00005 0.00004
## Cumulative Proportion  0.99962 0.99970 0.99977 0.99983 0.99988 0.99993 0.99997
##                           PC15
## Standard deviation     0.07759
## Proportion of Variance 0.00003
## Cumulative Proportion  1.00000

## [1] 15 15
```

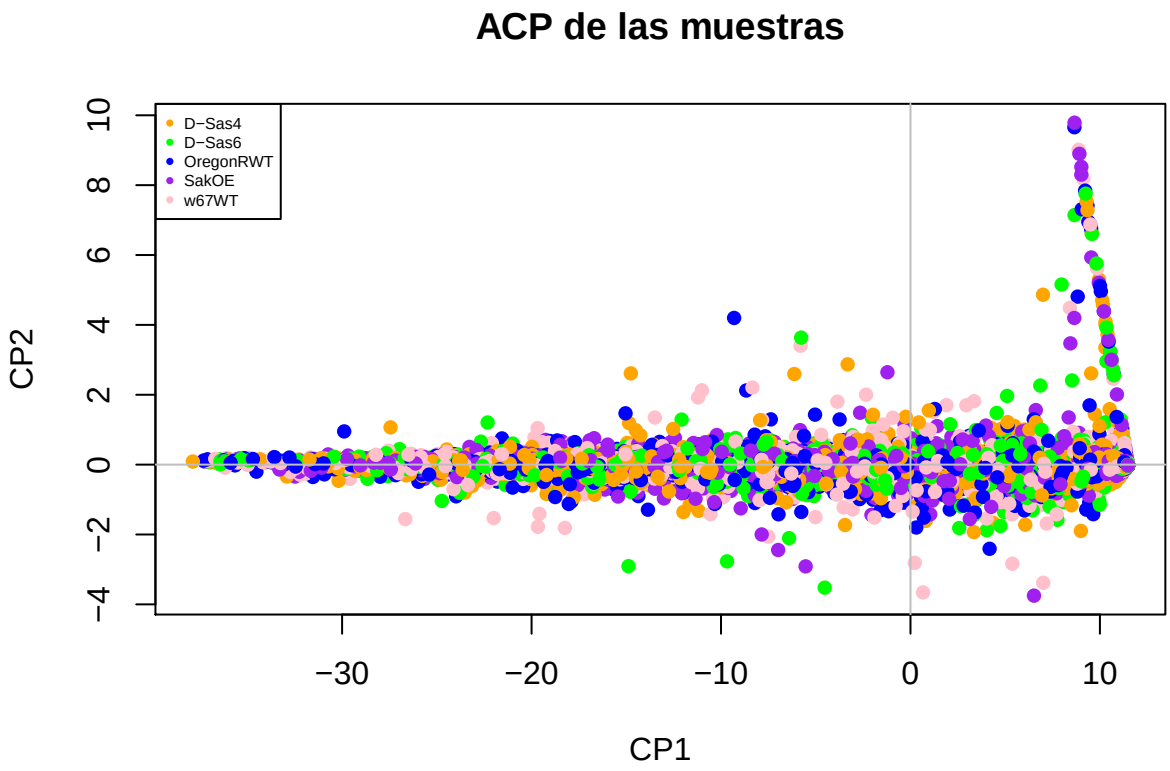

Figure 3: Dos primeras componentes principales de los datos sin procesar utilizando como variables la expresión en las muestras

Pero lo que nos interesa es considerar las diferentes muestras como observaciones, de forma que para cada muestra tenemos el perfil de expresión sobre todos los genes. De esta forma, podremos localizar rápidamente cuál es la principal fuente de variabilidad.

```
#convertimos en matriz los datos para poder analizar los CP
#transponemos la matriz para indicar que las muestras son las observaciones
#y los genes las variables
trawData_ACP <-prcomp(t(as.matrix(rawData)), center = TRUE, scale=FALSE)
summary(trawData_ACP)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
```

```
## Standard deviation     14.7325 13.3284 12.2417 8.43790 7.44894 6.83193 5.35075
## Proportion of Variance  0.2577  0.2109  0.1779 0.08452 0.06587 0.05541 0.03399
## Cumulative Proportion   0.2577  0.4685  0.6464 0.73096 0.79683 0.85224 0.88622
##                            PC8     PC9    PC10   PC11    PC12   PC13    PC14
## Standard deviation     4.60939 4.15177 3.84438 3.5661 3.38020 3.2057 2.85676
## Proportion of Variance 0.02522 0.02046 0.01754 0.0151 0.01356 0.0122 0.00969
## Cumulative Proportion  0.91145 0.93191 0.94945 0.9646 0.97811 0.9903 1.00000
##                            PC15
## Standard deviation     3.497e-13
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```
dim(trawData_ACP$rotation)
```

```
## [1] 18952    15
```

El gráfico de la figura @ref(fig:plot_trawData_ACP) nos muestra las dos primeras componentes principales de la expresión sobre los genes de cada muestra. Se puede ver

```
## [[1]]
## NULL
##
## [[2]]
## NULL
```

```
## Importance of components:
##                            PC1     PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     13.832 12.9369 9.1107 7.93837 7.19580 5.56387 4.78452
## Proportion of Variance  0.277  0.2423 0.1202 0.09124 0.07497 0.04482 0.03314
## Cumulative Proportion   0.277  0.5193 0.6395 0.73072 0.80568 0.85050 0.88364
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     4.30954 3.98952 3.70366 3.50786 3.32671 2.96494
## Proportion of Variance 0.02689 0.02304 0.01986 0.01782 0.01602 0.01273
## Cumulative Proportion  0.91053 0.93358 0.95343 0.97125 0.98727 1.00000
##                            PC14
## Standard deviation     4.789e-13
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```
## [1] 18952    14
```

```
## Importance of components:
##                            PC1     PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     13.832 12.9369 9.1107 7.93837 7.19580 5.56387 4.78452
## Proportion of Variance  0.277  0.2423 0.1202 0.09124 0.07497 0.04482 0.03314
## Cumulative Proportion   0.277  0.5193 0.6395 0.73072 0.80568 0.85050 0.88364
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     4.30954 3.98952 3.70366 3.50786 3.32671 2.96494
## Proportion of Variance 0.02689 0.02304 0.01986 0.01782 0.01602 0.01273
## Cumulative Proportion  0.91053 0.93358 0.95343 0.97125 0.98727 1.00000
```

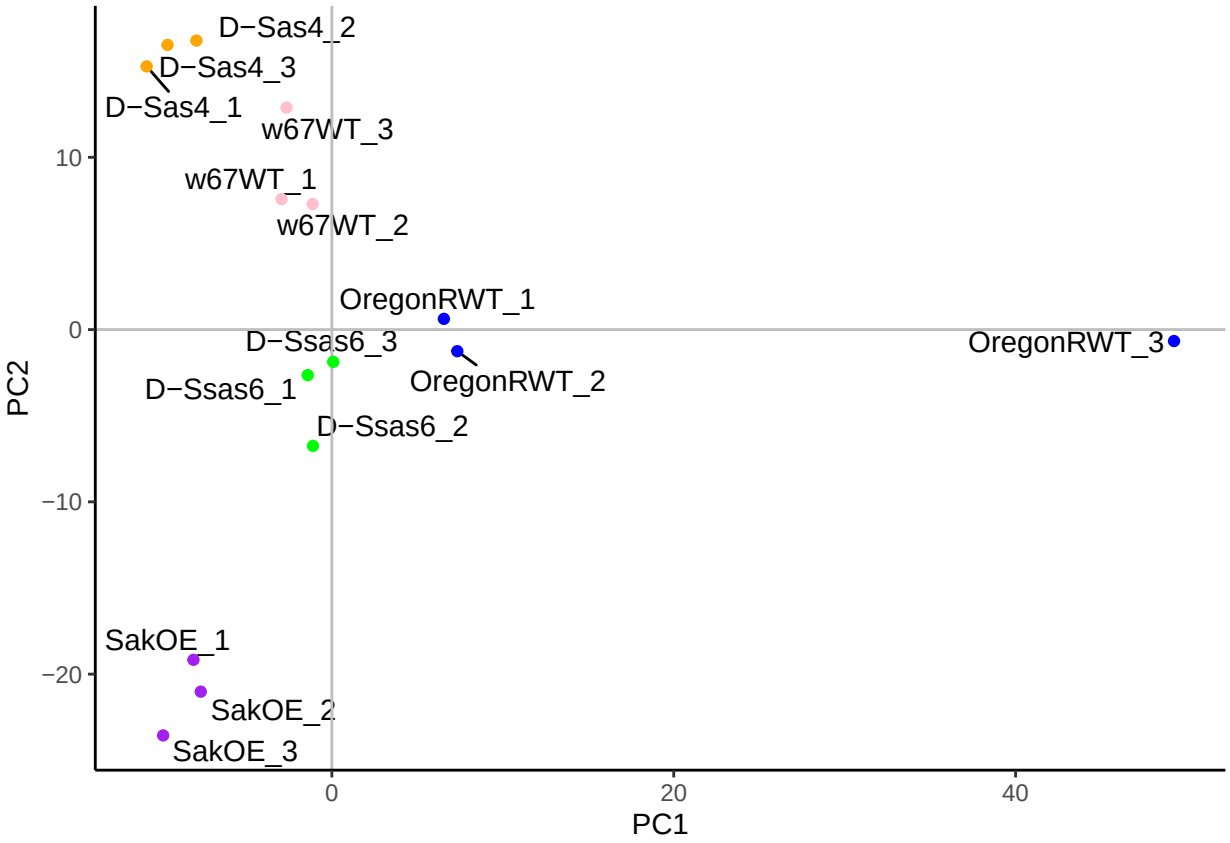

Figure 4: Dos primeras componentes principales de los datos sin procesar utilizando como variables la expresión de los genes

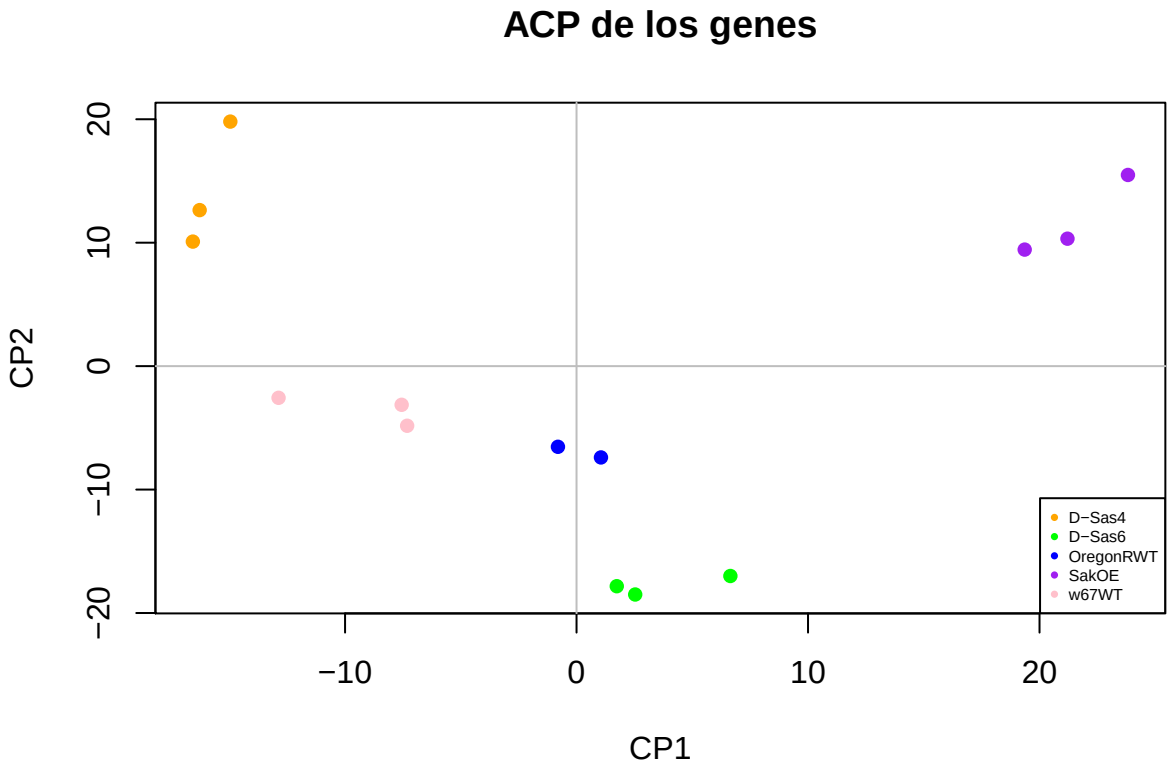

Figure 5: Dos primeras componentes principales de los datos sin procesar utilizando como variables la expresión de los genes

```
##                             PC14
## Standard deviation     4.789e-13
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```
## [1] 18952    14
```

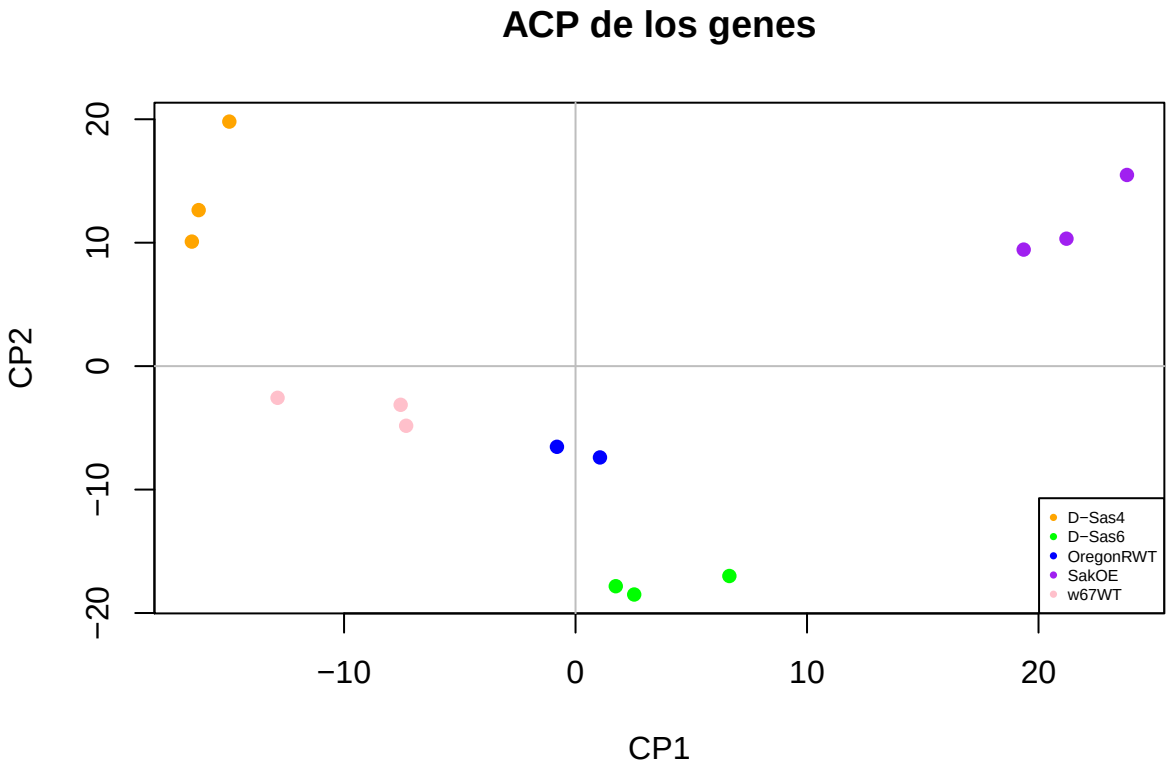

Figure 6: Dos primeras componentes principales de los datos sin procesar utilizando como variables la expresión de los genes

Baumbach, Janina, Mitchell P Levesque, and Jordan W Raff. 2012. "Centrosome Loss or Amplification Does Not Dramatically Perturb Global Gene Expression in Drosophila." *Biology Open* 1 (10). The Company of Biologists Ltd: 983–93.