

Da abundancia á evidencia: fluxo de traballo para a análise diferencial de datos ómicos en R

Xornadas de R

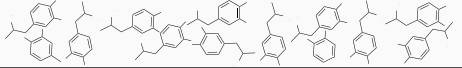
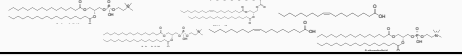
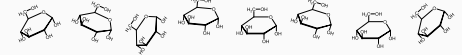
Julia García Currás
Biostatech | CITIC-UDC
16/10/2025

Introducción

As ómicas

Nova disciplina en bioloxía que estuda, de maneira global e masiva, diferentes moléculas dun organismo.

- Xéranse datos biolóxicos de grande dimensionalidade que requiren dunha **análise estatística: R software**.
- Tipos de moléculas e as súas correspondentes ómicas:

Tipos de molécula	Ómica	Datos finais	
ADN	Xenómica		
ARN	Transcriptómica		
Proteínas	Proteómica		
Metabolitos	Metabolómica		
Lípidos	Lipidómica		
Glúcidos	Glicómica		

R e as ómicas

Vantaxes do uso de R na análise de datos ómicos:

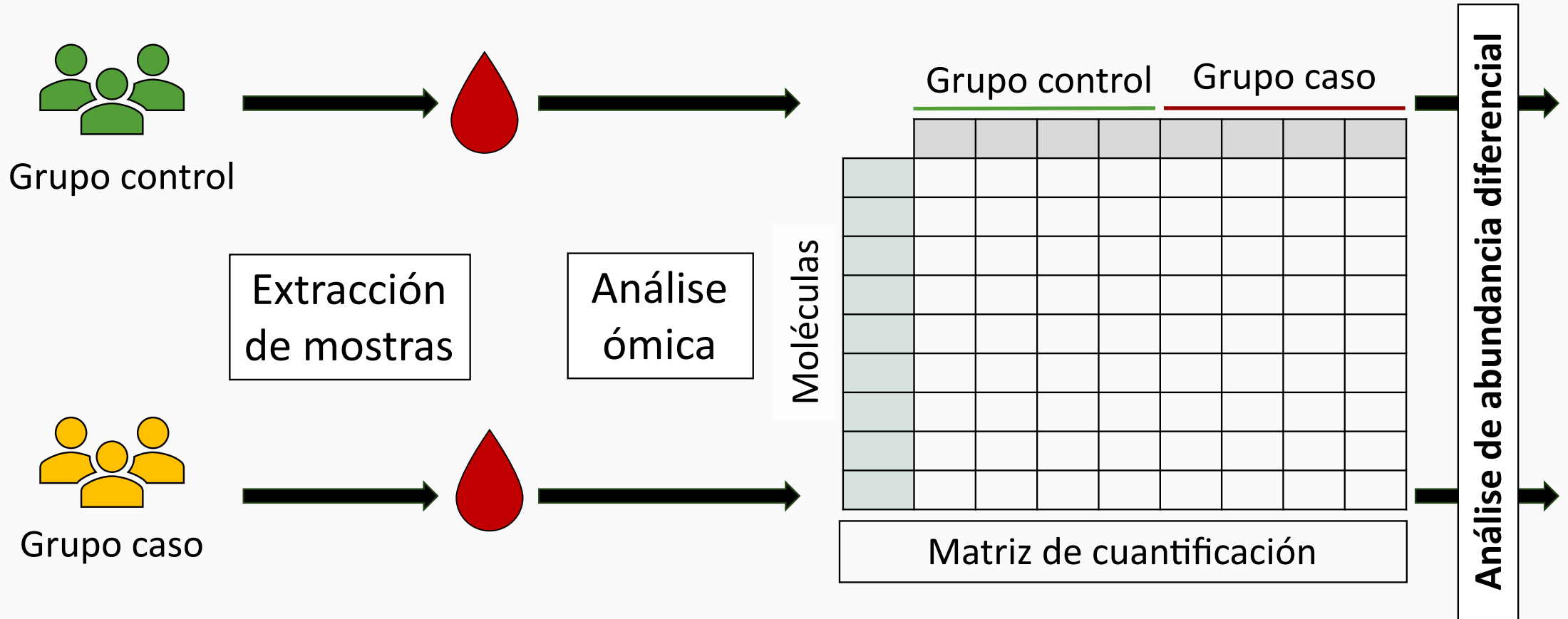
- Software libre e gratuito
- Ampla e activa comunidade de usuarios:
 - Desenvolvimento, mantemento e mellora constante dos diferentes paquetes de análise
- Estandarización e reproducibilidade de análises
- Código aberto e compartido:
 - Transparencia
 - Flexibilidade



Open source software for
Bioinformatics

Análise de abundancia diferencial

Diseño experimental común en análisis de abundancia



Fluxo xeral na análise de abundancia diferencial



Exemplo



Grupo control:

Suxeitos sans

N= 8



Tipo de mostra:

Bágoas



Grupo DM:

Pacientes con diabetes tipo 2

N= 8



Tipo de molécula:

Proteínas

Proteómica

TOTAL MOSTRAS: 16

[illegible]

TOTAL PROTEÍNAS: 1407

Amorim, M.; Martins, B.; Caramelo, F.; Gonçalves, C.; Trindade, G.; Simão, J.; Barreto, P.; Marques, I.; Leal, E. C.; Carvalho, E.; Reis, F.; Ribeiro-Rodrigues, T.; Girão, H.; Rodrigues-Santos, P.; Farinha, C.; Ambrósio, A. F.;

Silva, R.; Fernandes, R. Putative Biomarkers in Tears for Diabetic Retinopathy Diagnosis. *Front Med (Lausanne)* 2022, 9, 873483. <https://doi.org/10.3389/fmed.2022.873483>.

Preprocesado

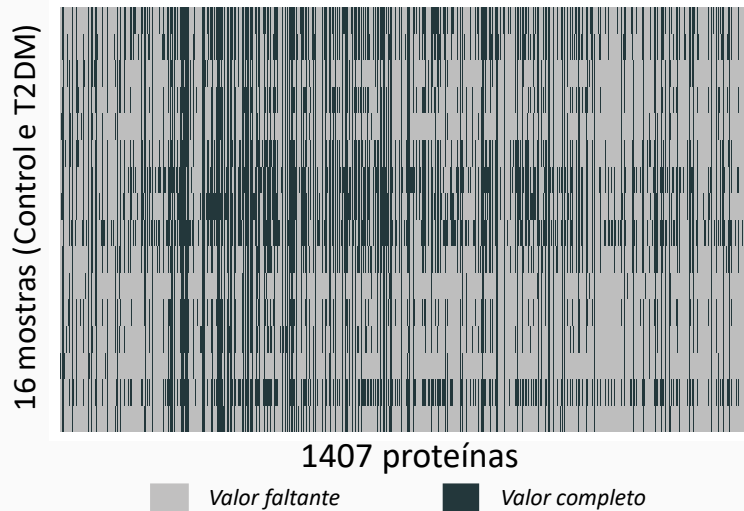
1. Transformación logarítmica da matriz de datos.
2. Análise de datos faltantes:
 - **Filtrado**: eliminación de moléculas con moitos valores faltantes. Recomendación: **filtrar por grupos**.
 - **Imputación**: substitución de valores faltantes por valores estimados.
3. **Normalización** para reducir o erro sistemático presente nos datos.

Método	Paquete R
Imputación	
KNN	<code>impute::impute.knn</code>
RF imputation	<code>missForest::missForest</code>
• Bayesian PCA • Local Least Squares	<code>pcaMethods::pca</code> • <code>method = bpca</code> • <code>method = llsImpute</code>
Normalización	
Quantile	<code>preprocessCore::normalize.quantiles</code>
VSN	<code>vsn::justvsn</code>
CyclicLoess	<code>limma::normalizeCyclicLoess</code>
Varias normalizacións con evaluacóns gráficas e métricas	<code>NormalyzerDE</code>

Preprocesado (exemplo)

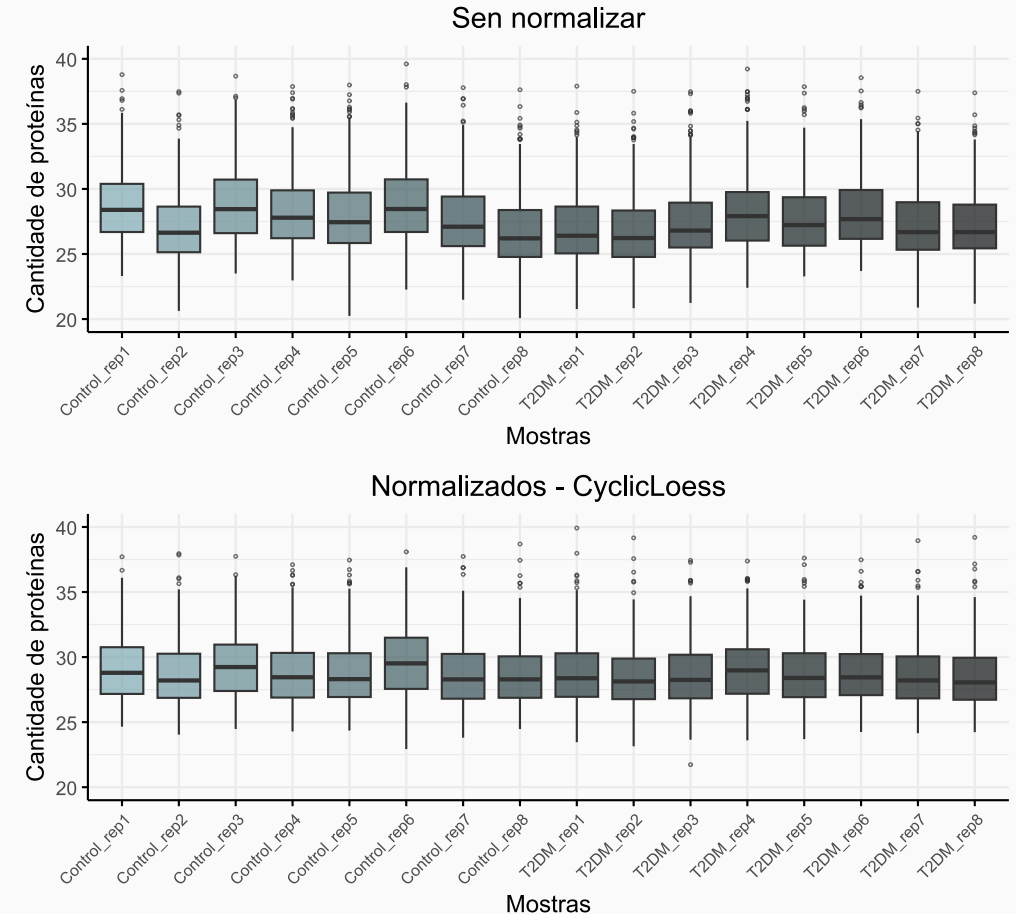
Datos faltantes:

- Total: 67.9 %
- Grupo NC: 27.35% | Grupo DM: 23.28%
- **Filtro:** máximo 50% datos faltantes por grupo.



291 proteínas de 1407 proteínas totais

Normalizaci3n:



Comparacións entre grupos

1. Comparación da cantidade de cada molécula entre grupos.

2. Corrección do p-valor por comparacións múltiples:

3. Cálculo do $\log_2$ fold change ou $\log FC$:

$$\log_2 FC = \log_2 \left(\frac{\text{mean}_{Case}(\text{molecule}_1)}{\text{mean}_{Control}(\text{molecule}_1)} \right)$$

4. Representación: gráfico de tipo volcán ou volcano plot.

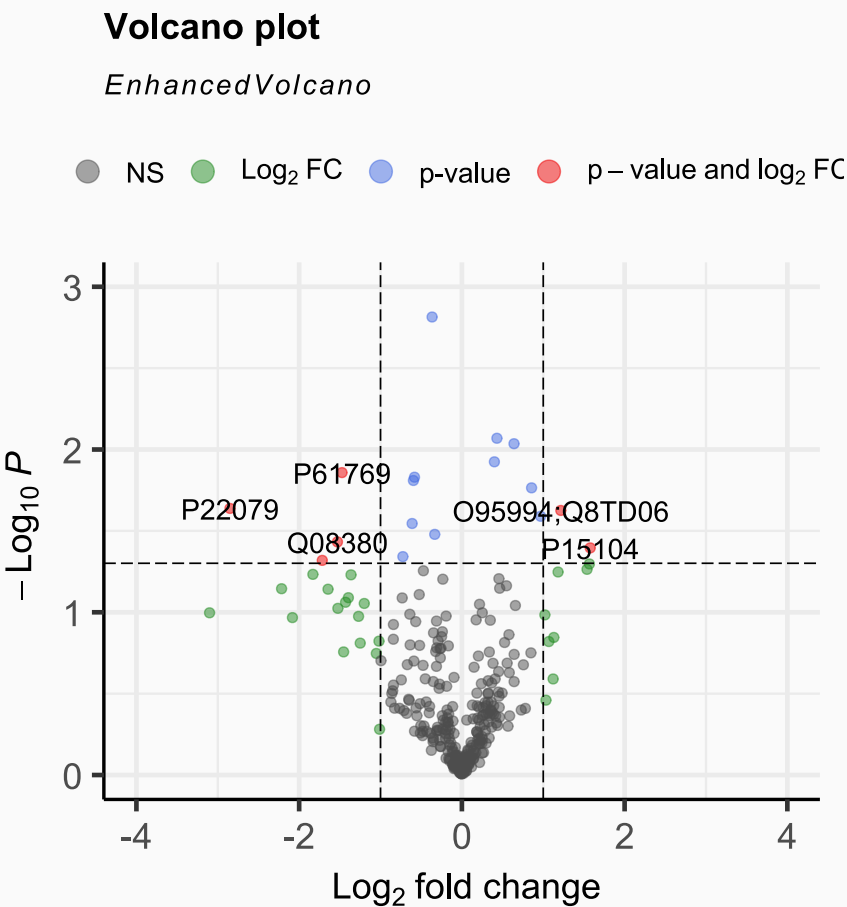
Método	Paquetes R
Comparacións	
Test clásicos (test t, test de Welch)	<code>stats::t.test</code>
Modelos lineais	<code>DESeq2</code> <code>limma</code>
Opción de ambos con correccións	<code>NormalyzerDE</code>
Corrección do p-valor	
Bonferroni, Benjamin-Hochberg, Holm...	<code>stats::p.adjust</code> <code>stats::p.adjust.methods</code>
Representación gráfica	
Gráfico de volcán (volcano plot)	<code>ggplot2</code> <code>EnhanceVolcano</code>

Comparacións entre grupos (exemplo)

Táboa de resultados:

ID	Proteína	logFC	FC	P-valor	P-valor ajustado
P61626	LYZ	-0.37	19.75	0.00	0.45
P01876	IGHA1	0.43	20.30	0.01	0.63
P0DOX7		0.40	20.28	0.01	0.63
P61769	B2M	-1.47	19.00	0.01	0.63
Q8NFU3	TSTD1	-0.58	19.60	0.01	0.63
O95336	PGLS	-0.60	19.59	0.02	0.63
Q6ZVX7	NCCRP1	0.86	20.60	0.02	0.63
P22079	LPO	-2.85	18.12	0.02	0.68

Volcano plot (EnhancedVolcano):

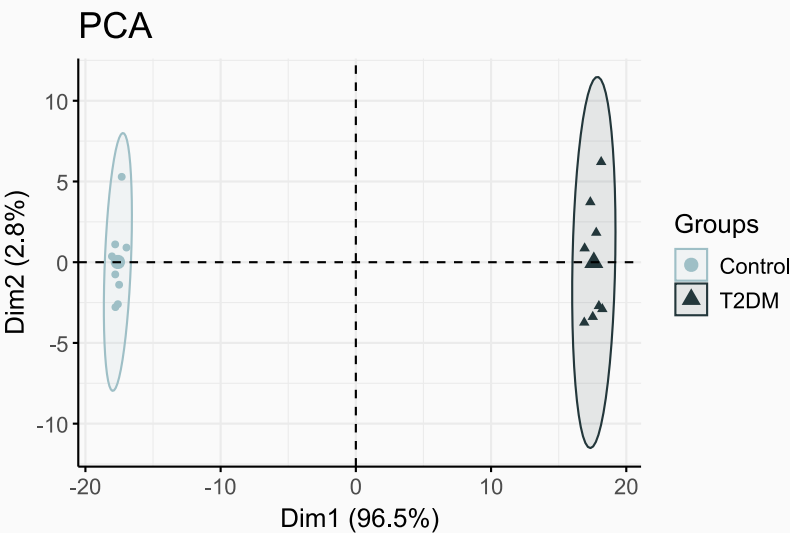


Análise de agrupación

As moléculas con abundancia diferencial permiten discriminar entre os grupos de estudo no seu conxunto?

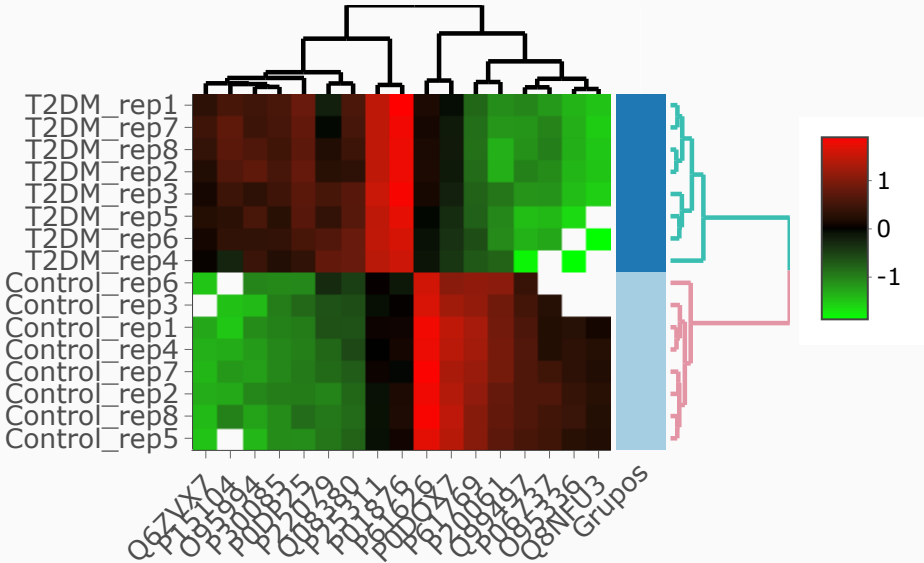
- Análise de compoñentes principais (PCA)

PCA	
Análise	<code>pcaMethods::pca stats::princomp</code>
Representación gráfica	<code>factoextra</code>



- Clustering xerárquico

Clustering xerárquico	
- Estimación de distancias - Aglomeración	<code>stats::dist</code> <code>stats::hclust</code>
- Heatmaps - Dendrogramas	<code>stats::heatmap, Heatmaply</code> <code>Dendextend, circlize</code>



Análise funcional

Permite coñecer aquelas funcións asociadas a moléculas con abundancia diferencial. Dise que estás funcións están sobrerrepresentadas nos resultados.

- Base de datos de termos funcionais: relación molécula - función.
- Lista de moléculas: significativas e non significativas.

	Con función A	Sen función A	Total
Moléculas significativas	n_{11}	n_{12}	n_{1+}
Moléculas non significativas	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- Constraste: test χ^2 de Pearson ou test hiperxeométrico.
- Paquetes: `clusterProfiler::enrichGO`, `AnnotationDbi`, `enrichplot`.

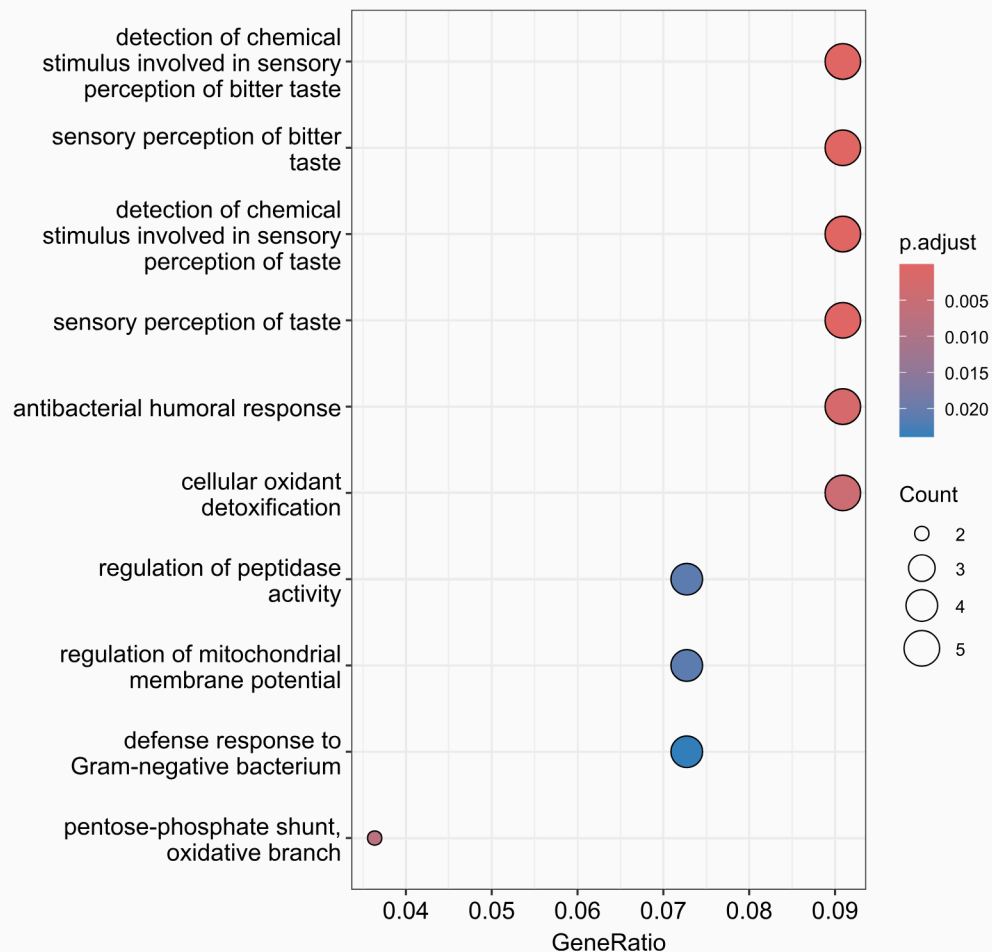
Análise funcional (exemplo)

Táboa de resultados:

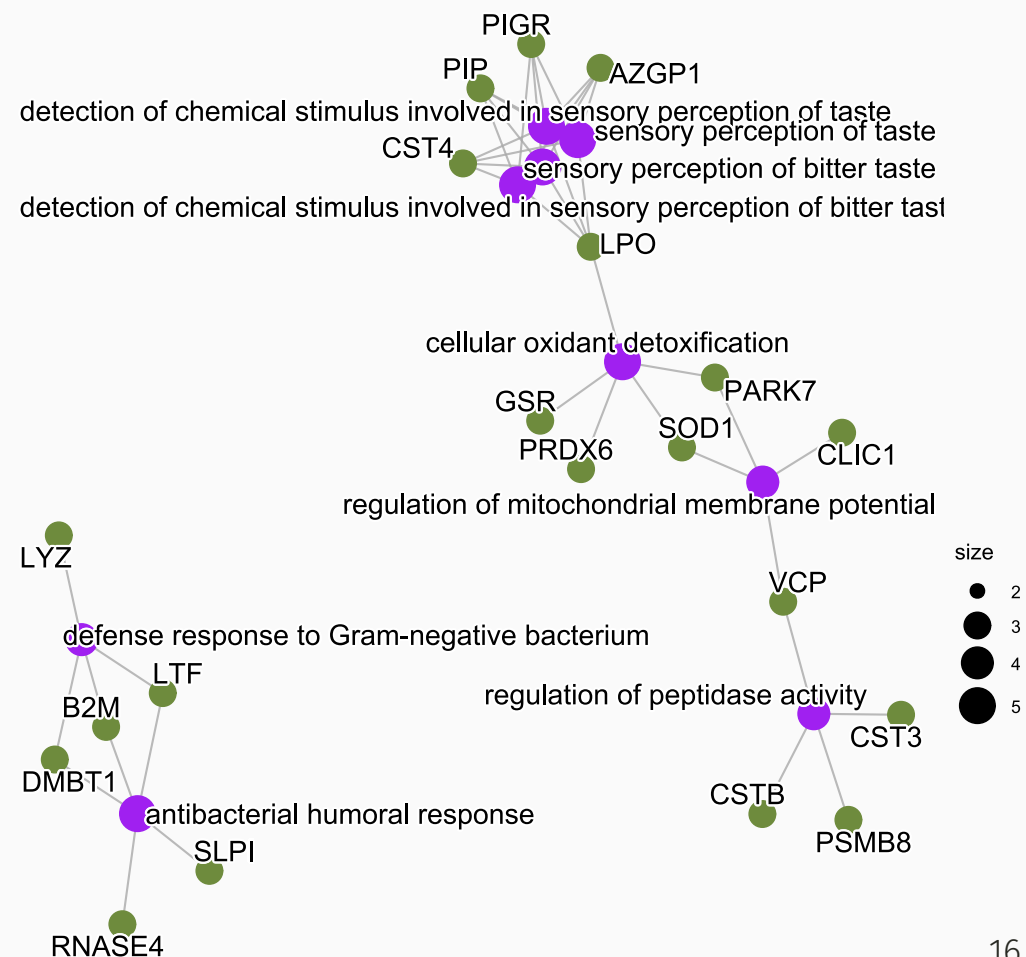
ONTOLOGY	ID	Description	GeneRatio	pvalue	p.adjust	geneID
BP	GO:0050909	sensory perception of taste	5/55	0.000	0.000	LPO/AZGP1/PIP/CST4/PIGR
BP	GO:0019731	antibacterial humoral response	5/55	0.000	0.002	B2M/DMBT1/RNASE4/SLPI/LTF
BP	GO:0098869	cellular oxidant detoxification	5/55	0.000	0.004	LPO/PARK7/GSR/SOD1/PRDX6
BP	GO:0052547	regulation of peptidase activity	4/55	0.000	0.021	CSTB/VCP/CST3/PSMB8

Análise funcional (exemplo)

Resultados enriquecimiento (`enrichplot::dotplot`):



Rede proteínas-funções (`enrichplot::cnetplot`):



Conclusiones

Conclusións

- Con este fluxo de traballo pódese obter a partir de datos ómicos de abundancia:
 - **Biomarcadores** candidatos
 - **Funcións biolóxicas** chave
- Fluxo de traballo **estensible e personalizable**
- O uso de R neste tipo de análises é moi importante para asegurar a **reproducibilidade, transparencia e transferencia** dos coñecementos.

Moitas grazas!

Julia García Curras



julia.garcia@biostatech.com



juliagcurras

BIOSTATΣCH



UNIVERSIDADE DA CORUÑA



