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

In the context of multi-omics data integration, our proposal relies on the idea that incorporating biological annotations into datasets before proceeding with integrative analysis enriches the 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 whole process.

1.1 Data Format Review and Quality Assessment

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.

Data, whether obtained directly from TCGA or from specific data sets used as examples in specific R packages[data source: http://mixomics.org/mixdiablo/diablo-tcga-case-study/], has to be reviewed by performing a basic descriptive analysis, as is customary in single-omics data studies.

Data source: The Cancer Genome Atlas Network (Network et al., 2012) veure [@koboldt_comprehensive_2012] data original source: https://portal.gdc.cancer.gov/projects/TCGA-BRCA

BRCA subtypes main ref: https://www.pnas.org/doi/full/10.1073/pnas.191367098 veure[@sorlie_gene_2001]

The Cancer Genome Atlas, often abbreviated as TCGA, is a landmark project funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in the United States. It's a joint effort to comprehensively understand the molecular basis of cancer by using genome analysis technologies. TCGA curates a vast collection of real-world data, covering diverse cancer types. This data encompasses omics datasets, such as transcriptomics and proteomics, offering researchers a comprehensive perspective on the molecular landscape of tumors. Stringent quality control and standardization measures ensure the data within TCGA is compatible and readily integrable with other datasets. This

facilitates robust multi-omics analyses. This meticulous approach ensures data Compatibility, given that standardized formats eliminate inconsistencies between datasets generated using different platforms or technologies. This allows for seamless integration of data from various sources, crucial for multi-omics analyses like the one presented in this thesis. The integration of transcriptomics and proteomics data, heavily relies on the ability to combine these kind of datasets for a more holistic understanding. Standardized formats within TCGA remove a significant hurdle in this process. In addition to that, standardized data collection and processing protocols minimize technical variations that could introduce bias into the data. This enhances the reliability and generalizability of findings derived from TCGA records. The public availability of TCGA data empowers researchers worldwide to leverage this rich resource in their investigations and, consequently, this fosters open science collaborations and accelerates advancements in cancer research.

In the vast majority of cases, the information is structured directly in tables or matrices (where, for example, the columns contain patient samples or experimental individuals, while the rows represent the values of the measured features). These matrices can be encapsulated in structures such as Expression Set or similar, where information about the omic measurements is accompanied by metadata related to the samples themselves or to the details of the technology used for the analysis.

Samples in rows (200G, 111P); Features in columns (150S). Tenim 200 x 150 i 111(154 originalment) x 150.

It is relatively frequent that the datasets available for multi-omics studies present information on the same samples, analyzed by means of two or more different technologies (e.g., microarrays or RNA-seq for mRNA gene expression, plus quantification for proteins) while the information related to the omic molecules analyzed is, obviously, different. Not only that, but it is also common that there is no direct mapping between the different types of features, so that, for example, not all genes analyzed in an RNA expression experiment are unambiguously represented by their corresponding proteins.

The task of recognizing the labels of the molecules analyzed in each dataset and consequently determining which ones are suitable for proceeding with the integrative analysis is not a light one. It often requires to implement a semi-automatic general identification of their names or ID codes, followed by a validation and filtering of the resulting non-obvious cases. If, in addition, the integrative analysis aims to map the different omics in some way at the biological process level (e.g., microRNAs against their target genes), then we are faced with an additional challenge of critical importance for the rest of the process.

Mostrar Figure 1.2 i Figure 1.1 PERO POTSER MILLOR COM A TAULES INTEGRADES AMB MARKDOWN?

	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 1.1: Example of proteomics input data, viewed as a table in RStudio

```
##            A0FJ     A13E     A0G0     A0SX     A143     A0DA      A0B3     A0I2
## RTN2    4.362183 1.984492 1.727323 4.363996 2.447562 4.770798 3.3520618 1.810382
## NDRG2   7.533461 7.455194 8.079968 5.793750 7.158993 8.748061 5.0984040 3.791965
## CCDC113 3.956124 5.427623 2.227300 3.544866 4.691256 4.305401 0.5932056 2.719169
## FAM63A  4.457170 5.440957 5.543480 4.737114 4.808728 5.307480 5.2175851 4.355919
## ACADS   2.256817 4.028813 2.629855 4.269101 2.442135 3.239909 3.8851534 4.200249
## GMDS    6.017940 4.341692 6.363030 4.001104 7.029723 4.236539 5.9178858 4.830286
## HLA-H   5.006907 6.178668 6.039563 7.087633 5.936138 6.909727 8.0433411 9.130370
## SEMA4A  3.217812 2.864659 5.946028 5.007565 5.901459 6.591109 6.5328925 4.982386
## ETS2    4.734446 5.411029 5.651670 5.902449 6.641225 5.858016 6.3091167 5.304488
## LIMD2   5.099598 4.211397 3.304513 5.479451 5.508654 3.766283 4.1138727 5.149344
```

```
##                A0FJ        A13E        A0G0        A0SX        A143        A0DA         A0B3        A0I2
## YWHAE     0.04913078 -0.07998211 -0.03284989 -0.20532949  0.06019021  0.03076171 -0.107861537  0.64984396
## EIF4EBP1  0.44748623  0.60521842  0.89460973 -0.14132292  0.13176899  0.03299680 -0.037124691 -0.52148657
## TP53BP1   0.91783419  0.05910121  0.51704453 -0.31372867  0.33091238 -0.22027100 -0.544743061 -1.60203535
## ARAF      0.02274147 -0.45985298 -0.19182192 -0.07482347 -0.02435747  0.41861665  0.430503500 -0.18714658
## ACACA    -0.08626782 -0.59269183  0.41117190 -0.85148060  0.76975143 -0.71430870 -0.363474049  1.07761482
## ACCB     -0.41662442 -0.06226840  0.82582859 -0.66341044  0.87347870 -0.21752677 -0.269313837  1.58998239
## PRKAA1    0.28527039 -0.27523360  0.06774184  0.02956373 -0.21653182 -0.06306506 -0.077581092 -0.07753959
## ANLN      0.17231110  0.22210598  0.12199399  1.05494810  0.01378422  0.06025690  0.008872461 -0.05187936
## AR       -1.30760569 -1.62047596 -1.07789444 -1.26705469 -0.60132744 -1.20803848 -1.016297633 -0.42122691
## ARID1A    0.50509449  0.33958160  0.22718066  0.35529767  0.54412514 -0.11094480 -0.233223615 -0.35537533
```

Figure 1.2: Example of gene expression and protein quantification data, viewed as loaded arrays in R

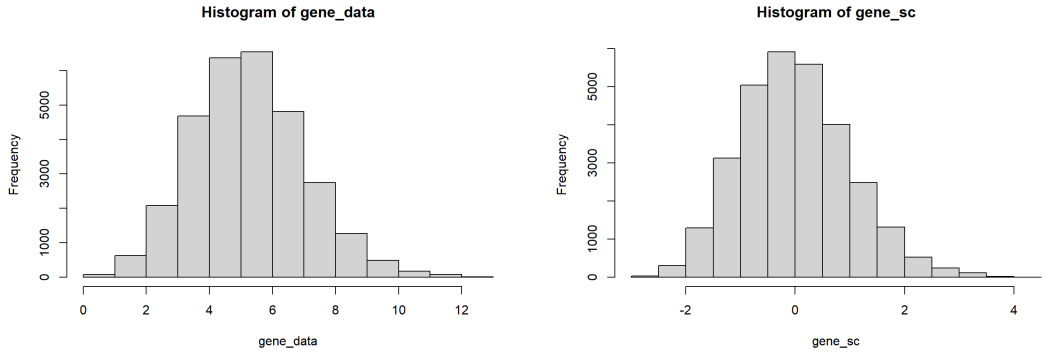


Figure 1.3: Histogram of the gene expression values coming from TCGA-BRCA dataset, before (left) and after (right) data centering

COMENTAR TAMBE REQUERIMENTS DE FORMAT (headers, value types...)

EXPLICAR QUALITY CHECKS APLICATS (grafiques per deteccio d'outliers, data centering...)

El proces s'ha de repetir, obviament, amb tots els input datasets.

1.2 Preprocessing for Integration of Biological Knowledge

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],[@thegeneontologyconsortium_gen 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 Datasets", 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:

- Selection of biological knowledge sources to feed annotations. Starting from the most commonly used biological knowledge databases in tasks of omics data analysis and interpretation, the goal is to access those that are most complete and applicable to the different types of data that must be annotated. GO and KEGG are excellent choices for that purpose.

- Selection of R packages specialized in the integration of biological information. The choice of the appropriate packages for the integration of biological information will depend on the specific needs of the project. In general, it is important to consider factors such as the type of data that will be integrated, the sources of the data, the integration methods that will be used and the desired level of complexity. In this case, it is appropriate to use R libraries that can work with gene and protein identifiers reliably and completely, without adding too much complexity to the process.

- Data-Annotation Mapping. Each omics dataset is mapped to the biological information collected based on its identifiers (for example, gene or protein

names) using the capabilities of the selected R packages. This step facilitates the relationship of each of the elements of the omics data with the biological knowledge entities, creating certain temporary objects that collect the information of these links. So this step allows to relate the elements of omics data with biological knowledge entities, such as genes, proteins, metabolic pathways, etc. The mapping is performed using the identifiers of the elements of the omics data. For example, genes can be identified by their name, their symbol, or their Ensembl ID. Proteins can be identified by their name, their sequence, or their UniProt ID.

- Annotation Integration. The most relevant annotation elements resulting from the previous step can be integrated into the matrix structure of the original omics dataset that has been used for its biological annotation, resulting in an expanded data matrix that combines the initial quantitative omics measurements with new values associated with the biological annotations obtained in the process. This step is implemented by executing new R functions specifically developed for this purpose. The resulting data matrices (which contain the integration of the most relevant biological annotations) are called 'Expanded Matrices', and will be the basis for the subsequent application of integrative analysis methods of the different omics analyzed.

1.2.1 Selection of the sources for biological annotation

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 1.4: List of gene symbols used as example

1.2.2 Selection of the annotation packages

PAS QUE ES FA PRACTICAMENT AL MATEIX TEMPS QUE L'ANTERIOR

Destacar criteris de fiabilitat i senzillesa

Apuntar llista o referencia principal important

1.2.3 Biological annotations mapping

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 dataset, if annotations are not already provided, two distinct basic annotation methods can be performed:

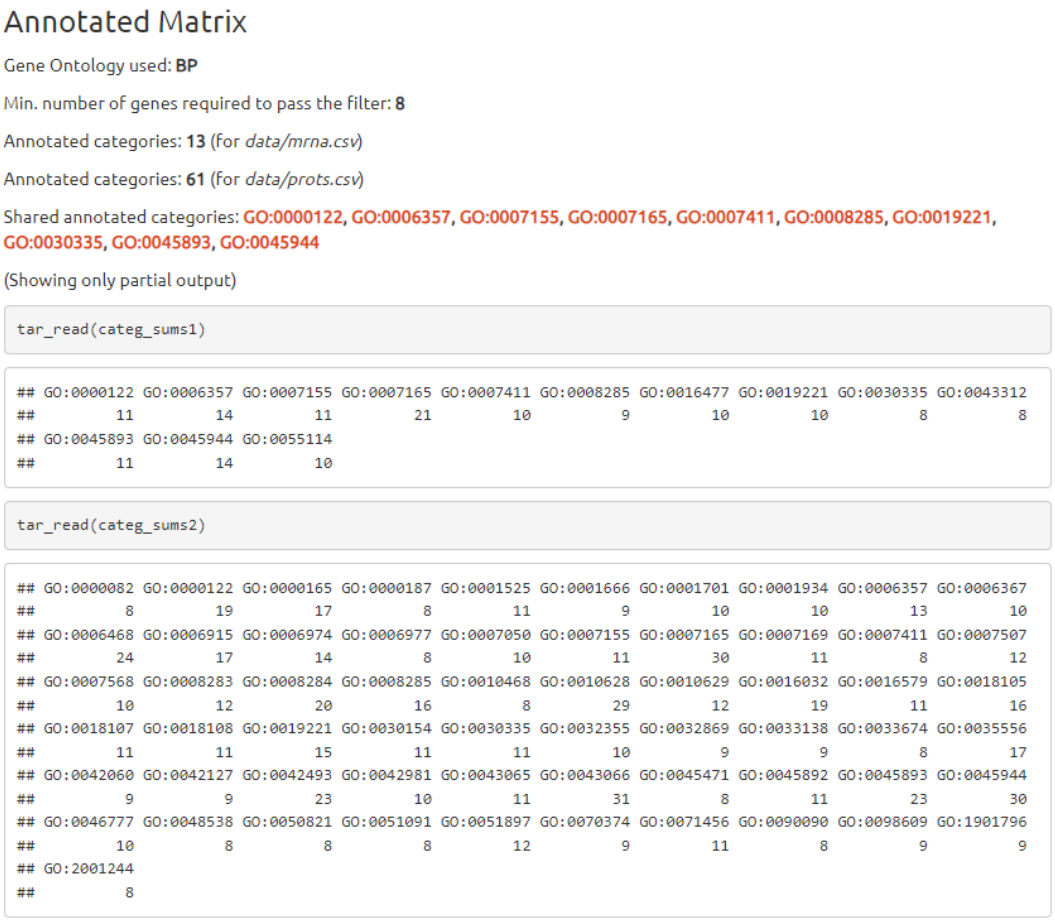
(i) a basic GO mapping, returning annotations to those GO entities for which we find more than a certain number of features (gene ids coming from our dataset, see Figure 1.4 for an example) annotated to them,

(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]



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 1.5: 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

es mostra exemple en Figure 1.5 POSSIBLE INTEGRAT EN MARKDOWN?

COMENTAR AQUI l'opció d'afegir les anotacions com a individus suplementaris enlloc de variables

Figure 1.7 is an example.]

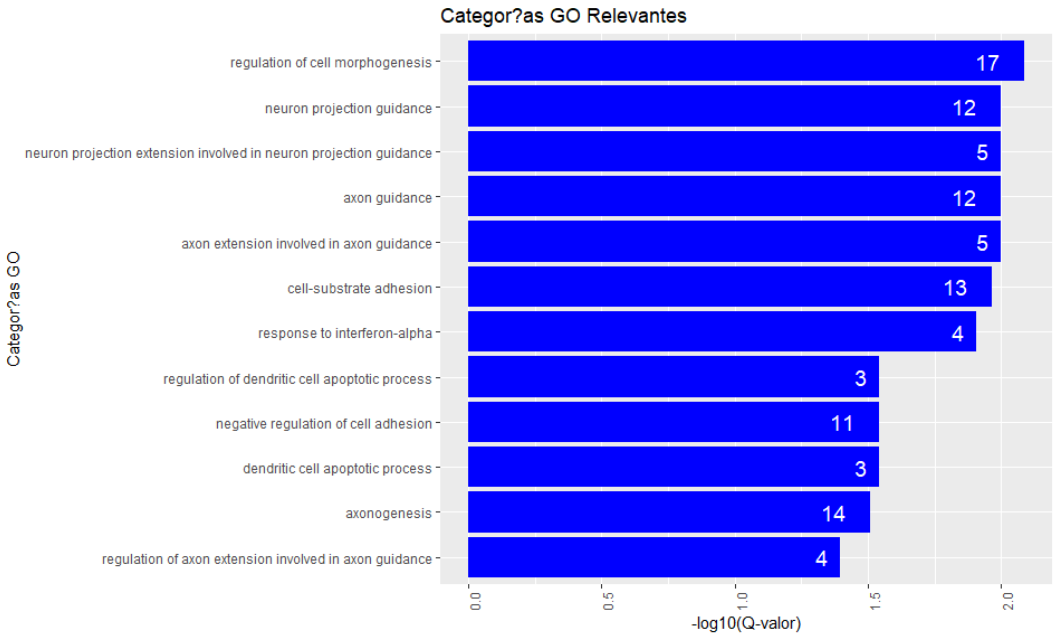


Figure 1.6: Example of results from GO annotation. Results of the biological significance analysis performed with the lists of genes against GO through clusterProfiler

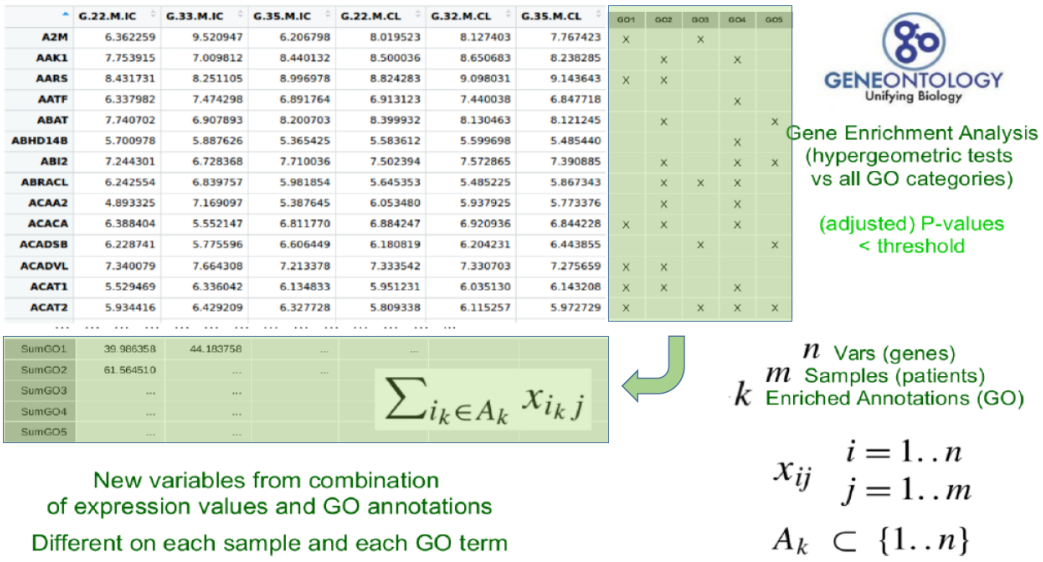


Figure 1.7: Addition of GO terms

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 datasets 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 datasets 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.

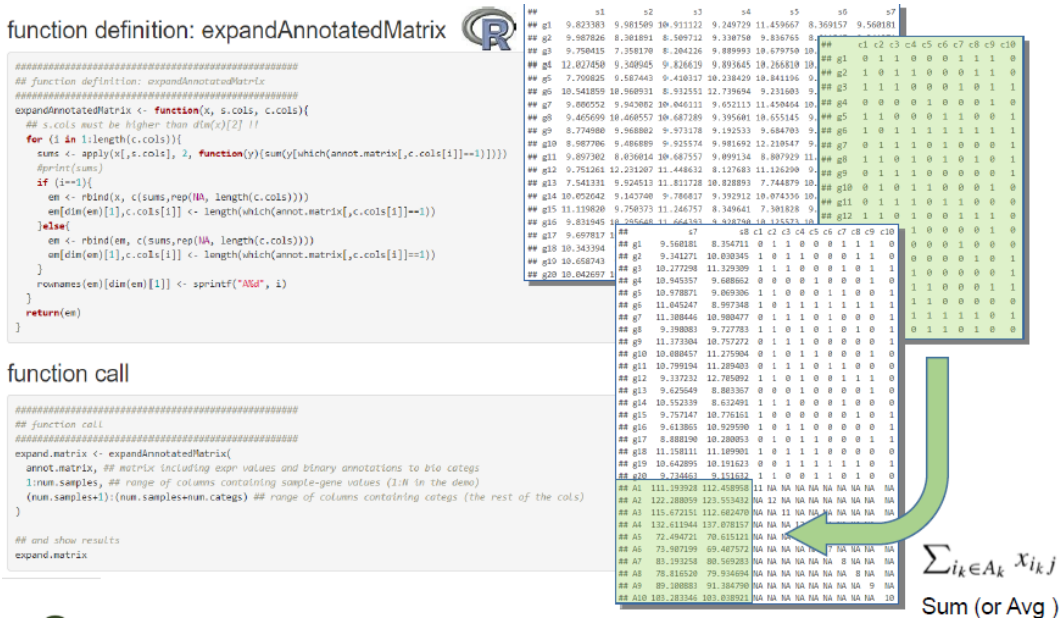


Figure 1.8: Addition of news feats

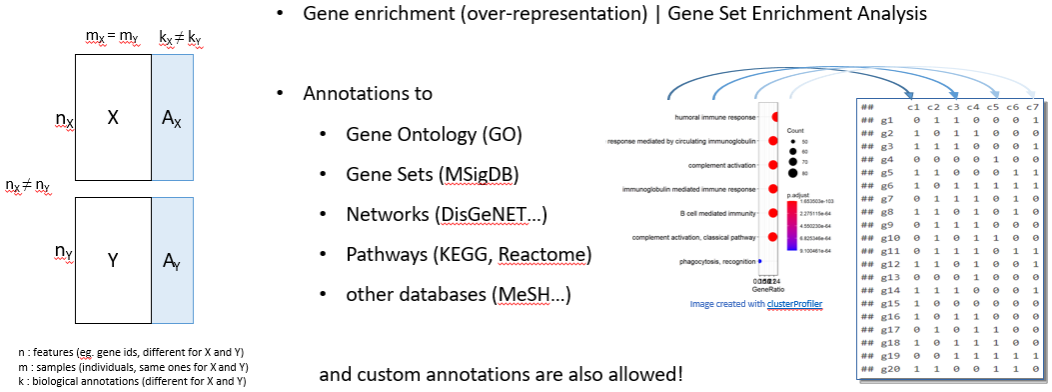


Figure 1.9: Gene enrichment diagram

1.2.4 Annotation Integration

Once the annotations are already computed, mapping each feature of the input dataset 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 1.10, highlighted in blue) we proceed to generate the Expanded matrices (in green) by casting these annotations as

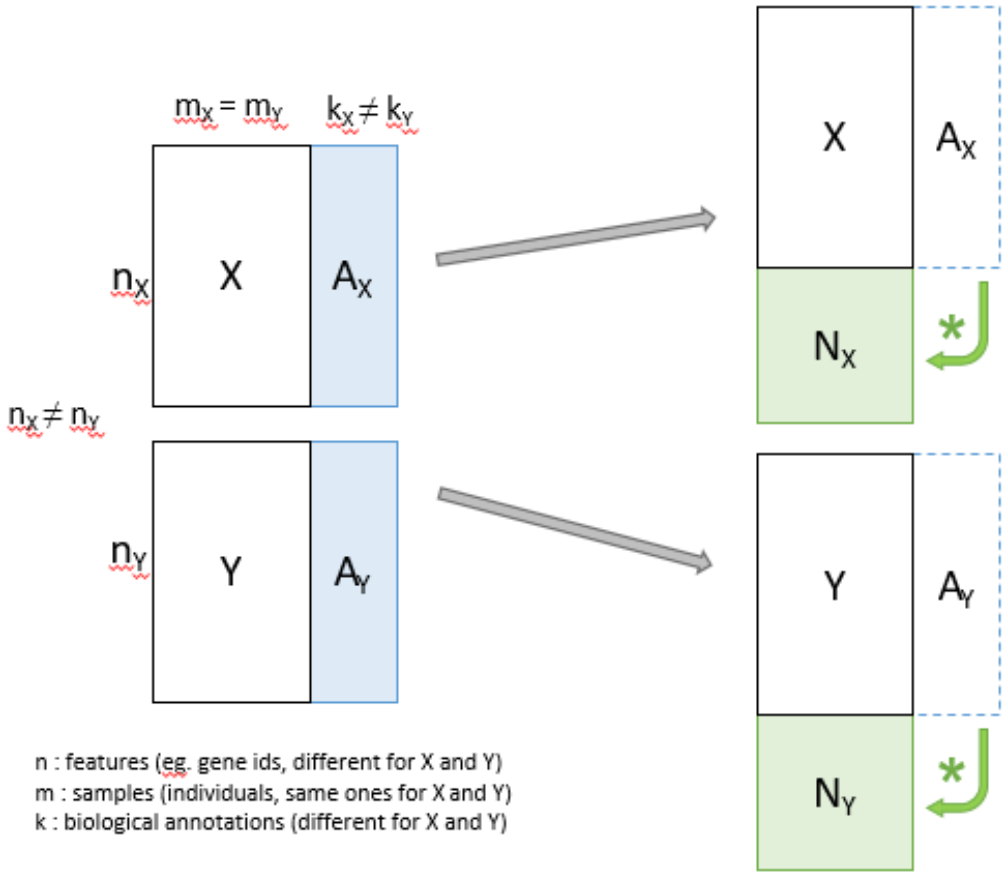


Figure 1.10: Matrix expansion diagram

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 datasets 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 1.11: Addition of new feats (2)

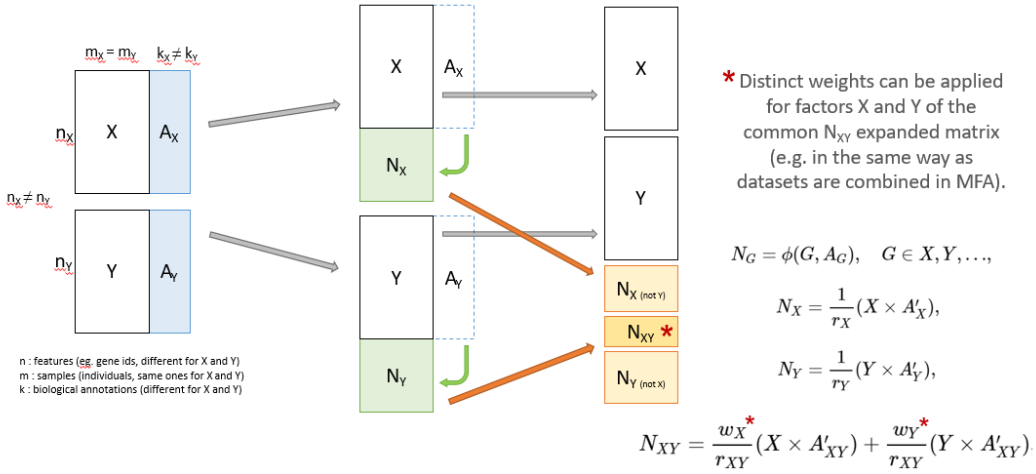


Figure 1.12: Matrix expansion diagram (2)

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 data sets.

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 [@lovino_survey_2021]
- 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 data sets 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.

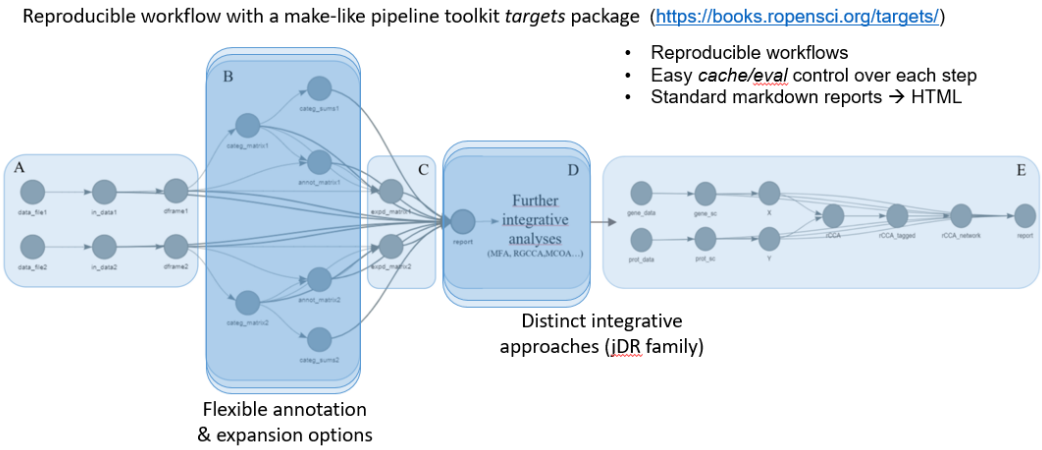


Figure 1.13: Workflow overview

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 1.13) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input datasets (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.