
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

# Identificación de biomarcadores genéticos asociados a diversos estreses en Glycine max (soja) utilizando datos públicos de RNA-Seq

Tesis de Licenciatura en Ciencias de Datos

Guido Freire

Director: Maximo Rivarola
Codirector: Diego Zavallo
Buenos Aires, 2024

# IDENTIFICACIÓN DE BIOMARCADORES GENÉTICOS ASOCIADOS A DIVERSOS ESTRESES EN GLYCINE MAX (SOJA) UTILIZANDO DATOS PÚBLICOS DE RNA-SEQ

La soja es uno de los cultivos más importantes a nivel global. En Argentina, ocupa un lugar preponderante en la matriz exportadora, siendo el tercer productor mundial y el principal exportador de aceite y harina de soja. Sin embargo, su producción enfrenta importantes desafíos ambientales.

A partir de los años 2000, gracias al advenimiento de técnicas de secuenciación genética masiva (NGS en inglés por *Next Generation Sequencing*), se produjo un salto exponencial en el volumen de datos genómicos de soja disponibles. Se destaca el uso de estos en estudios comparativos y de variabilidad genética para identificar mutaciones asociadas a la resistencia de factores ambientales.

Si bien la mayoría de estos datos son abiertos, nunca han sido explorados holísticamente hasta ahora. Esto se debe en parte a la falta de recursos humanos especializados en analizar integralmente esta información. Mas aún, la rigurosidad con la que se describe cada muestra subida a un repositorio público dista de ser óptima. En este trabajo categorizamos cada muestra por su estrés específico y tratamiento, y demostramos que se pueden utilizar técnicas de aprendizaje automático modernas para analizarlas, alcanzando resultados consistentes con la literatura existente.

**Palabras claves:** Genómica, Soja, Aprendizaje automático, Datos abiertos, Estrés, RNA.

# IDENTIFYING GENETIC BIOMARKERS IN GLYCINE MAX (SOY) ASSOCIATED TO DIFFERENT STRESSES USING OPEN RNA-SEQ DATA

Soy is one of the most important crops worldwide. In Argentina, it plays a key role in the export economy, being the third largest producer and main exporter of soy oil and soy meal. However, its production faces important environmental challenges.

Since the 2000s, owing to the usage of NGS (Next Generation Sequencing) technologies, the volume of available genomics data for soy grew exponentially. Its widespread usage in comparative and genetic variability studies has helped identify biomarkers associated to the resistance of environmental stresses.

Despite the open nature of most of this data, they have not been explored as a whole until now. This is in part due to the lack of specialized human resources capable of exploring this information integrally. Moreover, the rigurosity with which each sample is described is far from optimal. In this work we categorize each sample according to its specific stress and treatment, and demonstrate that it is possible to use modern machine learning techniques to analyze them, reaching results that are consistent with the existing literature.

**Keywords:** Genomics, Soy, Machine learning, Open data, Stress, RNA.

# AGRADECIMIENTOS

Agradecido con mi familia por acompañarme en la decisión de estudiar Ciencias de Datos, una carrera que al inscribirme aún no tenía graduados. Les agradezco mucho incentivar mi curiosidad en la ciencia desde que tengo memoria.

Para la Universidad de Buenos Aires también tengo agradecimientos; los excelentes profesores y dirección, por diseñar y gestionar una carrera que al momento de recibirme rompe récords de inscripciones.

Entre estos inscriptos muchos de mis compañeros, amigos y alumnos, a cada uno de ellos les agradezco por contribuir un punto de vista diferente, algo fundamental en la formación de todo científico.

Agradezco a mis directores de tesis Maxi y Diego por ayudarme a armar el rompecabezas enorme de datos y bioinformática. A Fede de APOLO Biotech por ponernos en contacto y a Manu, por ocuparse de la generación de anotaciones con LLMs.

*La verdadera tesis son los modelos que entrenamos en el camino.*

# Índice general

# 1. INTRODUCCIÓN

## 1.1. Motivación

Identificar biomarcadores, entendidos como genes con perfiles de expresión asociados a factores de estrés bióticos y abióticos, permite entender los procesos que originan respuestas a los estreses, diseñar estrategias de manejo de cultivos e incluso generar nuevas variedades de plantas capaces de prosperar en entornos adversos o de ofrecer un mayor rendimiento nutricional.

La relevancia de esta investigación se magnifica al considerar el rol fundamental de la soja en la economía argentina, donde representa aproximadamente el 30 % de las exportaciones totales del país. Este cultivo no solo es crucial para el sector agroindustrial nacional, sino que también posiciona a Argentina como un actor estratégico en la seguridad alimentaria global. La optimización de su producción mediante la identificación de biomarcadores de resistencia resulta particularmente significativa dado que gran parte de las áreas de cultivo se encuentran expuestas a diversos factores de estrés ambiental, como sequías, salinidad y enfermedades patógenas. La mejora en la resistencia de la soja a estas condiciones adversas tendría un impacto directo en la sostenibilidad del sector agrícola argentino.

El proceso típico de investigación de biomarcadores consiste en tres partes fundamentales:

- Se cultiva la planta en condiciones de control y estrés.
- Se secuencian muestras de la planta en laboratorios (muchas veces tercerizados).
- Se comparan los datos transcriptómicos de las plantas utilizando métodos estadísticos.

En los últimos años, la secuenciación (generación de información genética) se ha vuelto mucho más accesible en precio y tiempo gracias a avances tecnológicos conocidos como NGS (*Next Generation Sequencing*). Las técnicas de machine learning gozaron avances muy parecidos, gracias al abaratamiento del poder de cómputo, que permite entrenar modelos más sofisticados con un presupuesto y tiempo menores.

A pesar de estos avances en ambos campos, el costo de secuenciación sigue siendo el principal cuello de botella a la hora de implementar machine learning moderno en la identificación de biomarcadores. Tener un tamaño de población mayor permitiría al investigador acceder a mejores modelos estadísticos, mejorar su eficacia e identificar más biomarcadores.

Es por esto que aprovechar los datos genéticos ya disponibles plantea un modelo de investigación *en seco* completamente distinto, donde el enfoque no está en cultivar plantas para cada experimento, sino que agregar los datos ya existentes. Los puntos principales del nuevo paradigma de investigación bioinformática son:

- Normalizar los datos de diversas fuentes en lugar de generarlos (haciendo la búsqueda de biomarcadores más económica).
- Generar anotaciones automáticamente usando sus metadatos (ahorrando tiempo).
- Usar modelos avanzados de machine learning para analizar los datos.

## 1.2. Objetivos

En este trabajo, se busca identificar biomarcadores genéticos usando técnicas de machine learning diferentes a las tradicionales. A este fin, es necesario recopilar la mayor cantidad de datos genómicos de soja posible: implementamos una suite de herramientas que interactúan con las APIs de los principales repositorios de datos genómicos (SRA, GEO, etc.) para crear una base de datos de miles de muestras de plantas de soja.

También es central la curación y consolidación de metadatos de cada muestra, muchas veces fragmentados en las distintas plataformas. Además de la información genómica, es necesario poseer información sobre las condiciones de crecimiento y tratamiento de cada una de las muestras. Los modelos de machine learning van a intentar predecir estas condiciones basandose únicamente en la información genética.

El objetivo de un trabajo de modelado con machine learning suele ser producir un modelo con un alto poder predictivo. En este caso, si bien se utilizan métricas de exactitud como en aquellos trabajos, se busca explicar usando las variables que el modelo selecciona los genes e interacciones responsables de la resistencia a estreses. En estadística esto se conoce como inferencia.

## 1.3. Trabajos previos

Utilizar más y mejores técnicas para analizar las expresiones genómicas es una temática central de la biología molecular. En los últimos años, el machine learning emergió como una herramienta poderosa para el análisis de datos genómicos, ofreciendo nuevas perspectivas más allá de los métodos estadísticos tradicionales. Los métodos clásicos para identificar biomarcadores genéticos en soja típicamente se han basado en estudios de asociación del genoma completo (GWAS) y análisis de expresión diferencial. Depeng et al. [1] realizaron un extenso estudio utilizando GWAS para identificar loci asociados con la resistencia a la sequía en soja, estableciendo un punto de referencia para los métodos tradicionales.

Sin embargo, las limitaciones de estos enfoques han llevado a la exploración de técnicas de machine learning. Nazari et al. [2] realizaron un meta-análisis con inteligencia artificial para identificar fenotipos de resistencia a partir de datos transcriptómicos, logrando identificar nuevos genes candidatos que los métodos estadísticos tradicionales no detectaron. De manera similar, Zhou et al. [3] aplicaron Random Forests para diferenciar infecciones de Covid y *Mycoplasma pneumoniae*.

Un trabajo particularmente relevante es el de Venancio et al. [4], quienes desarrollaron una pipeline integrada que combina datos genómicos de múltiples fuentes públicas para expresión diferencial por tejidos. Su análisis destaca la importancia de la integración de datos, aunque se limitaron a métodos estadísticos convencionales. En cuanto a la curación y consolidación de datos genómicos, Brancato et al. [5] crearon un framework para la normalización y estandarización de metadatos de expresión génica provenientes de diferentes repositorios públicos.

# 2. DATOS

## 2.1. Recopilación

### 2.1.1. Expresión genómica

Existen muchas formas de generar y almacenar información genética. En este trabajo utilizamos únicamente datos de secuenciación de ARNm también conocidos como *RNA-Seq*. Estos suelen tener formato `.fastq` y consisten de lecturas individuales de pequeños pedazos de ARN. Para realizar el análisis buscamos construir una matriz de expresión $K^{n \times m}$, donde $n$ es el número de muestras, $m$ el número de genes y $K_{ij}$ la cantidad de veces que la muestra $i$ expresó el gen $j$.

Es imposible armar una matriz de expresión sin antes alinear las lecturas individuales. Este proceso, así como la maquinaria utilizada para secuenciar el transcriptoma y la manipulación de cada muestra, pueden introducir sesgos en la matriz final. Para mitigar esto, nos basamos en las matrices de expresión crudas generadas en Soy Atlas [4]. El mismo contiene más de 5000 muestras de soja de 60 proyectos de distintas temáticas, y la matriz de expresión generada es de $5376 \times 52837$ (muestras×genes).

| **Proyecto** | **Muestras** | **Título** |
|---|---|---|
| PRJNA706999 | 214 | Comparing early transcriptomic responses of 18 soybean (Glycine max) genotypes to iron stress |
| PRJNA389558 | 102 | Transcriptomes of 102 soybean accessions |
| PRJNA514200 | 47 | Characterization of interaction between soybean cyst nematode and soybean aphids in soybean |
| PRJNA544698 | 22 | Dynamic Gene Expression Changes in Response to Micronutrient, Macronutrient, and Multiple Stress Exposures in Soybean |
| PRJNA564957 | 20 | A Transcriptional Regulatory Network of Rsv3-mediated Extreme Resistance Against Soybean Mosaic Virus |
| PRJNA615913 | 8 | Differential gene expression in response to water deficit in leaf and root tissues of soybean genotypes with contrasting tolerance profiles |
| ... | | |

*Tab. 2.1:* Algunos proyectos usados para construir las matrices de expresión

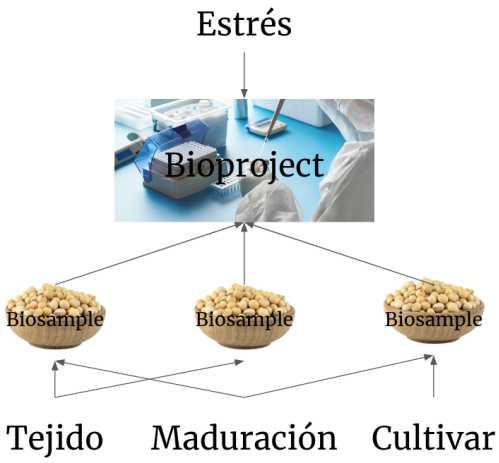
Para que la matriz sea compatible con los métodos de análisis tradicionales, eliminamos: los genes con varianza cero, los proyectos con menos de dos muestras y los estreses con menos de 10 muestras. Las dimensiones finales de $K$ son $4723 \times 51575$.

### 2.1.2. Metadatos

Los metadatos se extrajeron de la página web del *National Center for Biotechnology Informatics* (NCBI), pero existen en tres dimensiones [6]: *BioProject* (proyecto) es el colectivo de muestras que se utilizan en una publicación científica determinada y *BioSample*

(muestra) es una extracción de un tejido de una planta de soja que tiene una o más *Run* (corridas) asociadas, que son secuenciaciones. A veces, se secuencia una muestra más de una vez para asegurar la calidad de la data.

Para nuestro análisis juntamos toda la metadata posible, lo que implica interactuar con las tres plataformas y unir los datos usando sus identificadores. Esto implica definir una ontología donde cada atributo posible pertenece a sólo una dimensión. El conjunto de datos final contiene muestras y sus metadatos, sumados a todos los atributos de su proyecto y sus respectivas corridas.




*(a)* Ontología


```
{
    "acc": "SRR391537",
    "experiment": "SRX112518",
    "sample_name": "Soy 4",
    "sample_acc": "SRS282441",
    "sra_study": "SRP009826",
    "library_name": "Leaf tissue (elevated)",
    "primary_search": "['765513', 'Legume Comparison', 'SAMN00765513'
    "title_x": "Comparing the response to chronic ozone of five agric
    "BioSample": "SAMN00765513",
    "Sample name": "Soy 4",
    "SRA": "SRS282441",
    "name": "Key Laboratory of Biotechnology in Plant Protection of M
    "title_y": "The RNA-seq of Cavariella aegopodii",
    "description": "The specimens of the willow-carrot aphid utilized
},
{
    "acc": "SRR391538",
    "experiment": "SRX112520",
    "sample_name": "Soy 5",
    "sample_acc": "SRS282442",
    "sra_study": "SRP009826",
    "library_name": "Leaf tissue (elevated)",
    "primary_search": "['765514', 'Legume Comparison', 'SAMN00765514'
    "title_x": "Comparing the response to chronic ozone of five agric
    "BioSample": "SAMN00765514",
    "Sample name": "Soy 5",
    "SRA": "SRS282442",
    "name": "Key Laboratory of Biotechnology in Plant Protection of M
    "title_y": "The RNA-seq of Cavariella aegopodii",
    "description": "The specimens of the willow-carrot aphid utilized
},
```


*(b)* Metadata

*Fig. 2.1:* Organización de la metadata

## 2.2. Anotación

El punto de recopilar toda la metadata disponible al nivel de muestra es anotar el estrés, tratamiento y tejido de cada una de ellas. En particular, el estrés es la variable dependiente que nuestros modelos tradicionales y alternativos van a intentar predecir, basándose en las expresiones genómicas de cada una.

Para crear las anotaciones, se uso el modelo Gemini 1.0 Pro con la metadata completa como entrada y tres prompts:

> 1. What is the tissue used in this sample? Answer with one word indicating the tissue (tissue clases: leaf, seed, crown, shoot, stem, hypocotyl, pod, root, cotyledon, petiole, radicle, root, ovule, ovary, embryo).
>
> 2. What is the stressing agent? (stressing_agent classes: bacteria, fungus, virus, sporus, parasite, flood, drought, nutrient deficit, nutrient poisoning, acid, mineral deficiency, mineral poisoning, cold, heat, senescense, transgenic, none).
>
> 3. Is this biosample a control or a treatment? Answer with a concise word indicating the group class (group classes: control, treatment)

### 2.2.1. Estrés

El estrés no existe a nivel de muestra, sino que de proyecto como en Fig. 2.1a. No obstante, al estar vinculada cada muestra a un proyecto, podemos heredarle esta información. Ejemplos de campos que contribuyen a la anotación del estrés específico son: abstract, título del proyecto, nombre de la muestra.

Además, el estrés no es un campo predeterminado que los investigadores pueden llenar al momento de cargar sus datos, sino que es atributo que anotamos. Al ser un campo heredado del proyecto, no distingue entre muestras de control y tratadas, todas las muestras de un proyecto se consideran estresadas en esta anotación.

| | flw | hyp | leaf | nod | pod | rad | root | seed | sdl | sht | stm |
|---|---|---|---|---|---|---|---|---|---|---|---|
| acid | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 3 | 0 | 0 |
| bacteria | 0 | 0 | 21 | 19 | 0 | 0 | 30 | 0 | 0 | 0 | 0 |
| cold | 0 | 0 | 6 | 0 | 0 | 0 | 4 | 4 | 0 | 4 | 0 |
| control | 31 | 43 | 683 | 10 | 55 | 3 | 390 | 410 | 25 | 94 | 130 |
| drought | 0 | 0 | 131 | 0 | 0 | 0 | 42 | 9 | 3 | 21 | 1 |
| fungus | 0 | 11 | 25 | 0 | 30 | 0 | 91 | 7 | 31 | 0 | 128 |
| heat | 7 | 0 | 4 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 |
| mineral | 0 | 3 | 2 | 0 | 0 | 0 | 35 | 0 | 0 | 0 | 0 |
| nutrient | 0 | 0 | 119 | 15 | 0 | 0 | 172 | 87 | 0 | 0 | 0 |
| other | 4 | 0 | 48 | 0 | 4 | 0 | 6 | 4 | 0 | 0 | 0 |
| parasite | 0 | 0 | 82 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 3 |
| senescence | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| transgenic | 2 | 4 | 86 | 0 | 8 | 4 | 41 | 274 | 0 | 6 | 5 |
| virus | 0 | 0 | 172 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Tab. 2.2:* Distribución de estreses por tejido.

### 2.2.2. Tratamiento

Una muestra puede estar tratada o no. Una muestra tratada es una planta afectada por alguno de los estreses específicos de la sección anterior. Aquellas muestras no tratadas se consideran de control, y se espera que hayan crecido en condiciones saludables. Existen muestras cuyo estado de tratamiento es desconocido.

### 2.2.3. Tejido

El tejido de origen de la muestra también es fundamental en el análisis de expresión génica, ya que la expresión varía significativamente entre diferentes partes de la planta. Esto se debe a que cada tejido cumple funciones especializadas y, por lo tanto, requiere la activación de distintos conjuntos de genes. Por ejemplo, los genes relacionados con la fotosíntesis se expresarán principalmente en tejidos verdes como hojas, mientras que los genes involucrados en la absorción de nutrientes tendrán mayor actividad en las raíces. Esta especialización tisular también influye en cómo la planta responde a diferentes tipos de estrés, por lo que la selección del tejido a secuenciar depende tanto del estrés estudiado como de las preguntas biológicas que se busca responder.

## 2.3. Normalización

Para minimizar el impacto de los sesgos inducidos por la maquinaria utilizada, manipulación de muestras, reactivos, etc. se emplean técnicas de remoción de efecto de lote o *batch effect removal* [7]. Este paso es esencial en cualquier proyecto de bioinformática, incluso cuando el dataset entero es producido por el mismo laboratorio que realiza la investigación. El objetivo es modelar el impacto de las condiciones de crecimiento y tratamiento de la planta en la expresión génica, un impacto que muchas veces es pequeño y por ende se debe prestar especial atención a cualquier ruido que tenga el dataset, por más insignificante que parezca.

Los métodos tradicionales, como DESeq2 y edgeR consideran a la normalización como parte del análisis, usando las matrices crudas directamente. Por más que esto sea así, la amplia mayoría de las investigaciones tienen un paso de normalización separado para crear visualizaciones y reportar métricas como distancias medias entre categorías. En este trabajo, la normalización se realiza antes de cada experimento por separado, únicamente sobre los datos que van a ser utilizados, para mitigar la fuga de datos.

### 2.3.1. Estandarización

Una de las formas de lograr esto es tratar cada gen como una variable independiente, y estandarizarlo restando la media y dividiendo por la varianza. La ventaja es que se obtiene un conjunto de datos centrado en cero y es computacionalmente barato, pero la remoción del sesgo por lote no es tan efectiva. En Fig. 2.2 se aprecia el efecto confusor de la variable proyecto, viendo la primer fila (reducción de dimensionalidad sobre matriz de expresión cruda) parece que las muestras están bien separadas por tejido y estrés (segunda y tercer columna). Sin embargo, esta relación no es causal, porque esta influida por el proyecto.

### 2.3.2. ComBat

ComBat modela cada gen de la matriz de expresión como una binomial negativa, y hace regresión sobre los parámetros de esta distribución con un enfoque empírico sobre los mismos. Al ser una herramienta específica para RNA-Seq logra resultados mejores que el resto de las técnicas cuando se evalúa la preservación de los efectos genómicos originales [8]. Como se observa en la Fig. 2.2, la normalización por ComBat logra reducir el efecto por lote mientras preserva las diferencias biológicamente relevantes entre condiciones experimentales.

### 2.3.3. TMM

El método TMM (*Trimmed Mean of M-values*) fue desarrollado específicamente para datos de RNA-Seq. Se basa en la suposición de que la mayoría de los genes no están diferencialmente expresados entre muestras. El procedimiento calcula factores de normalización entre pares de muestras, donde una se toma como referencia, típicamente aquella cuyo recuento total está más cerca de la media. La ventaja principal de TMM es su robustez ante la presencia de genes altamente expresados o con expresión diferencial extrema, ya que estos son excluidos del cálculo de los factores de normalización. Esto es particularmente importante en estudios de estrés [9].

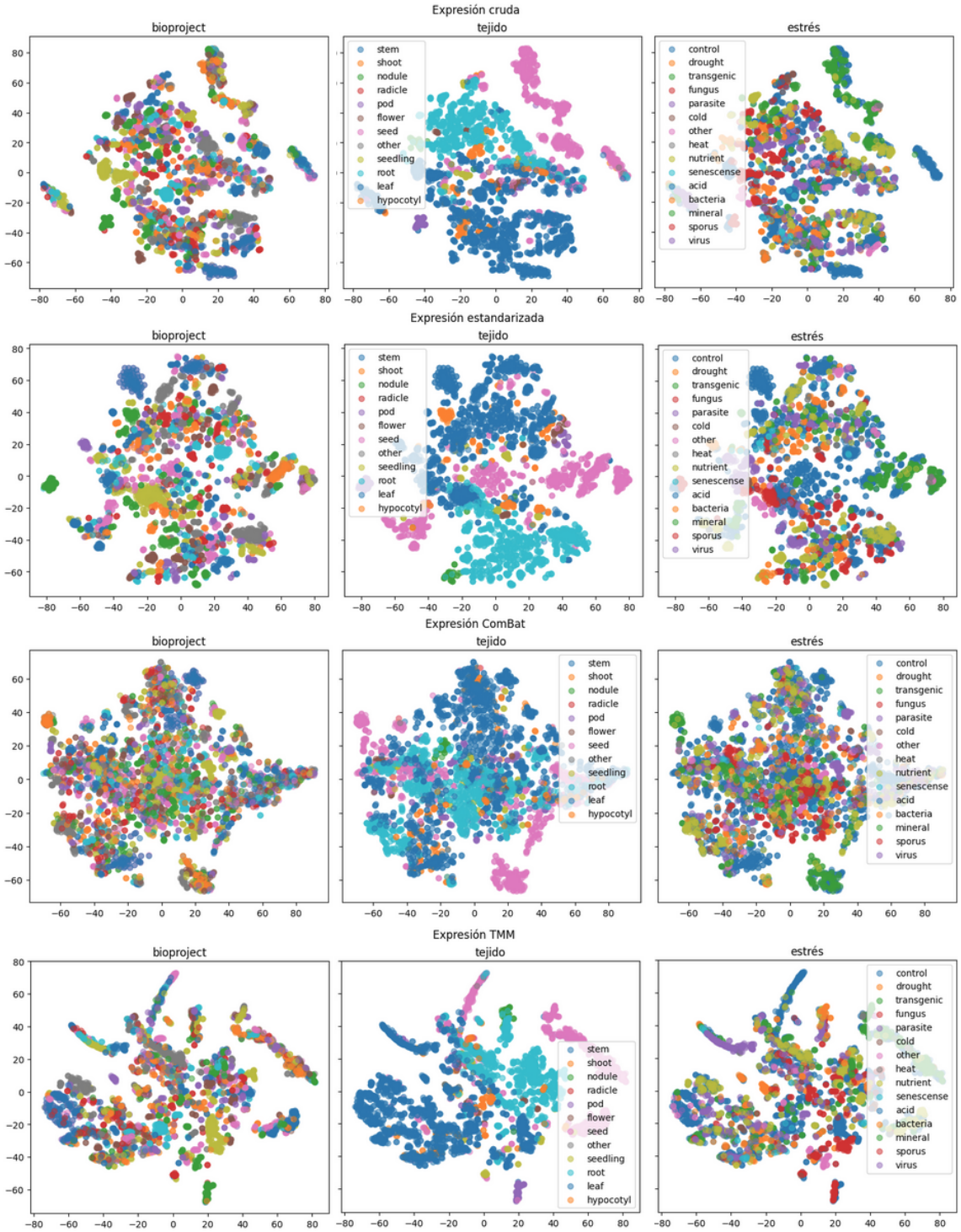



*Fig. 2.2:* t-SNEs de matrices de expresión normalizadas

# 3. MODELOS

El problema de predicción de estrés puede definirse en términos matemáticos de forma simple:

$$\min_f l(Y, f(K)) \qquad K \in \mathbf{R}^{n \times m}, \quad Y \in \text{estreses}^n$$

Con $l$ una función de pérdida. Pero en nuestro modelado, lo que realmente nos interesa es encontrar los genes cuya expresión sea indicadora de algún tipo de estrés. Formalmente, buscamos una función $h$ que mapee un modelo entrenado $f$ y un gen $i$ a una medida de importancia:

$$h : f, g_i \to \text{imp}_i \in [0, 1]$$

El criterio para elegir la función de importancia $h$ es justamente lo que cambia de modelo a modelo. En el esquema tradicional, $f$ no es una función de predicción entonces $imp_i$ se calcula directamente con los datos del gen $g_i$. Para los modelos alternativos, usamos medidas como importancia de permutación o reducción de entropía/impureza

## 3.1. Tradicionales

### 3.1.1. DESeq2

DESeq2 es un método univariado muy establecido para el análisis de expresión diferencial que asume que las expresiones se distribuyen como binomiales negativas y usa métodos bayesianos para mejorar su rendimiento en pocas muestras, lo que hace la inclusión de este modelo interesante, ya que contamos con un conjunto de datos grande [10].

El modelo asume que los conteos $K_{ij}$ para el gen $i$ en la muestra $j$ siguen una distribución binomial negativa con media $\mu_{ij}$ y dispersión $\alpha_i$. La media se modela como:

$$\mu_{ij} = s_j q_{ij}$$

donde $s_j$ es el factor de normalización por tamaño para la muestra $j$, y $q_{ij}$ es proporcional al nivel de expresión real del gen en la condición experimental. En DESeq2, la importancia se modela en base al *log fold change* y los p-valores de descartar, con prueba de Wald, la hipótesis nula de que el gen no varía su expresión frente al estrés.

### 3.1.2. EdgeR

EdgeR es otro método ampliamente utilizado que, al igual que DESeq2, se basa en la distribución binomial negativa. Sin embargo, difiere en su enfoque para la estimación de la dispersión. EdgeR utiliza un método de máxima verosimilitud ponderada para estimar la dispersión común entre genes, y luego modera las estimaciones específicas de cada gen hacia esta dispersión común en forma iterativa [11].

El método implementa la normalización *Trimmed Mean of M-values* (TMM) para corregir las diferencias en la composición de las bibliotecas. La dispersión se modela como:

$$\phi_i = \phi_0 + \phi_{g_i}$$

donde $\phi_0$ es la dispersión común y $\phi_{g_i}$ es la dispersión específica del gen $g_i$.

## 3.2. Alternativos

### 3.2.1. Random forest

En random forest se construyen muchos árboles de decisión pero en cada split se permite usar solamente un subconjunto aleatorio de $p$ genes. El beneficio de este enfoque es que cada árbol esta decorrelacionado del resto, son predictores débiles individualmente pero en conjunto capturan patrones incluso en ambientes de muy alta dimensionalidad. Esta heurística para reducir la varianza y evitar el sobreajuste se conoce como *bagging* [12] y es muy relevante en problemas de genómica donde $m \gg n$

Los árboles aleatorios son de los modelos de machine learning más accesibles y utilizados en genómica, pero generalmente con poblaciones experimentales pequeñas. Para nuestro dataset, tuvimos mejores resultados con 500 árboles de altura máxima 30 y $p = \sqrt{m}$. Estos parámetros se buscaron con una búsqueda aleatoria con validación cruzada en 5 pliegues. La métrica utilizada fue el área bajo la curva ROC "uno contra todos", ya que es una clasificación multilabel [13].

### 3.2.2. Gradient boosted trees

Gradient boosted trees entrena un árbol pero de forma iterativa. Para esto modela los residuales, o el error, de cada instancia de entrenamiento. Este enfoque secuencial permite que cada nuevo árbol se especialice en las muestras más difíciles de clasificar, lo que resulta en modelos propensos al sobreajuste. En este problema de alta dimensionalidad, donde la cantidad de features supera por órdenes de magnitud a la cantidad de observaciones, un modelo que tiende a especializarse suele tener mejor rendimiento.

Para nuestro conjunto de datos, los mejores resultados se obtuvieron con 1000 árboles de altura máxima 6 y una tasa de aprendizaje de 0.01. El número reducido de niveles por árbol, comparado con random forest, es una estrategia común en boosting para controlar la complejidad del modelo. También implementamos early stopping monitoreando el rendimiento en un conjunto de validación, deteniendo el entrenamiento cuando no se observa mejora durante 50 iteraciones consecutivas.

### 3.2.3. Red neuronal

Las redes neuronales representan una aproximación fundamentalmente diferente al problema de expresión diferencial. Para nuestro análisis, implementamos una arquitectura feed-forward simple con dos capas ocultas, utilizando activación ReLU y dropout (0.3) para prevenir el sobreajuste e incentivar la exploración de genes. La capa de entrada tiene dimensión $p$ (número total de genes) y la de salida utiliza activación softmax para la clasificación multiclase.

Para manejar la alta dimensionalidad (característica de los datos de expresión génica), se intenta añadir una capa de reducción de dimensionalidad antes de la clasificación. Esto comprime las variables de entrada en un espacio latente de menor dimensionalidad.

El entrenamiento se realizó utilizando el optimizador Adam con una tasa de aprendizaje inicial de 1e-4 y programación de tasa de aprendizaje cíclica para evitar mínimos locales. Para abordar el desbalance de clases, implementamos ponderación de clases inversamente proporcional a su frecuencia en los datos de entrenamiento.

# 4. EXPERIMENTOS

| | Experimento 1 | Experimento 2 | Experimento 3 | Experimento 4 |
|---|---|---|---|---|
| Muestras | 214 | 507 | 3748 | 4302 |
| Estreses | 1 | 1 | 2 | 12 |
| Tejidos | 2 | 7 | 11 | 11 |
| Proyectos | 1 | 25 | 127 | 139 |
| Control | 50 % | 42 % | 33 % | 47 % |

*Tab. 4.1:* Diseño de cada experimento

Para analizar la utilidad de cada modelo en distintos entornos, diseñamos cuatro experimentos basados en situaciones típicas de investigación bioinformática (Tab. 4.1). El primer experimento toma un solo proyecto, y se parece a la forma más ortodoxa de hacer un análisis de expresión génica donde todas las muestras vienen del mismo laboratorio fuente y estrés. El segundo experimento también analiza un único estrés, pero para aumentar el número de muestras integramos datos de varios proyectos que comparten un estrés. En el tercer experimento usamos todos los proyectos posibles, pero en lugar de centrarnos en un estrés en particular los subclasificamos en biótico, abiótico o control. Finalmente, juntamos todos los proyectos, estreses y tejidos en un único conjunto de datos para el experimento cuatro.

## 4.1. Clasificar tratamiento en un solo proyecto

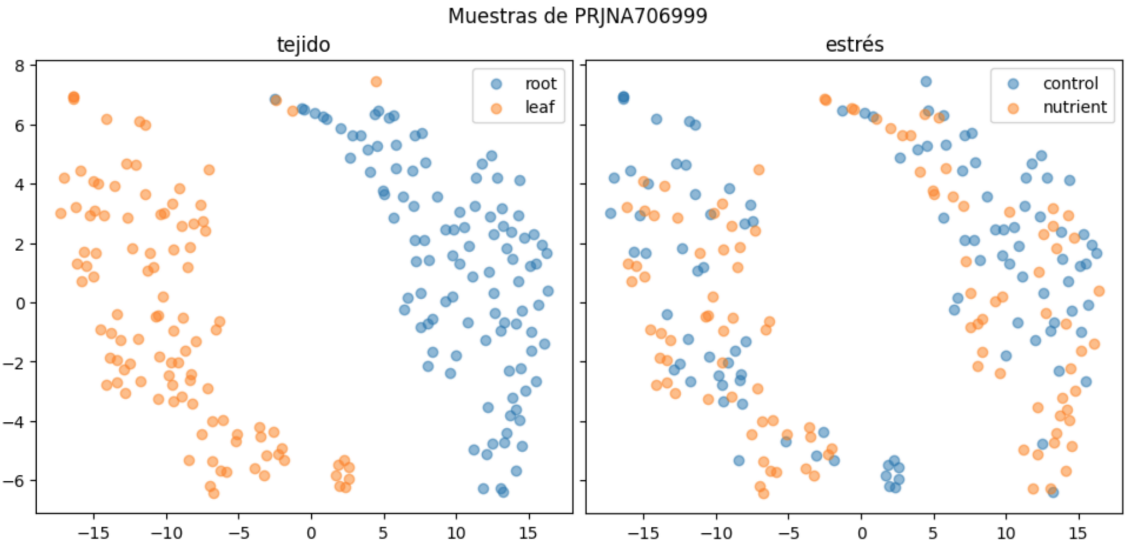



*Fig. 4.1:* t-SNE de expresión génica sobre muestras del proyecto seleccionado.

En este experimento, entrenamos los modelos descriptos en Sec. 3 usando las muestras del proyecto PRJNA706999 [14]. El mismo cuenta con 214 muestras de hojas y raíces

germinadas en condiciones de control y déficit de hierro en cantidades iguales (estrés anotado en nuestro dataset como *mineral*). Debido al tamaño del conjunto de datos, no se entrenó la red neuronal.

Para los modelos alternativos, buscamos que clasifiquen la variable binaria tratamiento explicada en Sec. 2.2.2. Es decir, si la muestra fue germinada con déficit de hierro o no. El único modelo alternativo que alcanzó resultados significativos en el conjunto de testeo fue el random forest, con exactitud del 92,5 % y $F_1$-score 0,928.

De los modelos tradicionales, DESeq2 reportó valores de expresión diferencial mayores que EdgeR para genes típicamente asociados a respuestas de estrés de origen mineral.

## 4.2. Clasificar tratamiento en un solo estrés

Para este experimento usamos todos los proyectos y tejidos cuyo estrés anotado sea *fungus* (Tab. 2.2), elegido por tener la mayor cantidad (321 tratadas y sus respectivos 236 controles por proyecto) y variedad de muestras (8 tejidos en 25 proyectos). Nuevamente entrenamos todos los modelos de la sección anterior sobre la variable binaria tratamiento.

Usamos ComBat (Sec. 2.3.2) para quitar el efecto del lote. A diferencia del experimento anterior, el mejor modelo para este experimento fue gradient boosted trees, con exactitud del 94,4 % y $F_1$-score 0,954. EdgeR y DESeq2 hallaron genes estadísticamente significantes pero ninguno asociado a estrés fúngico en específico.

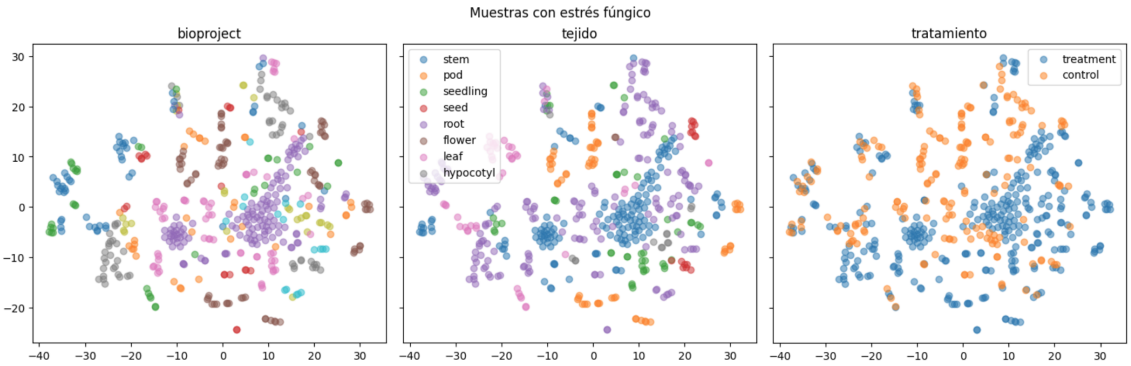



*Fig. 4.2:* t-SNE de expresión génica sobre muestras del estrés seleccionado.

## 4.3. Diferenciar estrés biótico o abiótico

En este experimento se usa el dataset completo de tejidos y estreses con todos los proyectos (Fig. 4.3). Sin embargo, la variable clasificada no es "tratamiento", sino que "biótico". Esta variable indica si la muestra sufrió un estrés biótico, abiótico, o es simplemente de control. Como el volumen de muestras es mayor: 3748, se pudo entrenar, además de los modelos de las secciones anteriores, una red neuronal.

En esta ocasión, la red neuronal tuvo el mejor desempeño con exactitud del 89,7 % y $F_1$-score de 0,912. El random forest y gradient boosted trees tuvieron desempeños parecidos entre sí pero inferiores a la red neuronal. DESeq2 y EdgeR no dieron resultados significativos, en gran parte debido a que esta tarea, si bien es de clasificación binaria, integra muestras de proyectos muy variados en cuanto a tejido y estrés (Tab. 5.1).

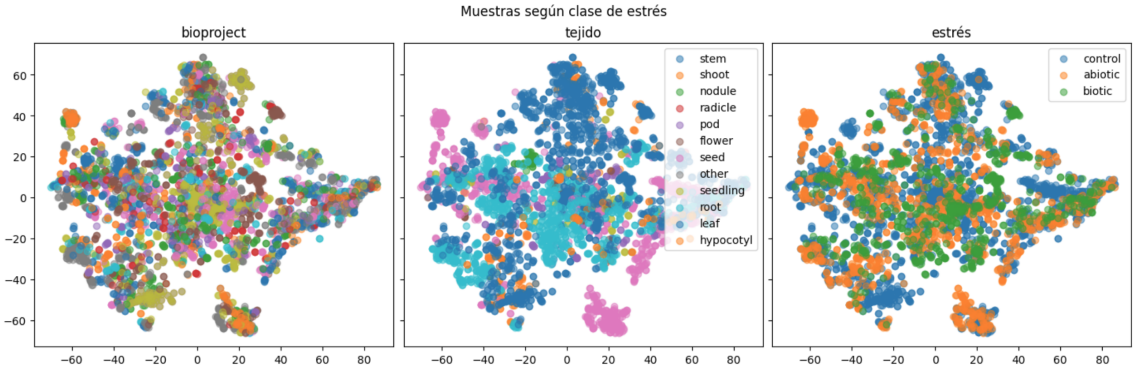



*Fig. 4.3:* t-SNE de expresión génica según clase de estrés.

## 4.4. Clasificar todos los estreses

Finalmente, usamos el conjunto de datos completo con la variable estrés (Fig. 4.4). Nuevamente entrenamos todos los modelos para hacer la comparativa, pero usando todas las categorías posibles de estrés. Al tratarse de un problema de clasificación en muchas clases, resulta más desafiante que los experimentos anteriores.

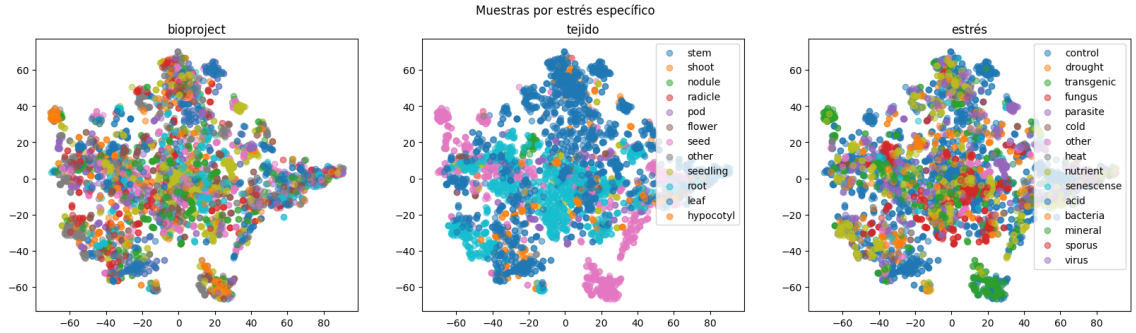



*Fig. 4.4:* t-SNE de expresión génica sobre estrés específico.

# 5. COMPARATIVA

## 5.1. Ajuste

| | Exp. 1 | | | Exp. 2 | | | Exp. 3 | | | Exp. 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Modelo** | $F_1$ | Exac. | T(m) | $F_1$ | Exac. | T(m) | $F_1$ | Exac. | T(m) | $F_1$ | Exac. | T(m) |
| Random forest | 0.928 | 92.5 % | 3 | 0.833 | 79.5 % | 10 | 0.814 | 82.07 % | 52 | 0.491 | 66.3 % | 78 |
| GBT | 0.813 | 79.6 % | 4 | 0.954 | 94.4 % | 12 | 0.743 | 75.5 % | 48 | 0.439 | 61.8 % | 79 |
| Red neuronal | - | - | - | - | - | - | 0.912 | 89.7 % | 123 | 0.823 | 81.2 % | 164 |

*Tab. 5.1:* Comparativa de rendimiento por modelo y experimento. *GBT: Gradient boosted trees, Exac.: Exactitud, T(m): Tiempo de entrenamiento en minutos.*

Para los modelos alternativos, la red neuronal fue consistentemente mejor que el resto en toda métrica. Sin embargo, precisa de una cantidad de datos mucho más elevada y típicamente imposible de producir en cualquier investigación individual. Sumado a esto, el tiempo de cómputo, tuneado y entrenamiento necesario la vuelven bastante menos accesible que los métodos menos complejos (Tab. 5.1).

Para escenarios simples, como los experimentos 1 (Sec. 4.1) y 2 (Sec. 4.2), no parece viable un modelo que supere a random forest y gradient boosted trees. En estos casos, alcanzan métricas de exactitud igualadas a la de un curador humano experto. Su rendimiento cae mucho en los escenarios más complejos, en estos el espacio de entrada es no solo altamente dimensional sino que altamente variable. En particular, cuando la clasificación deja de ser binaria e incluso se torna imbalanceada como en el experimento 4 (Sec. 4.4), los modelos no llegan a un rendimiento útil.

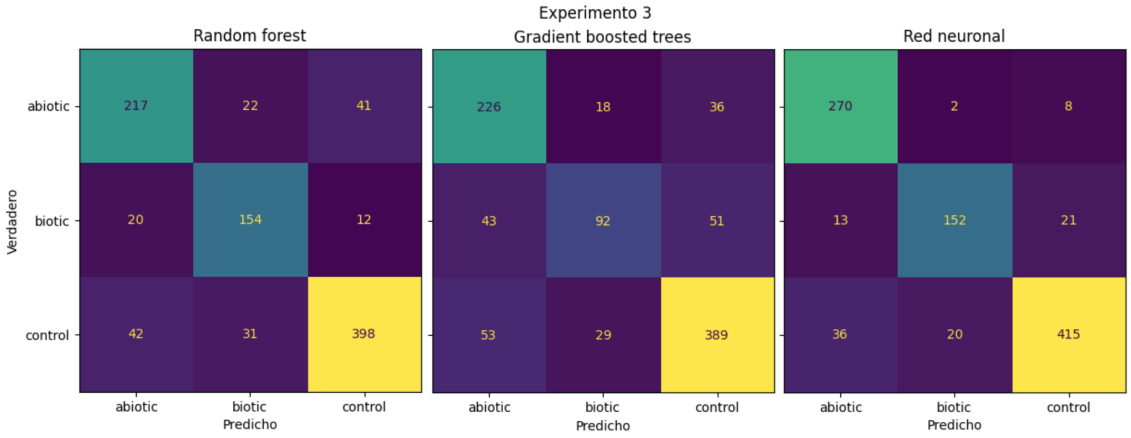



*Fig. 5.1:* Matrices de confusión por modelo para el experimento 3.

### 5.1.1. Selección de genes

Ningun análisis arrojó genes en común comparando los métodos tradicionales frente a los alternativos. Sí se observaron algunos genes en común al comparar los gradient boosted trees y random forest, probablemente por ser ambos modelos basados en árboles

de decisión. Sumado a esto, los métodos alternativos son no-determinísticos, por lo tanto seleccionan genes distintos en cada ejecución.

Los más parecidos fueron los genes seleccionados por EdgeR y DESeq2, ambos métodos tradicionales basados en estadística y semi-determinísticos (EdgeR tiene un enfoque iterativo y puede llegar a arrojar distintos resultados, pero en general es consistente).

En términos de cantidad de genes relevantes hubo más consenso entre los modelos, aún entre los tradicionales y alternativos. En general, solo los primeros 100 a 1000 genes son tomados en cuenta por los modelos a la hora de clasificar. Esto se condice con la literatura existente sobre soja.

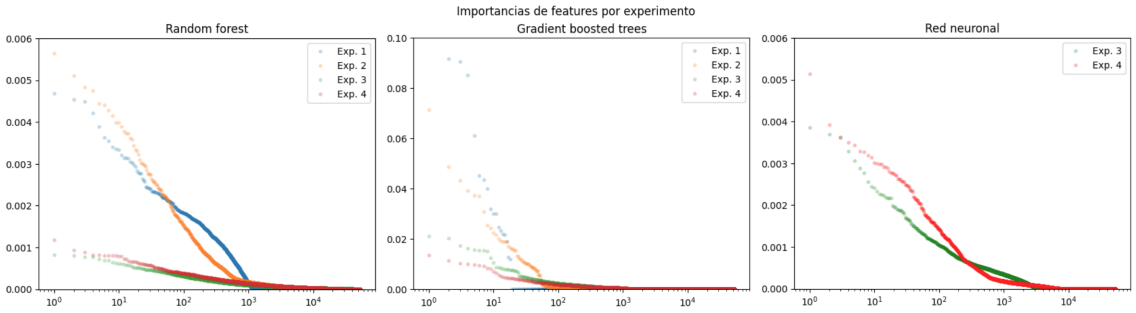



*Fig. 5.2:* Genes ordenados según importancia por experimento y modelo.

# 6. DISCUSIÓN

En este trabajo, se demostró la viabilidad técnica de integrar datos genómicos de soja provenientes de múltiples repositorios mediante herramientas automatizadas para la recopilación, curación y anotación de metadatos (Sec. 2). Este enfoque no solo permite un acceso más eficiente a la información ya disponible, sino que también establece las bases para explorar nuevas metodologías en bioinformática, como el uso de machine learning para la identificación de biomarcadores.

Un tema central en la preparación de datos fue la remoción del efecto de lote, un desafío conocido en estudios que integran datos de diversas fuentes. Si bien es un campo con avances significativos en los últimos años, sigue siendo objeto de discusión en la comunidad científica. Los resultados obtenidos con ComBat destacan su capacidad para preservar las diferencias biológicamente relevantes entre muestras de control y tratadas, minimizando simultáneamente el impacto del origen de los datos (instrumento, proyecto, tejido, entre otros). Esto es evidente al comparar las visualizaciones de reducción dimensional de experimentos con un solo proyecto (Fig. 4.1) y con varios proyectos de un mismo estrés (Fig. 4.2). Además, técnicas más simples, como la estandarización por gen, también demostraron ser útiles en ciertos contextos, como se observa en la (Fig. 2.2).

La diversidad de muestras disponibles, tanto en términos de estrés como de tejido, permitió realizar análisis específicos y explorar modelos alternativos de machine learning. Por ejemplo, el rendimiento superior de los métodos basados en bagging, como random forest, refleja su capacidad para segmentar el espacio de entrada en regiones manejables, lo que resulta particularmente útil en datasets con alta heterogeneidad, como el utilizado en este trabajo. Este enfoque contrasta con los métodos tradicionales univariados (DESeq2 y EdgeR), que tienden a modelar patrones globales, mostrando limitaciones en tareas complejas o altamente dimensionales. Los resultados de métricas como exactitud y F1-score (Tab. 5.1) respaldan esta observación, destacando la adaptabilidad de los modelos basados en árboles para manejar datos de múltiples fuentes y condiciones.

Por otro lado, los experimentos más complejos, como la clasificación de todos los tipos de estrés (Sec. 4.4), demostraron la superioridad de las redes neuronales en términos de desempeño predictivo. Esto pone en evidencia la capacidad de las redes para capturar patrones locales complejos en datos altamente dimensionales, un atributo que supera a cualquier otro modelo probado en este trabajo. Sin embargo, estos beneficios están condicionados a la disponibilidad de un volumen significativo de datos y a una adecuada mitigación de problemas como el sobreajuste y la influencia del efecto de lote. La correcta separación de los conjuntos de validación y test resultó crucial en este sentido.

Finalmente, la comparación entre modelos tradicionales y alternativos subrayó que, aunque las técnicas de machine learning ofrecen ventajas claras en escenarios con mayor cantidad y diversidad de datos, los métodos estadísticos tradicionales siguen siendo útiles en análisis más simples o cuando los recursos computacionales o la cantidad de datos son limitados. La ausencia de genes en común entre estos enfoques (Sec. 5.1.1) sugiere que ambos tienen fortalezas complementarias, dependiendo del contexto experimental y las preguntas biológicas que se deseen abordar.

En conclusión, este trabajo destaca la importancia de integrar enfoques tradicionales y alternativos en bioinformática, aprovechando al máximo los avances recientes en

secuenciación y análisis de datos genómicos. Esto no solo contribuye a la identificación de biomarcadores en soja, sino que también establece un marco metodológico que puede aplicarse a otros cultivos y desafíos en la agricultura moderna

## Bibliografía


[1] W. Depeng. A nuclear factor y-b transcription factor, gmnfyb17, regulates resistance to drought stress in soybean. *International Journal of Molecular Sciences*, 2022.

[2] Nazari L. Integrated transcriptomic meta-analysis and comparative artificial intelligence models in maize under biotic stress. *Scientific Reports*, 2023.

[3] X. Zhou. Using random forest and biomarkers for differentiating covid-19 and mycoplasma pneumoniae infections. *Scientific Reports*, 2024.

[4] Almeida-Silva F. The soybean expression atlas v2: A comprehensive database of over 5000 rna-seq samples. *The Plant Journal*, 2023.

[5] V. Brancato. Standardizing digital biobanks: integrating imaging, genomic, and clinical data for precision medicine. *Journal of Translational Medicine*, 2024.

[6] Leinonen R. The sequence read archive. *Nucleic Acids Research*, 2010.

[7] J. Leek. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 2010.

[8] Zhang Y. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2020.

[9] Robinson M. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 2010.

[10] M.I. Love. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 2014.

[11] Robinson MD. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010.

[12] Tibshirani R. *An Introduction to Statistical Learning*. Springer, 2013.

[13] Guido Freire. *https://github.com/freire-guido/soy-biomarkers*, 2024.

[14] D. Kohlhase. Comparing early transcriptomic responses of 18 soybean (glycine max) genotypes to iron stress. *International Journal of Molecular Sciences*, 2021.