Èlia Rovira Figueras
Anàlisi de Dades Òmiques
6 novembre 2024

Informe de la pràctica: Creació d'un Contenidor SummarizedExperiment i Anàlisi Inicial de Dades

Taula de continguts

1. Resum executiu

Aquest informe documenta el procés de creació d'un contenidor de dades utilitzant l'objecte SummarizedExperiment en R, una eina dissenyada per treballar amb conjunts de dades complexos que inclouen tant dades principals com metadades associades. Aquest exercici parteix d'un estudi biomètric relacionat amb els metabotips de resposta a la cirurgia bariàtrica, amb l'objectiu de comprendre millor les variacions metabòliques entre individus sotmesos a aquest tipus d'intervencions, independentment de la magnitud de la pèrdua de pes.

L'informe cobreix el procés complet, des de la descàrrega i preparació de les dades fins a l'anàlisi descriptiva i la visualització inicial de les dades. La preparació de dades inclou la revisió i harmonització de noms de variables entre els diferents fitxers, així com la transposició de dades per adaptar-les al format SummarizedExperiment. L'objecte final creat facilita una gestió unificada de les dades principals i les metadades, millorant així l'eficiència i organització de les anàlisis posteriors.

Per tal de tenir una visió inicial de les dades, s'han generat visualitzacions descriptives, com histogrames, boxplots i un mapa de calor de les correlacions, que han permès observar la distribució de les variables, la presència de valors extrems i les relacions entre variables. Aquestes visualitzacions són útils per identificar patrons generals en les dades, oferint pistes per a anàlisis més profundes.

Finalment, el projecte complet ha estat reposat a GitHub, incloent-hi l'informe detallat, el codi utilitzat, i les dades originals, facilitant així la seva revisió, ús i extensió per part d'altres investigadors. Aquest enfocament sistemàtic garanteix la transparència, la reproductibilitat i la col·laboració en l'àmbit de la bioinformàtica.

2. Objectiu de l'estudi

Els múltiples objectius de l'estudi es mostren a continuació:

1. Crear un contenidor de dades integrat: Utilitzar l'objecte SummarizedExperiment per crear un contenidor que permeti gestionar de manera integrada les dades principals de l'estudi i les metadades associades. Aquest contenidor facilita l'accés, manipulació i anàlisi de les dades en un únic objecte estructurat.
2. Realitzar una anàlisi descriptiva inicial: Proporcionar una visió general de les dades mitjançant estadístiques descriptives i visualitzacions. Això inclou l'anàlisi de la distribució de les variables, la identificació de valors extrems i l'avaluació de les relacions entre les variables. Aquest pas és essencial per comprendre les característiques generals del conjunt de dades abans d'aprofundir en anàlisis més complexes.
3. Identificar patrons i correlacions: Mitjançant visualitzacions com el mapa de calor de correlacions, es busca identificar possibles agrupaments de variables o relacions significatives entre elles. Aquesta informació és útil per a posteriors anàlisis, ja que permet detectar variables amb característiques similars o amb una estructura comuna.
4. Publicar el projecte de manera oberta: Reposar el projecte complet en un repositori públic de GitHub, incloent-hi l'informe, el codi i les dades originals. Això permetrà a altres investigadors revisar, reproduir i utilitzar el treball com a base per a futurs estudis, contribuint a l'avenç de la investigació en l'àmbit de la bioinformàtica i la biomedicina.
5. Desenvolupar habilitats d'anàlisi de dades bioinformàtiques: Familiaritzar-se amb l'ús d'eines bioinformàtiques en R per a l'anàlisi de dades complexes, una competència essencial en bioinformàtica. Aquesta pràctica posa en valor la capacitat d'integrar dades i metadades en un entorn organitzat i de presentar els resultats de manera clara i comprensible, una habilitat crucial en el camp de l'anàlisi de dades biomèdiques.

3. Materials i Mètodes

3.1 Descripció de les Dades

- Fitxer DataInfo_S013.csv: Inclou metadades per 695 variables amb les següents columnes:
    - VarName: Nom de la variable.
    - varTpe: Tipus de la variable (ex., categòrica, numèrica).
    - Description: Descripció breu de cada variable.
- Fitxer DataValues_S013.csv: Conté les dades principals amb 39 files (una per mostra) i 696 columnes (una per variable i una d'identificació).

3.2 Eines Bioinformàtiques

Les anàlisis es van realitzar amb el llenguatge de programació R, utilitzant les següents eines i llibreries:

- SummarizedExperiment: Per a la integració de dades principals i metadades.
- ggplot2: Per a la generació de visualitzacions de dades.
- ComplexHeatmap: Per al mapa de calor de correlacions.
- dplyr i tidyr: Per a la manipulació de dades.

3.3 Procediment de Preparació de Dades

1. Descàrrega de Dades: Es van descarregar els fitxers DataInfo_S013.csv i DataValues_S013.csv.
2. Comparació de Noms: Es va realitzar una comparació per assegurar que els noms de les variables coincidissin entre data_info i data_values.
3. Transposició de les Dades: Es va transposar data_values per ajustar l'estructura requerida per SummarizedExperiment.
4. Creació de l'Objecte SummarizedExperiment: Es va combinar data_values amb les metadades de data_info per crear l'objecte SummarizedExperiment.

4. Resultats

4.1 Creació del Contenidor SummarizedExperiment

L'objecte SummarizedExperiment es va crear amb les següents característiques:

- Dimensions: 695 files i 39 columnes.
- Classe: SummarizedExperiment, que permet treballar amb conjunts de dades complexos.
- Assays: Un únic assay anomenat counts, que emmagatzema els valors de les 695 variables per a cadascuna de les 39 mostres.
- Row Data: Inclou 4 columnes (VarName, varTpe, Description) que proporcionen informació addicional per a cada variable.

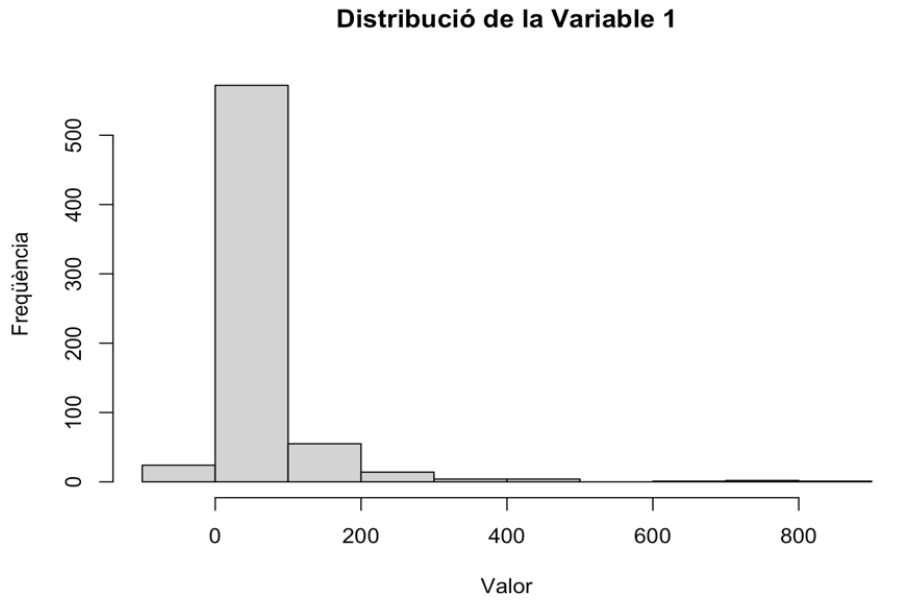
4.2 Exploració del Dataset

- Verificació de Dimensions: Es va verificar que les dimensions del dataset fossin consistents amb el nombre de variables i mostres.
- Conversió de Valors No Numèrics: Es van identificar columnes amb valors no numèrics que es van convertir a format numèric. Els valors no convertibles es van substituir amb NA per evitar errors en anàlisis futures.

5. Visualització de les dades

5.1 Histograma de la Distribució de Valors

L'objectiu de l'histograma és visualitzar la distribució de la Variable 1 i obtenir informació sobre la freqüència dels valors.



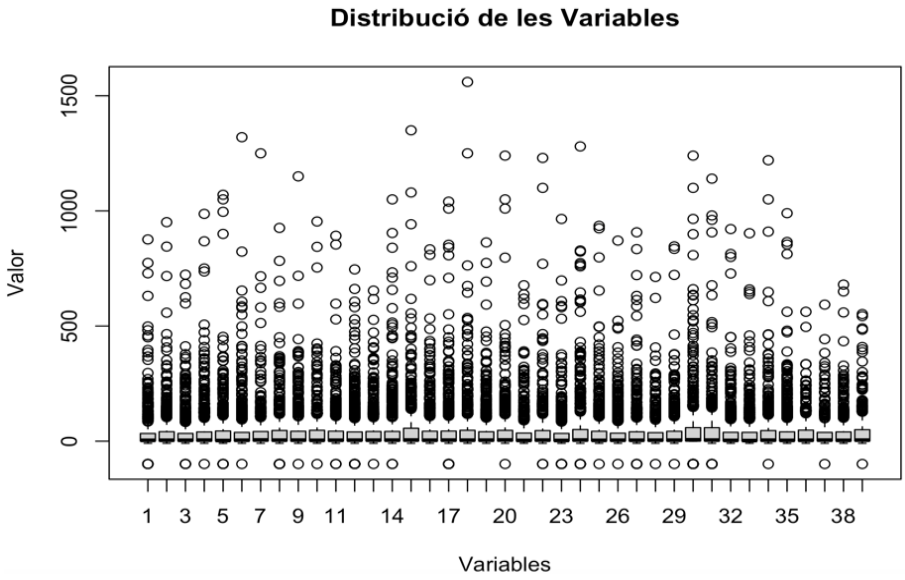
Imatge 1. Gràfic en forma d'histograma de la distribució de la variable 1.

- L'histograma mostra una concentració de valors propers a zero, amb una distribució que disminueix progressivament.
- La major part de valors es troben entre 0 i 200, amb alguns punts extrems fins a 800.

Aquesta distribució suggereix que la Variable 1 representa un fenomen amb una variabilitat significativa entre mostres.

5.2 Boxplot de la Distribució de Variables

L'objectiu del boxplot és examinar la distribució de cadascuna de les variables i identificar possibles outliers.



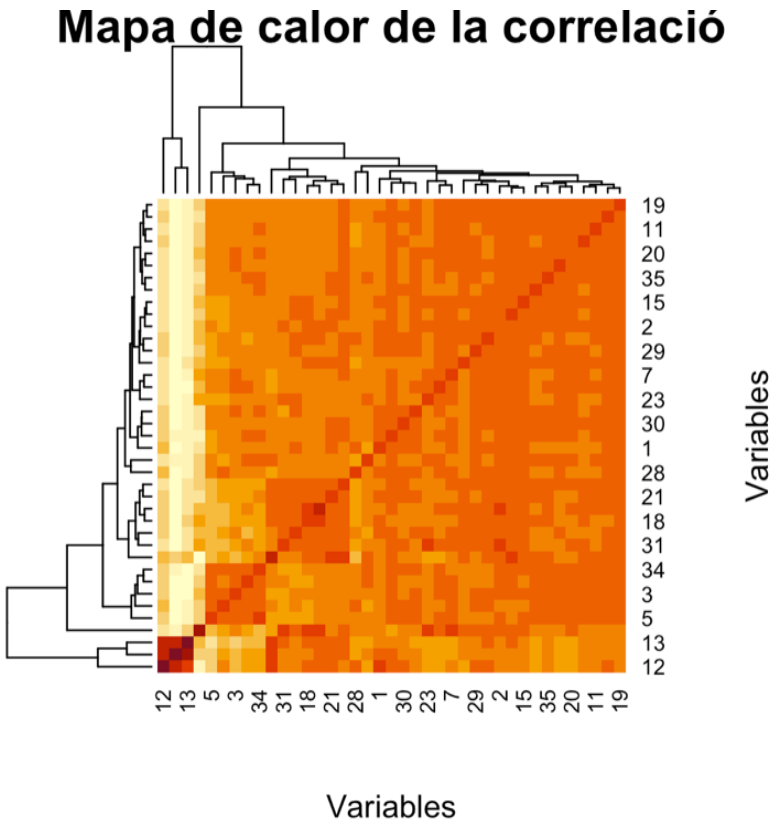
Imatge 2. Boxplot de la distribució de les diverses variables de l'estudi.

- Cada boxplot representa una variable, mostrant el rang interquartílic, valors mínims i màxims, i punts atípics.
- La concentració de valors prop de zero amb múltiples outliers suggereix que algunes variables tenen distribucions asimètriques.

Els boxplots permeten una ràpida identificació de les variables amb més variabilitat.

5.3 Mapa de Calor de la Correlació entre Variables

L'objectiu del mapa de calor és la identificació de les relacions de correlació entre les variables.



Imatge 2. Mapa de calor per observar la distribució de les diverses variables de l'estudi.

- Els colors foscos indiquen una correlació alta, mentre que els colors clars mostren una correlació baixa o nul·la.
- S'observen grups de variables altament correlacionades, suggerint que poden representar aspectes relacionats.

El mapa de calor facilita la identificació de grups de variables amb relacions significatives.

6. Discussió i Limitacions

En aquesta pràctica, s'ha utilitzat l'objecte SummarizedExperiment per gestionar un conjunt de dades complexes associat a un estudi biomètric sobre els metabotips de resposta a la cirurgia bariàtrica. Aquest enfocament ha permès integrar tant les dades principals com les metadades, facilitant així un accés organitzat i coherent a la informació. Aquesta estructura integrada és especialment útil per a l'anàlisi de dades biomèdiques, ja que permet treballar amb múltiples tipus de dades en un sol contenidor, afavorint la comparació i la identificació de patrons o relacions. Tot i així, aquest estudi presenta una sèrie de limitacions que convé considerar per a futures investigacions.

Una de les primeres limitacions trobades és la qualitat de les dades en si mateixes. Durant el procés de preparació, es van identificar diverses columnes amb valors no numèrics que van requerir un tractament especial, com la conversió a valors numèrics o la substitució de dades no vàlides per valors NA. Aquest procediment va ser necessari per assegurar la coherència de les dades, però també implica una possible pèrdua d'informació o una simplificació dels resultats. En estudis biomèdics, aquesta qualitat de les dades pot ser crucial, ja que els valors no numèrics poden representar característiques importants que, en ser substituïdes, poden alterar la interpretació final dels resultats. A més, la presència de valors NA pot afectar les anàlisis estadístiques futures, especialment en mètodes que no toleren dades incompletes. Per tant, en investigacions posteriors, seria recomanable explorar tècniques més sofisticades per al tractament de dades incompletes, com la imputació de valors, per reduir la possible pèrdua d'informació.

Durant la fase d'anàlisi descriptiva, es va observar una gran quantitat de valors atípics (outliers) en diverses variables. Aquesta presència d'outliers podria ser indicativa d'una variabilitat inherent dins les dades, però també pot dificultar l'anàlisi, ja que pot interferir en la identificació de patrons consistents. Tot i que els outliers poden proporcionar informació valuosa sobre variacions individuals en la resposta a la cirurgia bariàtrica, una elevada quantitat d'aquests valors podria ser problemàtica en models estadístics, especialment en aquells que assumeixen una distribució normal de les dades.

Una altra limitació significativa observada en aquesta anàlisi és la possibilitat de correlacions espúries. Tot i que el mapa de calor de correlacions proporciona informació visual valuosa sobre les relacions entre les variables, algunes d'aquestes correlacions podrien ser aleatòries o degudes a coincidències dins el conjunt de dades. En estudis biomèdics, on les dades poden provenir de fonts molt

diverses, les correlacions espúries poden conduir a conclusions errònies si no es revisen amb cautela.

Per evitar la interpretació equivocada de les correlacions, en futures investigacions seria recomanable aplicar tests estadístics de significació per validar les relacions observades i descartar aquelles que no siguin estadísticament significatives. També es podria considerar la utilització de mètodes de reducció de dimensionalitat, com l'anàlisi de components principals (PCA), per identificar les variables més rellevants i minimitzar el risc de correlacions espúries.

Finalment, tot i que l'objecte SummarizedExperiment és una eina potent per integrar dades i metadades en un únic contenidor, el seu ús també comporta algunes limitacions. Aquesta estructura pot ser complexa d'utilitzar per investigadors que no estiguin familiaritzats amb R o amb els mètodes de manipulació de dades bioinformàtiques. A més, l'objecte SummarizedExperiment té certes restriccions de format que poden limitar la flexibilitat en la preparació i manipulació de dades més enllà del format requerit.

En estudis futurs, es podria considerar l'ús d'alternatives o extensions de SummarizedExperiment que permetin una major flexibilitat i facilitat d'ús, o proporcionar formació complementària als investigadors per maximitzar el potencial d'aquesta eina.

7. Conclusions

Aquest informe ha demostrat l'eficàcia de l'objecte `SummarizedExperiment` per gestionar i analitzar dades complexes en estudis biomèdics, com l'anàlisi de metabotips en resposta a la cirurgia bariàtrica. La integració de dades principals i metadades en un únic contenidor ha facilitat la manipulació i l'accés a la informació, establint una base sòlida per a futures anàlisis. Aquesta estructura organitzada permet treballar de manera eficient amb grans volums de dades, fet que resulta especialment útil en bioinformàtica.

El procés de preparació i neteja de les dades ha posat en relleu la importància d'un preprocessament rigorós, ja que valors no numèrics o inconsistències poden comprometre la qualitat de l'anàlisi. Aquest aspecte és essencial, ja que una preparació inadequada pot afectar la validesa dels resultats i dificultar la interpretació final.

Les visualitzacions descriptives, com els histogrames, els boxplots i el mapa de calor de correlacions, han proporcionat una visió inicial de la distribució de les dades i de les relacions entre les variables. Tot i que aquestes eines han estat útils per identificar patrons generals, també s'han detectat limitacions, com la presència de valors atípics i correlacions espúries, que poden complicar la interpretació dels resultats.

Finalment, la publicació del projecte a GitHub ha promogut la transparència i la possibilitat de reproduir els resultats, contribuint a la col·laboració científica. Aquesta pràctica de fer els projectes accessibles a altres investigadors fomenta la revisió i reutilització dels resultats per a futurs estudis.

En conjunt, l'ús de `SummarizedExperiment` ha estat efectiu per a la gestió i l'anàlisi de dades complexes, establint una estructura sòlida per a futures investigacions en el camp de la bioinformàtica i la investigació biomèdica. Tot i les limitacions, aquesta metodologia ofereix una base robusta per a anàlisis posteriors.

8. Publicació del Projecte a GitHub

Per tal de fomentar la transparència i la col·laboració, el projecte complet es va reposar a GitHub, amb els següents components:

1. Informe en format .Rmd i .html/.pdf: Documentant el procés complet.
2. Objecte SummarizedExperiment (.Rda): Emmagatzemat com a objecte reutilitzable.
3. Codi per a l'exploració de dades: Inclou tot el codi utilitzat en l'anàlisi.
4. Fitxers de dades originals en CSV: Disponible per a la seva reproducció.
5. Arxiu README: Amb una descripció general del projecte i les metadades.

URL del repositori: https://github.com/erovirafigueras/Rovira-Figueras-Elia-PEC1.git