Ein Mann, der recht zu wirken denkt,
Muß auf das beste Werkzeug halten
The man who seeks to be approved,
must stick to the best tools for it

1 Methodology

Working phases, with the corresponding steps, followed in order to achieve the above objectives:

SHA DE SEPARAR ENTRE AQUI I ELS RESULTATS

1. Application of integrative multi-omics methods to (I) the analysis of specific data sets provided by research units from our former affiliation center, VHIR, and other research institutions that we collaborate with [@rodriguez-hernandez_cinacalcet_2016], [@rodriguez-fernandez_phosphatidyl_2018], [@simats_mouse_20 and (II) to the integrative analysis of larger data sets from public data bases, such as Breast Cancer samples from the TCGA project [TCGA Research Network: http://cancergenome.nih.gov/], [TCGA-BRCA Project: https://portal.gdc.cancer.gov/projects/TCGA-BRCA].

2. Development of methods, either in terms of new algorithms or in terms of combinative workflows, which will be able to improve, and facilitate, the analysis and biological interpretation of those data sets to be integrated.

3. Implementation of the methods developed for this study in the appropriate bioinformatics tools, such as an R package or a web-based application, to facilitate their use in the context of biomedical research projects.

Here follows a brief description of these main five activities, the methods in which they are initially based, the objectives that they are related to, and the corresponding results:

SEPARAR TAMBE AIXO EN ELS METODES DE CADA FASE (CLINICA NEUROCIENCIES, METODOLOGICA INTEGRACIO, DESENVOLUPAMENT R...)

1. Application of some state-of-the-art methods for integrative multi-omics data analysis to the study of human brain tissue samples, collected by the Neurovascular Diseases Laboratory at Vall d'Hebron Research Institute. This part is already finished, and led to publications in 2018 and 2021 [@simats_mouse_2020], [@ramiro_integrative_2021]. Researchers obtained different omics data from necropsies, which had been processed to obtain mRNA, microRNA and protein expression values. Each dataset had been first analyzed independently using standard bioinformatics protocols [R Development Core Team. 2008]. These analyses allowed selecting subsets of relevant features, for each type of data, to be used in the integrative analysis. Among all available options, we decided to use two distinct and complementary approaches: (I) Multiple Co-inertia Analysis implemented in Bioconductor packages made4 [@culhane_made4_2005] and mogsa [@singh_diablo_2016], and (II) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by mixomics R package [@rohart_mixomics_2017]. This work had been presented at some meetings [@brianso_ibc_2016], [@brianso_eccb_2016], [@garcia-berrocoso_scb_2016], [@garcia-berrocoso_scbf_2017] and in an already published extended abstract's series book [@brianso_integrative_2017]. This step had been obviously useful for the achievement of the objective number 3 explained in the previous section, which aims on the study of the regulome's response to ischemic stroke, but also useful for detecting the advantages and drawbacks of the methods applied, thus setting the basis for the work regarding to objective number 2.

2. Reproduction of the same analyses steps performed in point 1) above with publicly available databases, such as distinct omics data from 150 samples from the TCGA-BRCA collection. This data set contains the expression or abundance of mRNA, miRNA and proteomics for 150 breast cancer samples previously prefiltered, as explained in Rohart et al. [@rohart_mixomics_2017], and allows identifying a good multi-omics signature to discriminate between Basal, Her2 and Luminal A breast cancer subtypes. This work is already finished, and complies with objectives 3 and 2.

3. Use of all the data sets analyzed up to this point to make a comparison of results between the main implemented methods, and eventually some others, which is the aim of objective 1. This is based on quantitative and qualitative comparison and visualization methods, such as those explained by Thallinger [@pucher_comparison_2019] and Martin [@martin_bisogenet_2010], going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages [@r_core_team_2022] or Cytoscape [@cline_integration_2007]. The focus here is to use graphical visualization elements to compare the results of the analyses with and without the addition of biological information.

4. Development of new methods and/or workflows in order to improve and/or combine the benefits from the selected approaches, with focus in those allowing the addition of biological significance to the integration process. Here follows an overview of the methods developed to expand the original datasets (X, Y) with annotations (Ax, Ay) to obtain new blocks of data (Nx, Ny,and Nxy). And the workflow has been implemented adapting the integrative pipelines applied so far to the R targets package [@landau_targets_2021], a pipeline toolkit that improves reproducibility, skipping unnecessary steps already up to date and showing tangible evidence that the results match the underlying code and data. The development of this targets workflow is intended to comply with the objective number 2 of this working plan.

5. Implementation of the methods resulting from 4) as a new R package to be submitted to Bioconductor repository [@huber_orchestrating_2015], and, finally, to complete objective 4 of this thesis plan, as a web application [@shiny_2021] to be used in further steps of the current biomedical research projects in which our collaborators are implied, as well as in future studies.

AQUI COMENCEN PROPIAMENT ELS METODES

In the context of multi-omics data integration, our proposal relies on the idea that incorporating biological annotations into datasets before integrative analysis enriches outcomes and enhances their biological interpretability. Therefore, augmenting quantitative omics data with contextual biological knowledge will deepen our understanding of complex biological phenomena. To do so, we begin with meticulous data quality assessment and standardization, laying the foundation for reliable analyses. We then infuse biological knowledge using standard biological annotations, creating "Expanded Datasets" that provide context for comprehensive analysis. Advanced dimension-reduction techniques can be applied to illuminate hidden patterns and relationships between data sources or blocks, and the semi-automation capabilities of the Targets R package allow us to build an easy-to-use implementation of the process.

1.1 Data Quality Assessment and Format Review

Before initiating the integrative analysis, a meticulous evaluation of data quality and format compatibility was conducted to ensure the reliability of the input datasets. This crucial step aimed to identify and rectify discrepancies, inconsistencies, or errors that could potentially impact subsequent analyses. During this process, datasets spanning various omics technologies, including transcriptomics and proteomics, are selectively acquired from reputable sources and repositories. Emphasis was placed on meticulous source selection to guarantee consistency and adherence to standardized formats. Subsequently, the raw omics data underwent a comprehensive preprocessing phase, addressing issues such as missing values, outliers, and normalization. This

preprocessing step was indispensable for enhancing data quality and enabling comparability across diverse datasets. Additionally, a thorough review of data formats encompassing file types, column naming conventions, and units of measurement was conducted. Non-standardized data were systematically transformed into a uniform format to streamline the downstream integration processes. Through these procedures, a robust foundation was established for subsequent integrative omics analysis, ensuring coherence and validity of the synthesized biomedical insights.

FALTA DETALLAR AQUI COM SHAN VALIDAT ELS DATASETS DE TCGA (ELS DE STROKE ANIRAN A BANDA) Explicar aquí els requeriments de format dels data sets d'entrada

Mostrar Figure 1.1 i Figure 1.2 PERO AQUESTES MILLOR COM A TAULES INTEGRADES a markdown

Data loaded from files **data/mrna.csv** and **data/prots.csv**

Samples in rows; Features in columns

```
## # A tibble: 150 x 201
##    sample  RTN2 NDRG2 CCDC113 FAM63A ACADS  GMDS HLA.H SEMA4A  ETS2 LIMD2  NME3  ZEB1 CDCP1 GIYD2 RTKN2 MANSC1
##    <chr>  <dbl> <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
##  1 A0FJ    4.36  7.53   3.96    4.46  2.26  6.02  5.01   3.22  4.73  5.10  4.56  3.99  6.42  2.47  4.58   8.08
##  2 A13E    1.98  7.46   5.43    5.44  4.03  4.34  6.18   2.86  5.41  4.21  5.14  3.76  6.61  4.98  5.17   7.30
##  3 A0G0    1.73  8.08   2.23    5.54  2.63  6.36  6.04   5.95  5.65  3.30  4.45  3.85  6.41  3.20  4.24   7.99
##  4 A0SX    4.36  5.79   3.54    4.74  4.27  4.00  7.09   5.01  5.90  5.48  3.76  5.23  6.02  4.30  4.25   5.41
##  5 A143    2.45  7.16   4.69    4.81  2.44  7.03  5.94   5.90  6.64  5.51  4.24  3.54  7.09  5.71  3.96   8.49
##  6 A0DA    4.77  8.75   4.31    5.31  3.24  4.24  6.91   6.59  5.86  3.77  4.26  4.80  6.05  4.01  2.07   5.90
##  7 A0B3    3.35  5.10   0.593   5.22  3.89  5.92  8.04   6.53  6.31  4.11  4.79  4.26  7.27  4.64  3.53   5.03
##  8 A0I2    1.81  3.79   2.72    4.36  4.20  4.83  9.13   4.98  5.30  5.15  5.76  2.46  3.79  6.90  1.91   6.39
##  9 A0RT    2.09  6.33   2.36    4.04  4.13  4.29  7.59   5.94  6.57  7.07  5.13  5.55  3.85  5.66  4.27   3.71
## 10 A131    4.34  4.70   3.64    4.03  3.14  5.25  6.35   6.06  6.32  4.89  5.73  4.52  6.57  5.42  3.36   5.74
## # ... with 140 more rows, and 184 more variables: TAGLN <dbl>, IFIT3 <dbl>, ARL4C <dbl>, HTRA1 <dbl>,
## #   KIF13B <dbl>, CPPED1 <dbl>, SKAP2 <dbl>, ASPM <dbl>, KDM4B <dbl>, TBXAS1 <dbl>, MT1X <dbl>, MED13L <dbl>,
## #   SNORA8 <dbl>, RGS1 <dbl>, CBX6 <dbl>, WWC2 <dbl>, TNFRSF12A <dbl>, ZNF552 <dbl>, MAPRE2 <dbl>,
## #   SEMA5A <dbl>, STAT5A <dbl>, FLI1 <dbl>, COL15A1 <dbl>, C7orf55 <dbl>, ASF1B <dbl>, FUT8 <dbl>,
## #   LASS4 <dbl>, SQLE <dbl>, GPC4 <dbl>, AKAP12 <dbl>, AGL <dbl>, ADAMTS4 <dbl>, EPHB3 <dbl>, MAP3K1 <dbl>,
## #   PRNP <dbl>, PROM2 <dbl>, SLCO3A1 <dbl>, SNHG1 <dbl>, PRKCDBP <dbl>, MXI1 <dbl>, CSF1R <dbl>, TANC2 <dbl>,
## #   SLC19A2 <dbl>, RHOU <dbl>, C4orf34 <dbl>, LRIG1 <dbl>, DOCK8 <dbl>, BOC <dbl>, C11orf52 <dbl>, ...
```

```
## # A tibble: 150 x 112
##    sample   YWHAE EIF4EBP1 TP53BP1    ARAF   ACACA    ACCB  PRKAA1     ANLN     AR ARID1A   ASNS     ATM
##    <chr>    <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
##  1 A0FJ    0.0491   0.447   0.918   0.0227 -0.0863 -0.417   0.285   0.172   -1.31   0.505  0.811 -0.496
##  2 A13E   -0.0800   0.605   0.0591 -0.460  -0.593  -0.0623 -0.275   0.222   -1.62   0.340  1.18  -0.276
##  3 A0G0   -0.0328   0.895   0.517  -0.192   0.411   0.826   0.0677  0.122   -1.08   0.227  1.95   0.771
##  4 A0SX   -0.205   -0.141  -0.314  -0.0748 -0.851  -0.663   0.0296  1.05    -1.27   0.355  0.607  0.781
##  5 A143    0.0602   0.132   0.331  -0.0244  0.770   0.873  -0.217   0.0138  -0.601  0.544  0.539  0.0139
##  6 A0DA    0.0308   0.0330 -0.220   0.419  -0.714  -0.218  -0.0631  0.0603  -1.21  -0.111  0.312  0.0717
##  7 A0B3   -0.108   -0.0371 -0.545   0.431  -0.363  -0.269  -0.0776  0.00887 -1.02  -0.233  1.14  -0.210
##  8 A0I2    0.650   -0.521  -1.60   -0.187   1.08    1.59   -0.0775 -0.0519  -0.421 -0.355 -0.633 -0.924
##  9 A0RT   -0.0137  -0.635  -0.721  -0.375  -1.25   -0.901  -0.178  -0.0419  -0.952 -0.179  0.145  0.834
## 10 A131    0.431    1.05   -0.811   0.353  -0.883  -0.681   0.261  -0.00213 -0.180 -0.475  0.573 -1.44
## # ... with 140 more rows, and 99 more variables: AKT1 <dbl>, ANXA1 <dbl>, BRAF <dbl>, BAK1 <dbl>, BAX <dbl>,
## #   BCL2 <dbl>, BCLX <dbl>, BECN1 <dbl>, BID <dbl>, BCL2L11 <dbl>, RAF1 <dbl>, PECAM1 <dbl>, ITGA2 <dbl>,
## #   CDK1 <dbl>, CASP7 <dbl>, CAV1 <dbl>, CHEK1 <dbl>, CHEK2 <dbl>, CLDN7 <dbl>, COL6A1 <dbl>, CCNB1 <dbl>,
## #   CCND1 <dbl>, CCNE1 <dbl>, PARK7 <dbl>, DVL3 <dbl>, CDH1 <dbl>, EGFR <dbl>, ESR1 <dbl>, MAPK1 <dbl>,
## #   FOXO3 <dbl>, FN1 <dbl>, GAB2 <dbl>, GATA3 <dbl>, GSK3A <dbl>, ERBB2 <dbl>, ERBB3 <dbl>, HSPA1A <dbl>,
## #   IGFBP2 <dbl>, INPP4B <dbl>, IRS1 <dbl>, MAPK9 <dbl>, MAPK8 <dbl>, KRAS <dbl>, XRCC5 <dbl>, LCK <dbl>,
## #   MAPK1_1 <dbl>, MAP2K1 <dbl>, ERRFI1 <dbl>, MRE11 <dbl>, CDH2 <dbl>, NF2 <dbl>, NOTCH1 <dbl>, ...
```

Figure 1.1: Example of gene expression and protein quantification data as loaded in R

	A0FJ	A13E	A0G0	A0SX	A143	A0DA	A0B3	A0I2	A0RT
YWHAE	0.049130778	-0.079982106	-0.032849886	-0.205329492	0.060190211	0.030761714	-0.107861537	0.64984396	-0.013650441
EIF4EBP1	0.447486231	0.605218418	0.894609732	-0.141322924	0.131768992	0.032996799	-0.037124691	-0.52148657	-0.634850633
TP53BP1	0.917834192	0.059101206	0.517044530	-0.313728669	0.330912383	-0.220271002	-0.544743061	-1.60203535	-0.720723295
ARAF	0.022741468	-0.459852981	-0.191821916	-0.074823472	-0.024357467	0.418616650	0.430503500	-0.18714658	-0.374882996
ACACA	-0.086267822	-0.592691835	0.411171898	-0.851480596	0.769751430	-0.714308701	-0.363474049	1.07761482	-1.254491083
ACCB	-0.416624416	-0.062268404	0.825828592	-0.663410436	0.873478702	-0.217526770	-0.269313837	1.58998239	-0.901353585
PRKAA1	0.285270389	-0.275233600	0.067741840	0.029563729	-0.216531821	-0.063065064	-0.077581092	-0.07753959	-0.177636653
ANLN	0.172311102	0.222105981	0.121993985	1.054948103	0.013784220	0.060256895	0.008872461	-0.05187936	-0.041880238
AR	-1.307605693	-1.620475956	-1.077894436	-1.267054694	-0.601327437	-1.208038484	-1.016297633	-0.42122691	-0.952324860
ARID1A	0.505094485	0.339581595	0.227180664	0.355297672	0.544125136	-0.110944799	-0.233223615	-0.35537533	-0.179195256
ASNS	0.811462882	1.181015791	1.950922363	0.607423831	0.538762877	0.311949453	1.138875941	-0.63275876	0.145464752
ATM	-0.495944728	-0.275533386	0.770857796	0.781328690	0.013854306	0.071748319	-0.209624373	-0.92406461	0.833870191
AKT1	-0.001377255	-0.755547887	-0.067397666	0.056726701	0.238114357	0.193712038	-0.301495924	-0.47402849	-0.367759411
ANXA1	-0.092909287	0.194749839	1.252992383	0.575274185	-1.557003586	0.491015188	0.533878400	1.21076392	0.424004827
BRAF	0.476309798	0.143257789	0.224891925	-0.221859607	0.248234872	-0.195445933	-0.036284702	-1.07351919	-0.711215072
BAK1	0.112201063	0.111310840	-0.069962738	-0.036546549	-0.124839115	-0.257300059	-0.115681609	0.87744695	0.183770057
BAX	-0.156538756	-0.205462637	-0.047604780	0.085173319	0.151544397	-0.090041106	-0.041475148	-0.54119284	0.146947246
BCL2	1.060203513	-0.160826453	-1.771917375	0.345023494	-1.588878871	-0.782913123	-1.041432134	0.41442932	0.531749294
BCLX	-0.100950513	-0.171629248	-0.056202128	-0.096473309	-0.140526557	-0.099476757	-0.037510164	0.74769887	-0.275295151
BECN1	-0.019449441	-0.041253352	-0.076969142	0.963238561	-0.219198072	-0.208762350	-0.219048417	-0.65145061	-0.165614955
BID	-0.034821157	-0.298426931	0.073740813	-0.203558523	-0.147058902	-0.035583612	-0.166370155	0.94513564	0.310251463
BCL2L11	0.408337983	-0.442249202	-1.244877548	0.163042908	0.051760204	-0.434277577	-0.290144108	-0.52809090	0.227648621
RAF1	0.108839334	0.403923023	-0.157470172	0.037683828	0.219029836	0.120775889	0.149760561	-0.42642058	-0.284617195
PECAM1	0.096913816	-0.135688779	-0.229473098	0.073040655	-0.037514094	-0.172646324	-0.003427764	-0.57845063	0.081700778
ITGA2	0.056953664	-0.369652041	0.225919578	-0.377061206	0.032209215	-0.279859151	0.064025012	0.20592716	-0.132198583
CDK1	0.391893475	0.431874216	0.170553859	0.487608500	0.356320675	0.057826839	0.201249969	0.15530981	0.238982271
CASP7	-0.209648791	-0.442253479	-0.027768363	0.619517693	-0.113207256	-0.533360022	1.021124691	0.62764170	1.765265466
CAV1	0.533755894	-1.310081134	-2.024819193	-0.105724014	-1.723398601	-1.761346920	-0.396679637	2.68650874	1.696473472
CHEK1	0.160404592	0.223441176	0.376873438	0.004513815	0.227372000	-0.069237193	0.204323902	0.48476037	0.190864987
CHEK2	1.056926875	0.651510308	0.881431094	0.222821482	0.425296288	-0.258980977	0.163256893	-0.70436969	0.476877281
CLDN7	-0.620225952	0.780008039	-0.343776287	0.228050453	0.233078487	0.040042353	-0.287622816	-1.39960030	-1.606562159
COL6A1	-0.869405130	-0.262350291	-0.425013922	-0.159521178	-0.805978550	-0.535507295	-0.513767940	0.97852799	0.835688459
CCNB1	1.516735476	1.025776946	0.977360144	0.569273220	1.368956673	0.071681473	1.237889430	-1.32406936	0.245716912
CCND1	-0.310524605	-0.434476563	-0.226412035	-0.512848244	-0.875023630	0.099692628	-0.359145590	1.01357985	0.247347773
CCNE1	0.987850528	0.249589732	-0.329458663	1.425506793	1.406639282	-0.042159529	0.516007883	0.38413664	0.193681101

Figure 1.2: Example of proteomics input data

1.2 Preprocessing for Integration of Biological Knowledge: Generating the "Expanded Datasets"

The integration of biological knowledge into omics datasets can be achieved through a preprocessing step aimed at expanding the original data matrices with annotations accessed from specialized R libraries, which provided direct access to curated biological databases such as Gene Ontology (GO[@ashburner_gene_2000]) and biochemical pathways information (e.g., KEGG[@kanehisa_kegg_2000]). This process, that combines the annotation of the most significant biological entities with the quantification and integration of their annotation values to the data matrices, ends up with what we term "Expanded Data Sets", which include the original biological features (e.g., gene expression or protein quantification values) as well as new variables coming from the annotation of biological terms. The following steps explain this preprocessing procedure in more detail:

- Selected biological annotations: Specialized R libraries, dedicated to biological knowledge integration, are employed to access and retrieve up-to-date annotations from databases such as GO and KEGG.

- Data-Annotation Mapping: Each omics dataset are mapped to the retrieved biological annotations based on identifiers (e.g., gene or protein names) using the capabilities of the R libraries. This step facilitates the linking of omics data to biological knowledge.

- Annotation Integration: The annotated information is integrated, implementing new R functions, into the starting omics datasets, resulting in expanded data matrices that combined the original quantitative omics measurements with new quantified features associated with the given biological annotations.

1.2.1 Selecció de les fonts d'anotacions biològiques

PENDENT DE DETALLAR com escullo les fonts de les anotacions per defecte. Apuntar que es poden facilitar ja anotacions disponibles prèviament, sempre que

compleixin amb el format que s'explica al següent apartat. Aquestes poden ser estàndard o bé personalitzades a mida de l'usuari (tot i que si es aixi hi ha certes funcionalitats posteriors que no es podran aprofitar).

```
  [1] "RTN2"      "NDRG2"     "CCDC113"   "FAM63A"    "ACADS"     "GMDS"      "HLA.H"     "SEMA4A"    "ETS2"      "LIMD2"     "NME3"
 [12] "ZEB1"      "CDCP1"     "GIYD2"     "RTKN2"     "MANSC1"    "TAGLN"     "IFIT3"     "ARL4C"     "HTRA1"     "KIF13B"    "CPPED1"
 [23] "SKAP2"     "ASPM"      "KDM4B"     "TBXAS1"    "MT1X"      "MED13L"    "SNORA8"    "RGS1"      "CBX6"      "WWC2"      "TNFRSF12A"
 [34] "ZNF552"    "MAPRE2"    "SEMA5A"    "STAT5A"    "FLI1"      "COL15A1"   "C7orf55"   "ASF1B"     "FUT8"      "LASS4"     "SQLE"
 [45] "GPC4"      "AKAP12"    "AGL"       "ADAMTS4"   "EPHB3"     "MAP3K1"    "PRNP"      "PROM2"     "SLCO3A1"   "SNHG1"     "PRKCDBP"
 [56] "MXI1"      "CSF1R"     "TANC2"     "SLC19A2"   "RHOU"      "C4orf34"   "LRIG1"     "DOCK8"     "BOC"       "C11orf52"  "S100A16"
 [67] "NRARP"     "TTC23"     "TBC1D4"    "DEPDC6"    "ILDR1"     "SDC1"      "STC2"      "DTWD2"     "TCF4"      "ITPR2"     "DPYD"
 [78] "NME1"      "EGLN3"     "CD302"     "AHR"       "LAPTM4B"   "OCLN"      "HIST1H2BK" "HDAC11"    "C18orf1"   "C6orf192"  "AMPD3"
 [89] "COL6A1"    "RAB3IL1"   "APBB1IP"   "PSIP1"     "EIF2AK2"   "CSRP2"     "EIF4EBP3"  "LYN"       "WDR76"     "SAMD9L"    "ASPH"
[100] "RBL1"      "SLC43A3"   "HN1"       "TTC39A"    "MTL5"      "NES"       "APOD"      "RIN3"      "ALCAM"     "C1orf38"   "PLCD3"
[111] "BSPRY"     "NTN4"      "IL1R1"     "EMP3"      "ZKSCAN1"   "FMNL2"     "OGFRL1"    "IRF5"      "IGSF3"     "DBP"       "CNN2"
[122] "CAMK2D"    "SIGIRR"    "AKAP9"     "ICA1"      "FGD5"      "DSG2"      "E2F1"      "QSOX1"     "TOB1"      "CSF3R"     "SHROOM3"
[133] "CCDC80"    "FRMD6"     "CXCL12"    "CCNA2"     "TIGD5"     "ALDH6A1"   "POSTN"     "FZD4"      "NCAPG2"    "SDC4"      "SNED1"
[144] "PLEKHA4"   "KCNAB2"   "SH3KBP1"   "IGSF9"     "DNLZ"      "S1PR3"     "PTPRE"     "FLJ23867"  "PLSCR1"    "LMO4"      "IFITM2"
[155] "LRRC25"    "TST"       "NCF4"      "NCOA7"     "IL4R"      "CCDC64B"   "SGPP1"     "RUNX3"     "SLC5A6"    "IFIH1"     "PREX1"
[166] "PLAUR"     "CDK18"     "SLC43A2"   "GK"        "ICAM2"     "YPEL2"     "CBR1"      "MEX3A"     "ZNRF3"     "PTPRM"     "C1orf162"
[177] "GAS6"      "C1QB"      "PVRL4"     "CTSK"      "MRVI1"     "LEF1"      "PLCD4"     "ZNF37B"    "MEGF9"     "GINS2"     "FAM13A"
[188] "CPT1A"     "SNX10"     "TRIM45"    "ELP2"      "ALOX5"     "AMN1"      "CERCAM"    "SEMA3C"    "KRT8"      "TP53INP2"  "JAM3"
[199] "ZNF680"    "PBX1"
```

Figure 1.3: List of gene symbols used as example

1.2.2 Anotació de la info biològica

COM VAM PLANTEJAR fer l'anotació biologica. Quines opcions i amb quins metodes estadístics/bioinformatics... DUBTO SI LO QUE SEGUEIX NO ANIRIA A RESULTATS

For each input data set, if annotations are not already provided, two distinct basic annotation methods can be performed:

(i) a basic GO mapping, returning annotations to those GO entities for which we find more than a certain number of features (gene ids coming from our data set, see Figure 1.3 for an example) annotated to them,

(ii) a Gene Enrichment Analysis (based on Hypergeometric tests against all GO categories, with FDR correction) is performed in order to retrieve the most relevant annotations to that set of genes/features.[@yu_clusterprofiler_2012]

[mostrar exemple de llista de gens]

[punt de millora, que l'anotacio basica pugui ser tb a KEGG]

[mostrar formula]

es mostra exemple en Figure 1.4 POSSIBLE INTEGRAT EN MARKDOWN?

[comentar aquí l'opció d'afegir les anotacions com a individus suplementaris enlloc de variables Figure 1.6 is an example.]

Annotated Matrix

Gene Ontology used: **BP**

Min. number of genes required to pass the filter: **8**

Annotated categories: **13** (for *data/mrna.csv*)

Annotated categories: **61** (for *data/prots.csv*)

Shared annotated categories: GO:0000122, GO:0006357, GO:0007155, GO:0007165, GO:0007411, GO:0008285, GO:0019221, GO:0030335, GO:0045893, GO:0045944

(Showing only partial output)

```
tar_read(categ_sums1)
```

```
## GO:0000122 GO:0006357 GO:0007155 GO:0007165 GO:0007411 GO:0008285 GO:0016477 GO:0019221 GO:0030335 GO:0043312
##         11         14         11         21         10          9         10         10          8          8
## GO:0045893 GO:0045944 GO:0055114
##         11         14         10
```

```
tar_read(categ_sums2)
```

```
## GO:0000082 GO:0000122 GO:0000165 GO:0000187 GO:0001525 GO:0001666 GO:0001701 GO:0001934 GO:0006357 GO:0006367
##          8         19         17          8         11          9         10         10         13         10
## GO:0006468 GO:0006915 GO:0006974 GO:0006977 GO:0007050 GO:0007155 GO:0007165 GO:0007169 GO:0007411 GO:0007507
##         24         17         14          8         10         11         30         11          8         12
## GO:0007568 GO:0008283 GO:0008284 GO:0008285 GO:0010468 GO:0010628 GO:0010629 GO:0016032 GO:0016579 GO:0018105
##         10         12         20         16          8         29         12         19         11         16
## GO:0018107 GO:0018108 GO:0019221 GO:0030154 GO:0030335 GO:0032355 GO:0032869 GO:0033138 GO:0033674 GO:0035556
##         11         11         15         11         11         10          9          9          8         17
## GO:0042060 GO:0042127 GO:0042493 GO:0042981 GO:0043065 GO:0043066 GO:0045471 GO:0045892 GO:0045893 GO:0045944
##          9          9         23         10         11         31          8         11         23         30
## GO:0046777 GO:0048538 GO:0050821 GO:0051091 GO:0051897 GO:0070374 GO:0071456 GO:0090090 GO:0098609 GO:1901796
##         10          8          8          8         12          9         11          8          9          9
## GO:2001244
##          8
```

Figure 1.4: Example of basic Go annotation by raw count against GO Biological Processes, setting 8 as minimum number of genes included in the BP entity. Annotation performed separately for gene expression and protein quantification input files

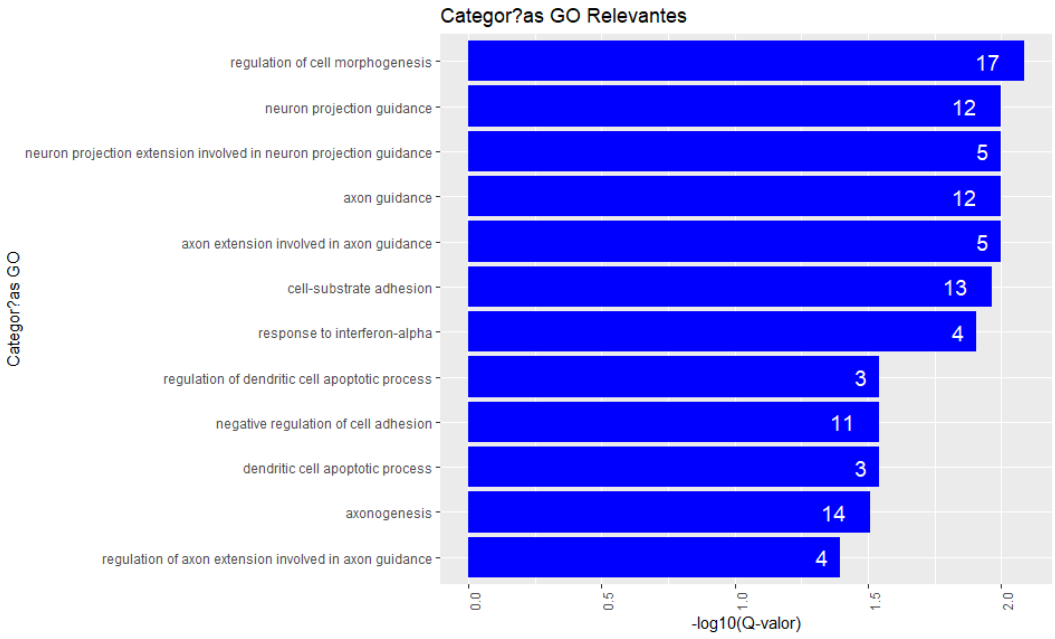

Figure 1.5: Example of results from GO annotation. Results of the biological significance analysis performed with the lists of genes against GO through clusterProfiler

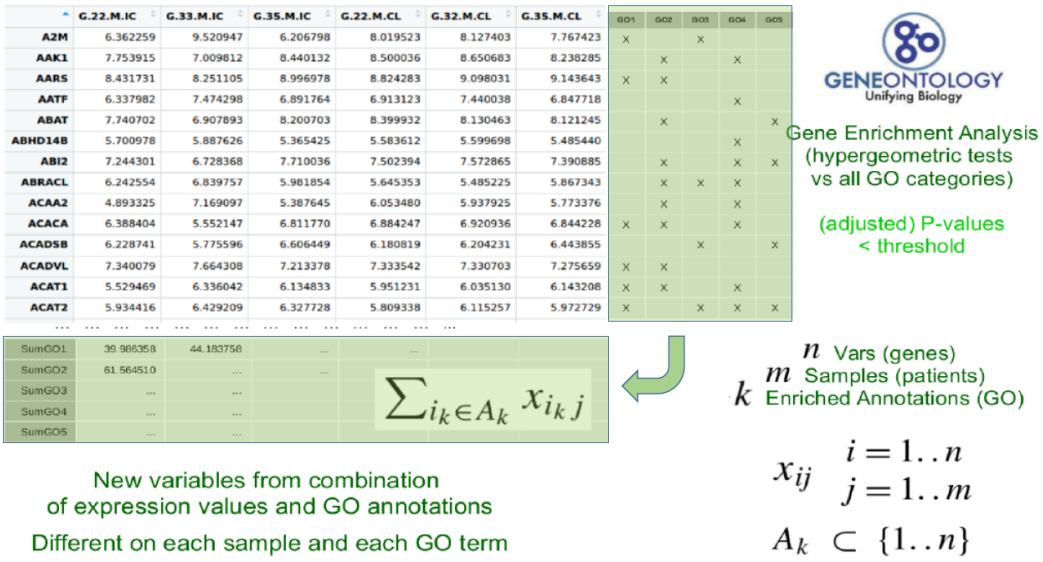

Figure 1.6: Addition of GO terms

Alternatively, manual annotations can be provided (eg. GO terms, canonical pathways, or even annotation to custom entities) as an optional input file.

[mostrar el format requerit].

Other annotation methods can be implemented, as functions to be used by the main pipeline, if more complex methods for biological information addition are required.

[Mostrar el format final de les anotacions, com a matrius dels data sets amb anotacions binàries 1/0 com a columnes extra]

EXPANSIO DE LES MATRIUS (numeritzar anotacions, creació de noves vars a partir de les anotacions)

The process starts already having a couple of data sets from distinct 'omics sources [punt de millora: admetre 3 o + inputs, comentar més tard a Discussion], mapped to gene ids (in the default case, where GO annotation have been performed), containing the results from a selection of differentially expressed genes or most relevant proteins analysis, or similar.

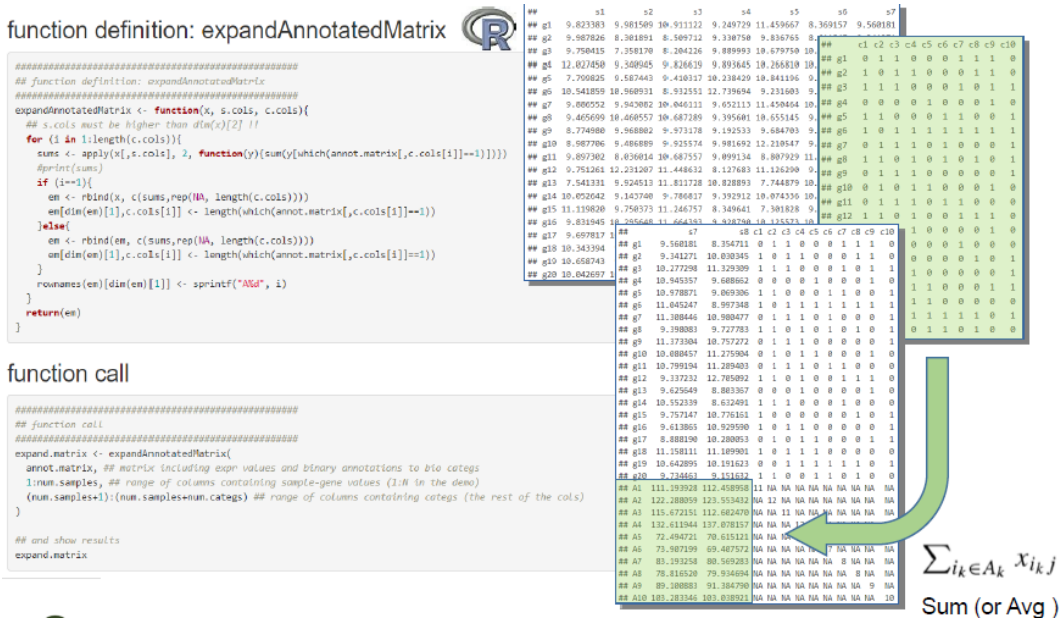

Figure 1.7: Addition of news feats

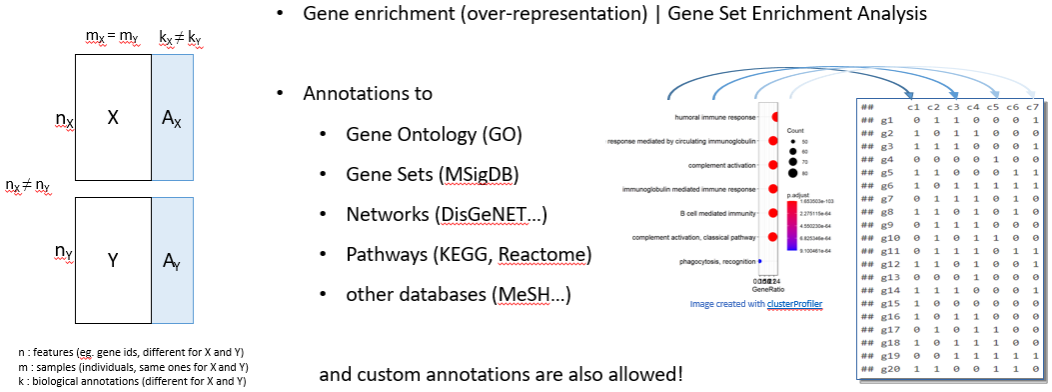

Figure 1.8: Gene enrichment diagram

1.2.3 Annotation Integration

Once the annotations are already computed, mapping each feature of the input data set to the corresponding biological entity, they can be used to generate new features (as new rows), computing the average value [punt de millora: funció de ponderació] of the expression/intensity values from all original features being mapped to the annotated biological entities.

Once we have the annotated matrices (Figure 1.9, highlighted in blue) we proceed to generate the Expanded matrices (in green) by casting these annotations as

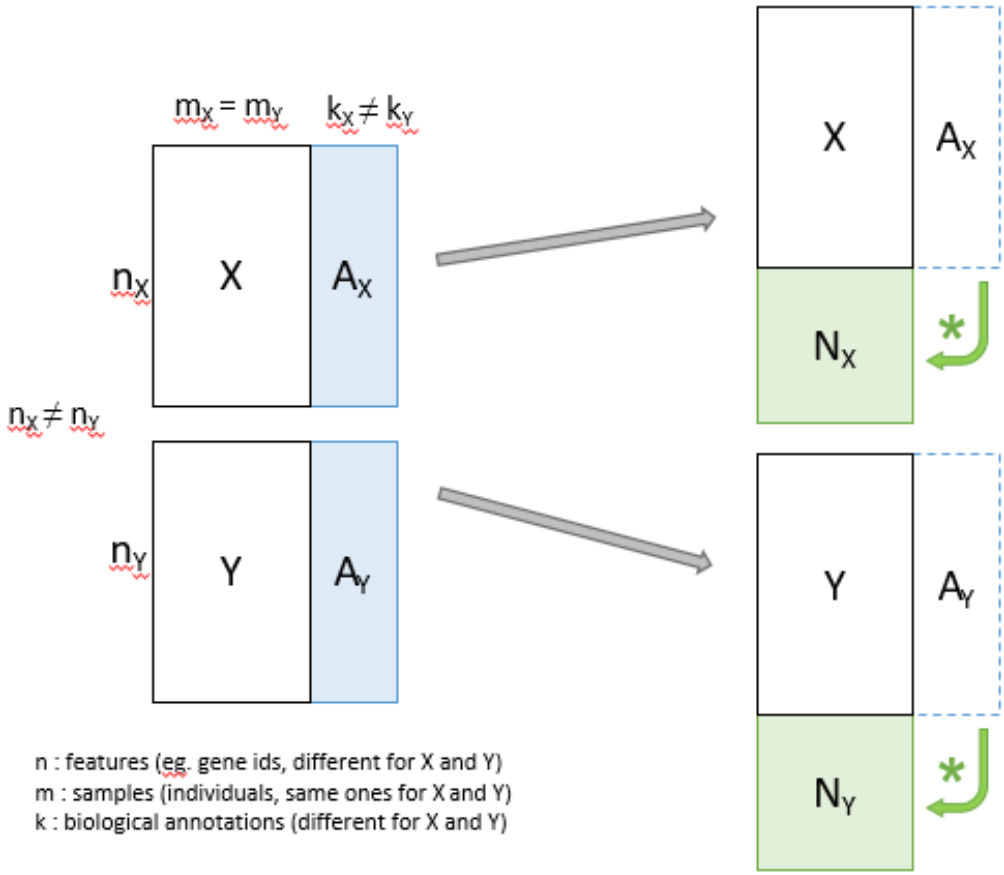

Figure 1.9: Matrix expansion diagram

numerical values, that is, calculating the average of the numerical expressions of each individual for the variables annotated to each category. This is done with the matrix product of the initial numerical values (expression, proteins...) with the transposed matrices of their annotations, and then with the inverse matrix of a diagonal matrix of the count of how many annotations each category or entity annotated has had.

1.3 Integrative Analysis with Joint Dimension Reduction Techniques

To uncover meaningful insights from the expanded data sets and extract relevant information from the integrated omics and biological knowledge, contrasted joint dimension reduction techniques were employed. These techniques enable the simultaneous analysis of multiple data types and facilitate the identification of

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$

$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

Figure 1.10: Addition of new feats (2)

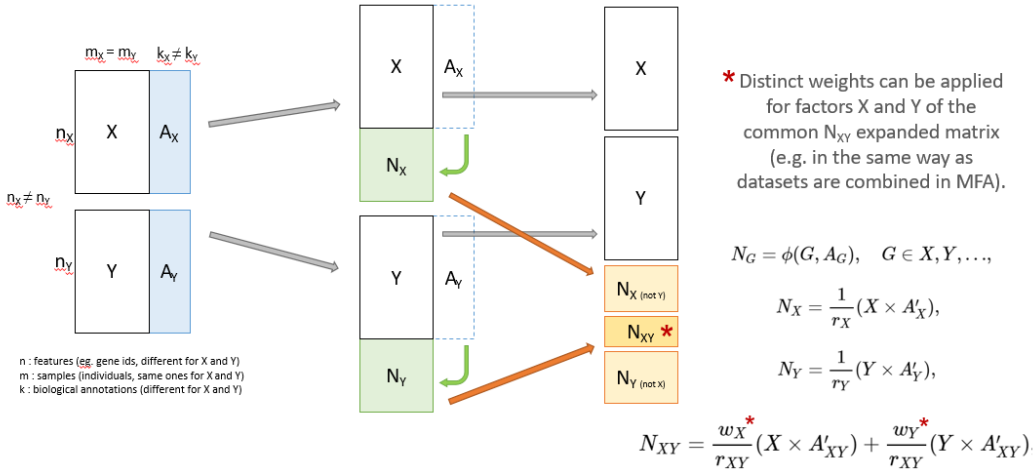

Figure 1.11: Matrix expansion diagram (2)

key patterns and relationships. The following methods were applied:

- Multiple Factor Analysis (MFA): MFA, adapted for multi-omics data, was utilized to identify sources of variability in the integrated dataset while considering both quantitative omics data and biological annotations. MFA aims to maximize relevant information within the data while accounting for the hierarchical structure of the biological knowledge.

- Multiple Co-Inertia Analysis (MCIA): MCIA, a technique that aligns the covariance structures of multiple datasets, was employed to explore relationships between omics measurements and biological annotations. MCIA seeks to identify common patterns and associations between these data sources.

- Regularized Generalized Canonical Correlation Analysis (RGCCA): RGCCA was used to identify latent variables that capture joint information from omics data and biological annotations. RGCCA extends canonical correlation analysis to handle multi-view data integration and helps reveal correlated features across datasets.

PUNTS A INCLOURE:

- Reducció de dimensió. Anàlisi factorial en detall (MFA), + MCIA + RGCCA

- incloure aquí % variabilitat explicat segons la estructura de la intersecció de les 2 taules (article Lovino 2022)
- avantatge del MFA és que podem definir blocs de variables!
- no mirem unicament si guanyem variabilitat, sino tambe si millorem interpretabilitat biologica

1.4 Semi-Automation using the Targets R Package

The semi-automation of the integrative analysis process was facilitated by leveraging the Targets R package, which provides an efficient and user-friendly framework for building and managing complex analysis pipelines. In the development of the Targets pipeline, careful management of functions and parameters was essential to ensure a systematic and reproducible workflow. The following principles were applied:

- Function Modularity: Functions within the Targets pipeline were designed to be modular, focusing on specific tasks or analyses. This modularity enhanced code readability and maintainability.
- Parameterization: Parameters for each function and analysis step were carefully defined, allowing for flexibility and adaptability in the pipeline. This parameterization enabled the adjustment of analysis settings without modifying the underlying code.
- Dependency Management: Dependencies between different analysis steps were explicitly defined within the pipeline. This ensured that each step was executed in the correct order, and dependencies were automatically managed by the Targets package.
- Error Handling: Error handling procedures were implemented to capture and address potential issues during pipeline execution. This included the ability

to handle errors, retries, and reporting of errors for troubleshooting. (NO APLICAT ARA PER ARA!)

PENDENT A AMPLIAR:

- Introduccio al paquet Targets en general i de les seves caracteristiques...

The R 'targets' package is a powerful tool for building and managing data science and data analysis pipelines. It is primarily designed for workflow automation, dependency management, and parallel processing in R projects. This package is useful for the following purposes:

1. Define and Manage Workflows: You can create a directed acyclic graph (DAG) that represents the workflow of your data analysis or machine learning project. Each node in the graph corresponds to a target, which can be a data file, an R script, or any other computational task.

2. Manage Dependencies: 'targets' allows you to specify dependencies between targets, ensuring that tasks are executed in the correct order. If a target depends on another target, it won't be executed until its dependencies are up-to-date.

3. Parallel Processing: One of the strengths of 'targets' is its ability to parallelize tasks. It can automatically determine which targets can be executed concurrently, improving the efficiency of your workflows, especially when working with large datasets or computationally intensive tasks.

4. Incremental Builds: When you make changes to your code or data, 'targets' can identify the minimal set of targets that need to be recomputed, saving time and computational resources. This is particularly useful for iterative development and experimentation.

5. Reports and Logging: 'targets' provides tools for generating reports and logging the progress of your workflow, making it easier to track and document your work.

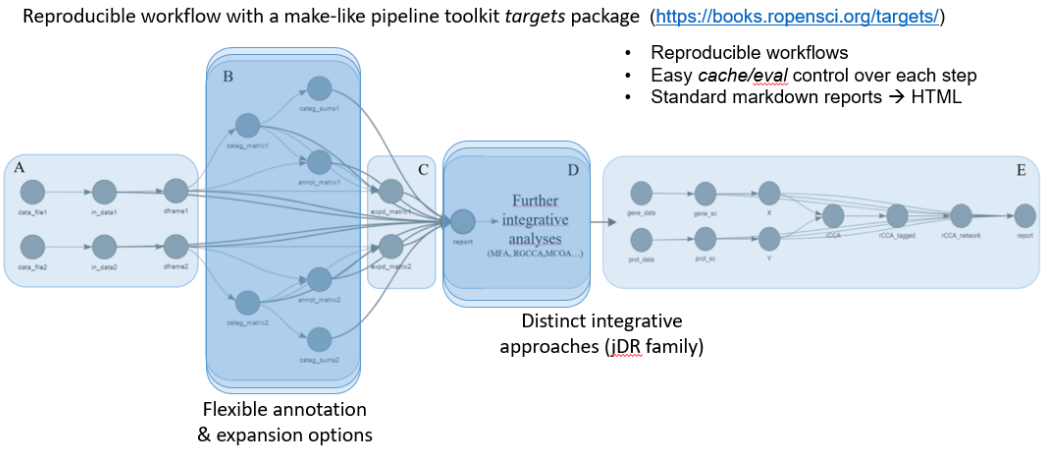

Figure 1.12: Workflow overview

6. Integration: It can be seamlessly integrated with other R packages and tools, such as 'drake' for more advanced data workflow management.

So, the 'targets' package is especially valuable for projects where data processing is a significant component, and you need a structured way to manage the various steps of your analysis or modeling pipeline. It helps ensure that your analyses are reproducible, efficient, and well-documented.

- Sistema que hem aplicat per crear el pipeline amb Targets...

Targets workflow diagram (Figure 1.12) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices). These are converted to R data frames with features in rows and samples in columns. Then, a data frame containing related annotations (B) is created, or loaded, for each given input matrix, and used to expand these original data, in order to end up with a pair of data frames (C) containing the original values plus the average expression/abundance values of the features related to each annotation as new features in additional rows. After that, distinct Dimension Reduction Methods are applied to perform the integrative analysis (D), and finally, an R markdown report (E) is rendered to show steps and main results of the full process.