***TASKS FOR PRESENTING***

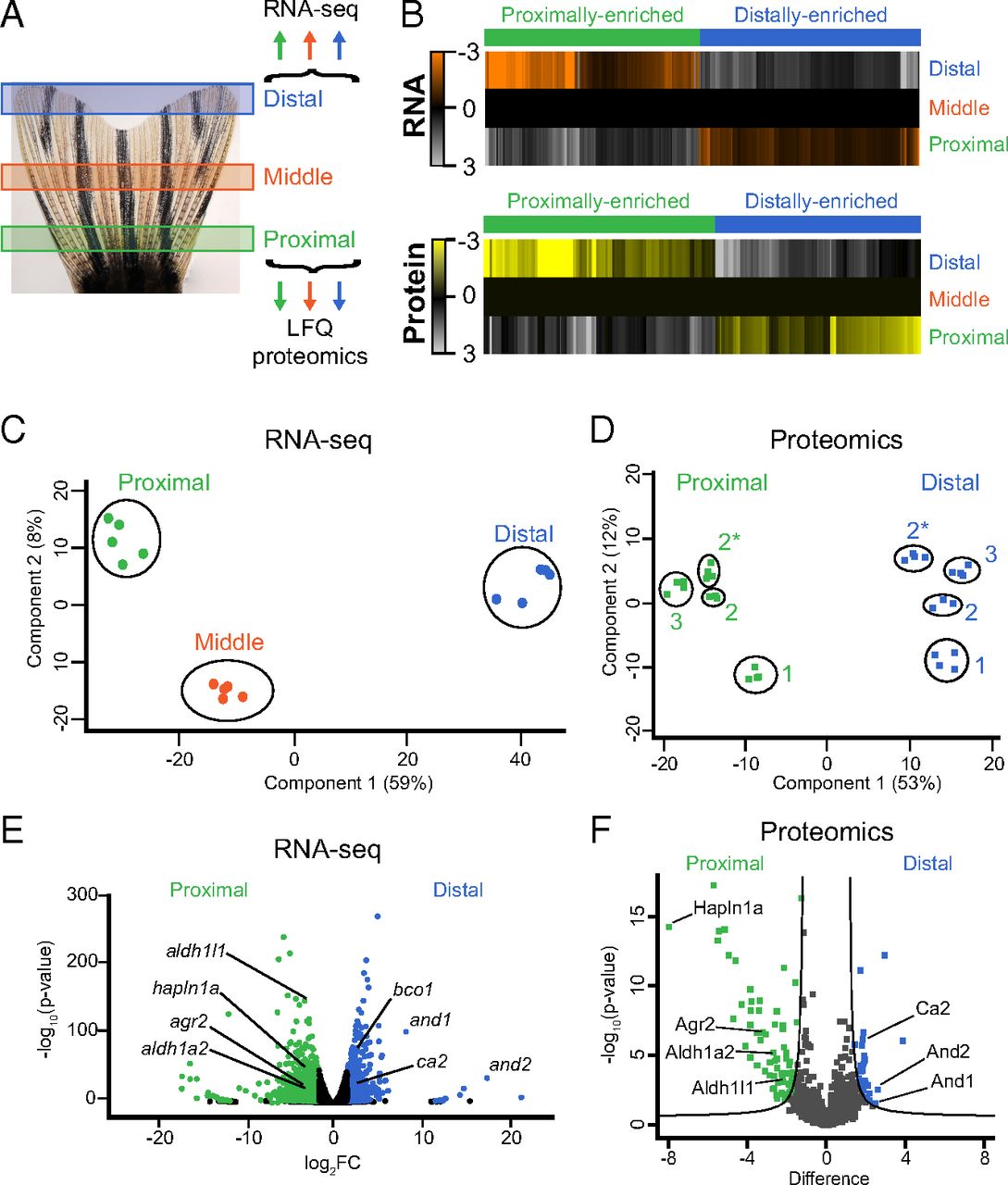
**Exploring transcriptomic, proteomic, and metabolomic landscape of positional memory in the caudal fin of zebrafish**

***Hint:****The tasks here are just for your reference of what you may want do to draw a conclusion based on analyses of the data provided. You can do whatever you think is appropriate to achieve the goal.*

**Problem Scenario**

In vertebrates, proper patterning during appendage regeneration is regulated by positional memory - a cellular property hypothesized to rely on gradients of molecules present in uninjured limbs. Only very few genes have been shown to regulate positional memory and be expressed in a gradient in the uninjured limb. To identify new candidate effectors of positional memory, Rabinowitz et al. mapped the abundance of transcripts, proteins, and metabolites along the uninjured zebrafish tail fin to generate a high-confidence list of genes and metabolites as candidate effectors of positional memory in zebrafish.

It was hypothesized that molecular effectors of positional memory are expressed in gradients along the proximodistal axis of uninjured appendages. To identify candidate molecules that might be involved in this process, the authors performed RNA sequencing (RNA-seq) and label-free quantification (LFQ) proteomics on proximal, middle, and distal regions of uninjured zebrafish caudal fins (Fig.1A) and quantified 23,926 transcripts and 3,061 proteins respectively.

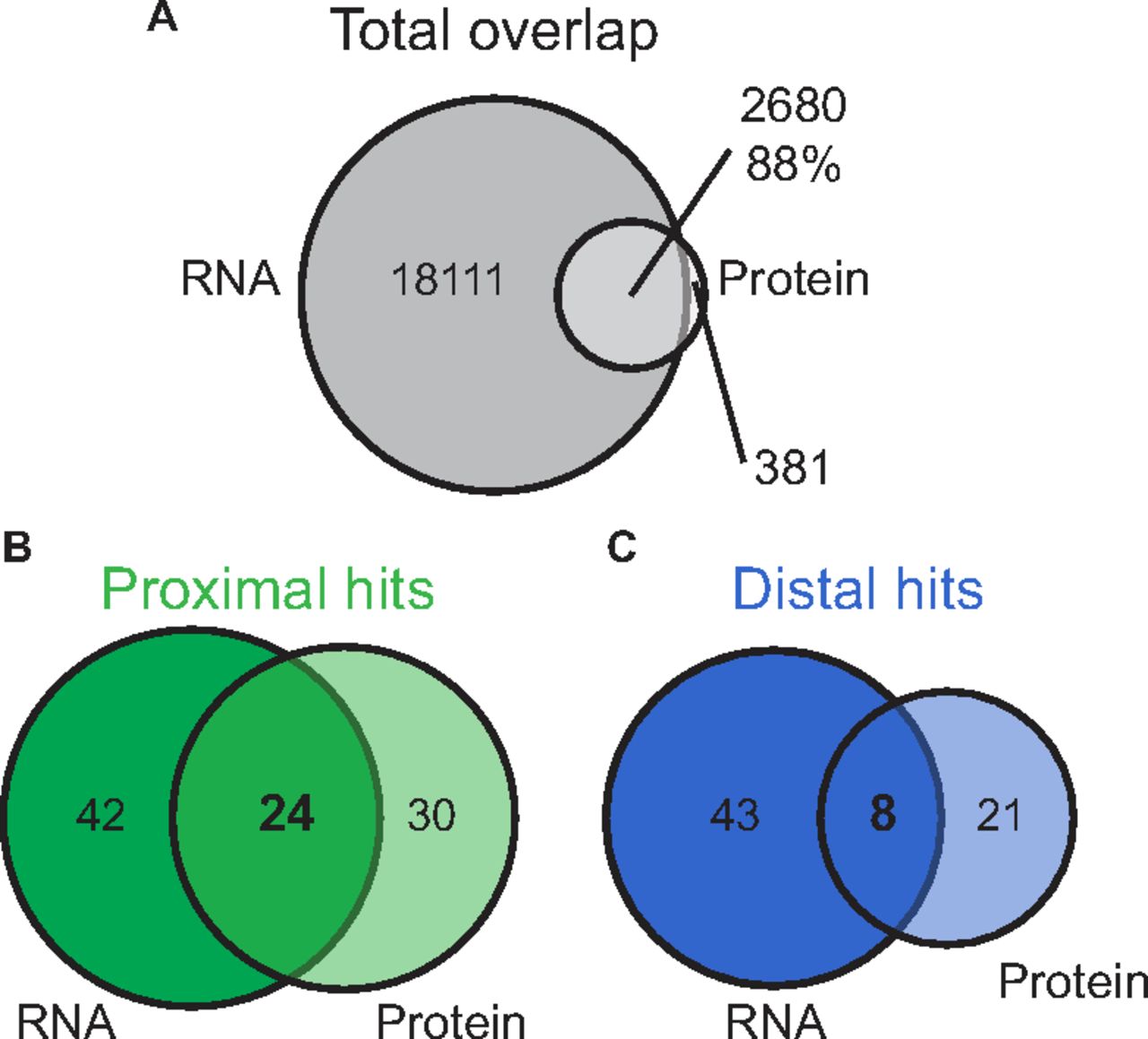


***Fig.1A |*** Illustration of the three regions of the fin that were harvested for RNA-seq or LFQ proteomics (see also Fig. S1).

1. **Identifying candidate genes whose levels differ along the proximodistal axis**

To generate a high confidence list of molecules whose levels differ along the proximodistal axis, the proximal and distal RNA-seq and LFQ proteomics datasets were compared. The analysis was restricted to transcripts and proteins present in similar patterns in both datasets. The RNA-seq screen was more comprehensive and quantified 88% of the proteins quantified by LFQ proteomics. Conversely, LFQ proteomics only quantified 13% of the transcripts quantified in the RNA-seq data (Fig.S3A). Transcripts were considered differentially expressed at FDR < 1% and fold change > 3 between proximal and distal regions, whereas proteins were considered differentially expressed at FDR < 5% and fold change > 2. Identified and quantified transcripts and proteins were assigned according to gene symbol.

***Fig.S3A |*** RNA-seq identified and quantified 20,791 transcripts, whereas LFQ proteomics identified and quantified 3,061 proteins.



As result, among the transcripts measured in both experiments, 66 and 51 transcripts were proximally and distally enriched respectively. Ways to identify significant genes from RNA-seq data were introduced in details previously in BIO211 Lab4.

***Task 1.1***

*Briefly explain how proximally and distally enriched transcripts were identified. Quantification data, including FoldChange and FDR, of transcripts can be found in PvD\_trans.csv. The complete set of gene symbols of quantified proteins can be found in PvD\_combine.csv.*

***Notes:***

* *Any ways that lead to the results are acceptable.*
* *Multiple transcripts may correspond to the same gene symbol.*

***Task 1.2***

*Identify proximally and distally enriched proteins that are measured in both experiments. An example R script to achieve this can be found in the following box. Report your key results and briefly explain your findings.*

# Read data

path <- '/Users/xin.liu/Desktop/BIO316/xl/'

setwd(paste(path))

rm(list = ls())

protein <- read.csv('PvD\_combine.csv', header = T, sep = ',', stringsAsFactors = F,

quote = "", row.names = 1)

LFQ <- protein[, -1]

P\_D <- 2^LFQ

# DEP analysis

pvalue <- apply(P\_D[, c(1:30)], 1, function(x) {

a <- factor(c(rep("Distal", 15), rep("Proximal", 15)))

t.test(x~a, var.equal = T, alternative = "two.sided", conf.level = 0.95)

})

result\_t.test <- data.frame(gene\_symbol=protein$Gene.Name,

Pvalue = as.numeric(unlist(lapply(pvalue, function(x) x$p.value))),

log2FC = log2(as.numeric(unlist(lapply(pvalue,

function(x) x$estimate[1]/x$estimate[2]))))

)

rownames(result\_t.test) <- rownames(protein)

t.test\_fdr <- p.adjust(result\_t.test$Pvalue, method = 'fdr')

result\_t.test\_fdr <- cbind(result\_t.test, t.test\_fdr)

dep <- result\_t.test\_fdr[result\_t.test\_fdr$t.test\_fdr < 0.05 & abs(result\_t.test\_fdr$log2FC) > 1, ]

dep1 <- dep[order(abs(dep$log2FC), decreasing=T),]

write.table(dep1, file = 'dep\_fdr\_raw.txt', sep = '\t',quote = F, row.names = F)

***Task 1.3***

*In statistics, a volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets. It plots significance versus fold-change on the y and x axes, respectively. More detailed explanation about Volcano plots and an example R script can be found in BIO211 Lab4. Generate a Volcano plot for the quantified proteins. Report key results and briefly explain your findings. Another example R script can be found in the following box.*

## Volcano plot

install.packages('ggplot2')

install.packages('cowplot')

install.packages('patchwork')

install.packages('ggplotify')

install.packages("ggrepel")

library(cowplot)

library(patchwork)

library(ggplotify)

library(ggplot2)

library(ggrepel)

logFC\_cutoff <- 1

result\_t.test\_fdr <- result\_t.test\_fdr[which (result\_t.test\_fdr$gene\_symbol != ""), ]

# omit rows with unknown ID

anno <- read.csv('volcano\_anno.csv', header = T, sep = ',', stringsAsFactors = F,

quote = "", row.names = 1)

result\_t.test\_fdr$change = as.factor(

ifelse (result\_t.test\_fdr$t.test\_fdr < 0.05 & abs(result\_t.test\_fdr$log2FC) > logFC\_cutoff,

ifelse (result\_t.test\_fdr$log2FC > logFC\_cutoff,"UP","DOWN"),"NOT")

)

up\_down <- table(result\_t.test\_fdr$change)

loc\_up <- which(result\_t.test\_fdr$change == 'UP')

loc\_down <- which(result\_t.test\_fdr$change == 'DOWN')

significant <- rep('normal', times=nrow(result\_t.test\_fdr))

significant[loc\_up] <- 'up'

significant[loc\_down] <- 'down'

significant <- factor(significant, levels = c('up', 'down', 'normal'))

result\_t.test\_fdr\_new = cbind(result\_t.test\_fdr, anno)

p <- qplot(x = result\_t.test\_fdr\_new$log2FC,

y = -log10(result\_t.test\_fdr\_new$t.test\_fdr),

xlab = 'log2(FC)', ylab = '-log10(FDR)',

size=I(2), alpha = I(1/3), colour=significant)

p <- p+scale\_color\_manual(values = c('up'='red', 'normal' = 'gray', 'down'='blue'))

# Add cutoff line

xline = c(-logFC\_cutoff,logFC\_cutoff)

p <- p+geom\_vline(xintercept = xline, lty = 2, size = I(0.2), color = 'grey11')

yline = -log(0.05,10)

p <- p+geom\_hline(yintercept = yline, lty = 2, size = I(0.2), color = 'grey11')

# Label gene names of interest

p <- p+geom\_text\_repel(aes(label = result\_t.test\_fdr\_new$label),

point.padding = unit(0.25, "lines"),

arrow = arrow(length = unit(0.01, "npc")),

nudge\_y = 0.1)

p <- p+theme\_bw() + theme(panel.background = element\_rect(colour = 'black',

size = 1, fill = 'white'),

panel.grid = element\_blank(),

axis.title.x = element\_text(size = 15) ,

axis.title.y = element\_text(size = 15),

axis.text.x = element\_text(size = 15),

axis.text.y = element\_text(size = 15),

legend.title = element\_text(size = 12),

legend.text = element\_text(size = 10))

ggsave('volcano\_dep.pdf')

***Task 1.4***

*Compare the proximal and distal differentially expressed transcripts and proteins, restricted to the 2,680 commonly measured by both screens. How many genes were proximally and distally enriched for both transcript and protein? Briefly explain how you did the analysis and report the key results* *e.g. numbers, Venn diagrams, el al.*

1. **Evidence for possible multifactorial contributions to positional memory**

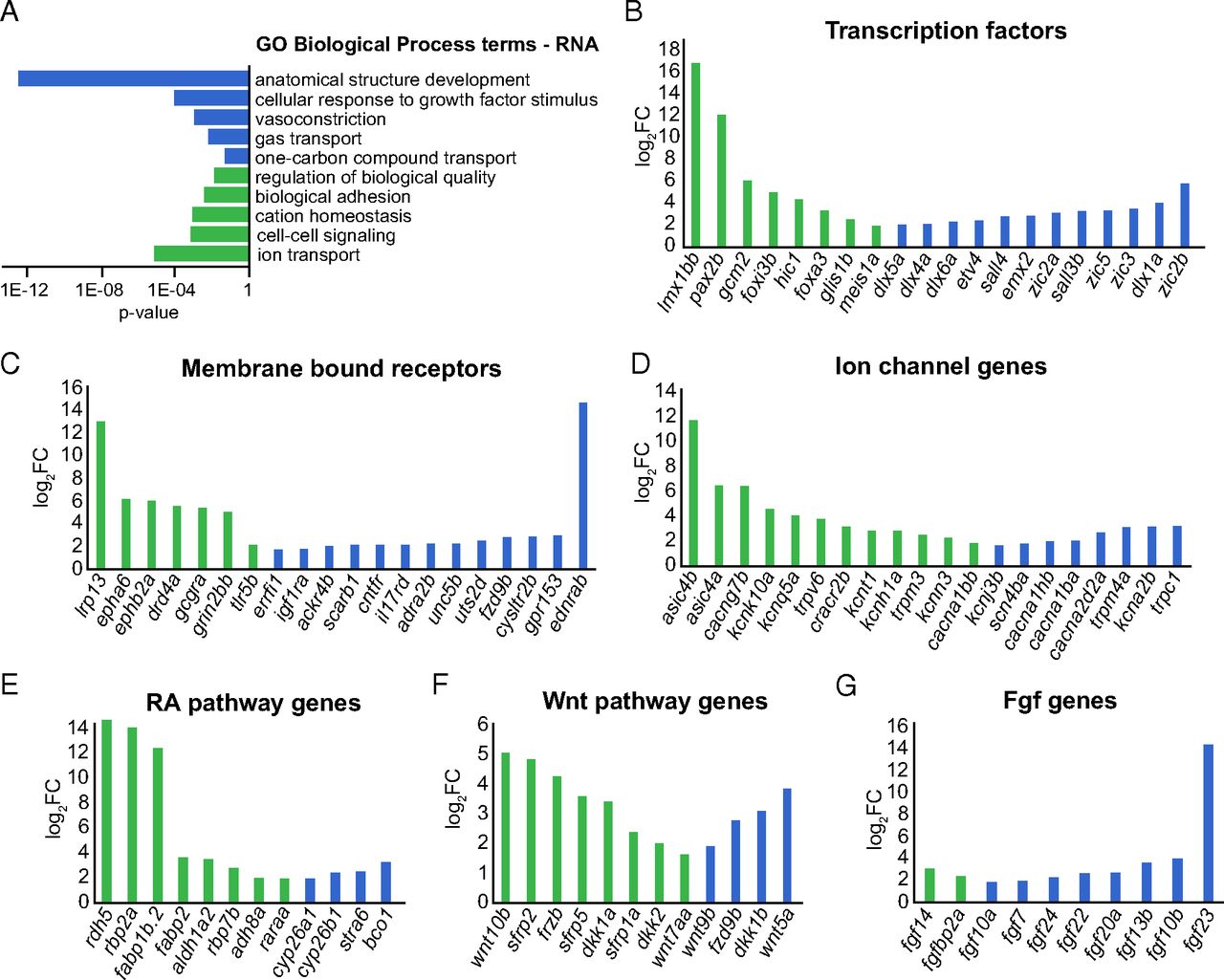
Experimental manipulation of effectors of positional memory was predicted to induce abnormal patterning during regeneration. Therefore, identifying the types of genes that are differentially expressed along the proximodistal axis of the caudal fin may shed light on the discovery of new candidate effectors of positional memory. This can be achieved by GO and KEGG pathway enrichment analysis.

To be concise, the **Gene Ontology (GO)[[1]](#footnote-1)** is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species, whereas **Kyoto Encyclopedia of Genes and Genomes (KEGG)[[2]](#footnote-2)** is a database resource for understanding high-level functions and utilities of the biological system from molecular-level information. **Enrichment** analysis identifies which pathways, diseases and GO terms are statistically significantly associated with the genes/proteins of interest. More details were provided during the lectures and can also be found in BIO211 Lab4.

***Task 2***

*Perform GO and/or KEGG pathway enrichment analysis to identify any types of genes/specific gene(s) that may be effectors of positional memory. Partially save your results and explain your findings. There is no best answer; you will be scored as long as your explanation makes sense.*

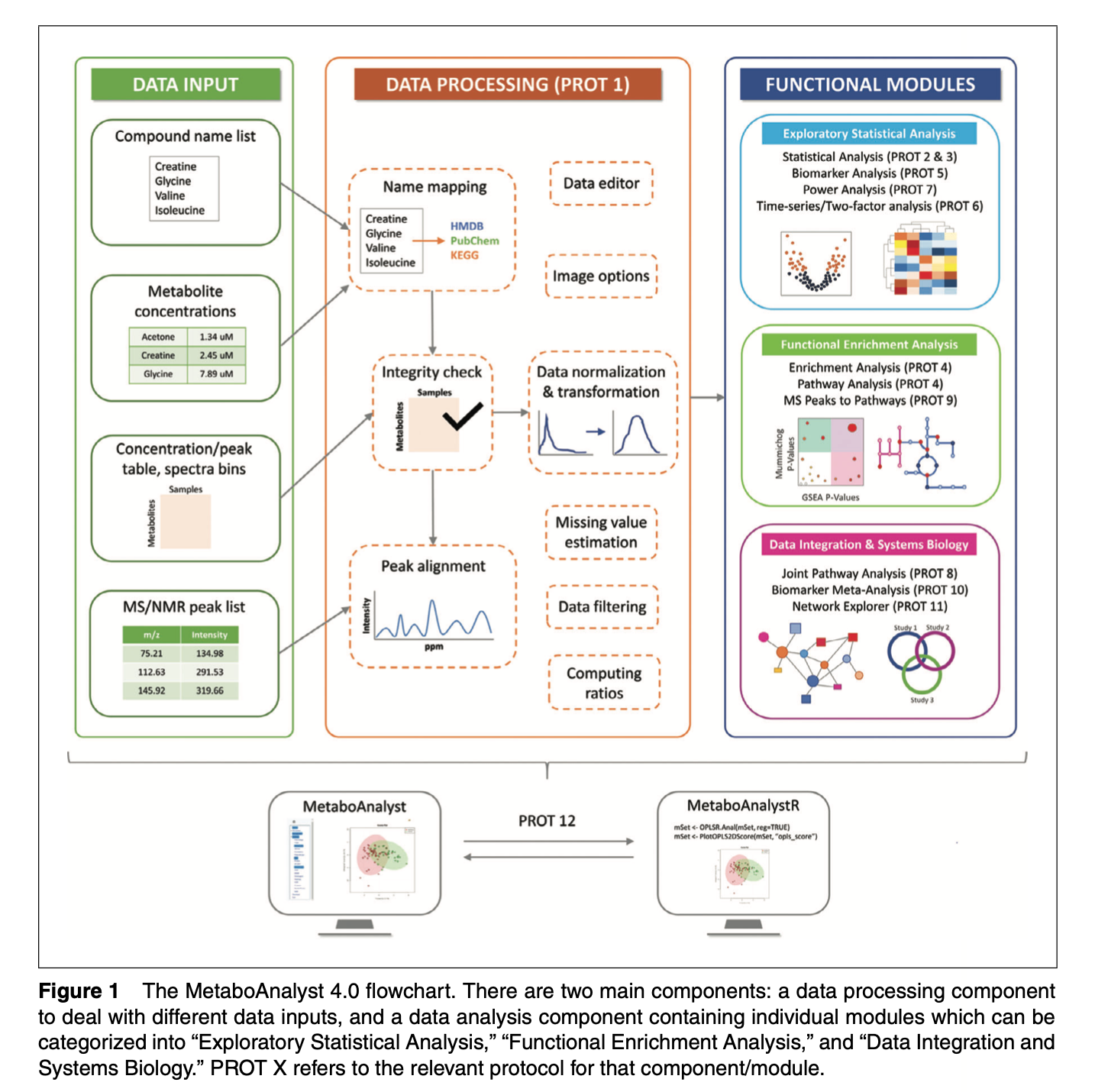
***One example:*** *Annotations for GO Biological Process terms were performed for transcripts. Part of the statistically significant GO terms with P < 0.05 were summarized in the following figures. One may conclude from the results that transcription factors are master regulators of cell identity and, thus, are top candidates for effectors of positional memory. Proximally enriched (green) and distally enriched (blue) transcripts identified by RNA-seq (FDR < 1% and fold change > 3) for transcription factors were also summarized below. Of the 693 transcripts encoding transcription factors (Dataset S7) quantified in the RNA-seq data, several were differentially expressed between proximal and distal regions of the fin. Examples of these transcription factors include dlx5a, dlx6a, meis1a, msx1a, hoxb13a, raraa, and lmx1bb. They are directly implicated in patterning developing appendages, consistent with the prediction that mechanisms from development may regulate positional memory in adult structures.*

**

1. **Relative metabolite abundance measured along proximodistal axis**

In addition to genetically encoded cues, metabolic products may contribute to positional memory. For example, vitamin D metabolism is important for establishing anteroposterior positional information in the pectoral fin. From the analysis results of RNA-seq and LFQ data, one may find that several metabolic enzymes, including *aldh1l1*, were differentially expressed along the proximodistal axis of caudal fins. Moreover, GO term analysis of proximally enriched proteins also identified proteins related to amino acid metabolism and lipid transport.

To explore metabolites that may be involved to function as effectors of positional memory, the authors used mass spectrometry to profile metabolites in proximal, middle, and distal regions of the uninjured caudal fin and were able to quantify 151 metabolites in individual fish fin samples (see P\_D\_meta.csv where we only focus on proximal and distal regions). The goal here is, by using methods provided in MetaboAnalyst (modules and general workflow are summarized in the following figure), to identify metabolites that are significantly different between these two groups (proximal vs. distal regions of the uninjured caudal fin). These metabolites could serve as potential effectors for maintaining proximal tissues in a metabolically poised state for more rapid proliferation.



***Task 3.1***

*For your convenience, the raw data were pre-processed and normalized. Only quantification of metabolites of proximal and distal regions of the uninjured caudal fin were retained to form P\_D\_meta.csv for further analysis. By MetaboAnalyst, try to find metabolite candidates consistently identified by both univariate analysis and PLS-DA analysis.*

***Notes:*** *Data were previously normalized i.e. no additional sample normalization or data transformation or scaling is necessary.*

Metabolomics is increasingly applied together with other omics technologies such as transcriptomics, proteomics, and metagenomics to gain functional insight into complex diseases/conditions. However, interpretation of multi-omics data at a systems level remains a significant challenge. A common strategy is to analyze each set of omics data individually, and then piece together the “big picture” using the significant features (genes, proteins, metabolites, etc.) identified from individual omics analysis. Biological networks are useful and flexible means to represent our knowledge at a systems level. By harnessing the power of networks and a priori biological knowledge, lists of significant features can be co-projected onto knowledge-based networks to reveal important links between the molecules of interest. The networks can also be used to identify their associations with diseases or other interesting phenotypes. Network visualization of datasets can be employed to gain novel insights or assist users with the development of new hypotheses.

***Task 3.2***

*For the consensus results, identify their biological function(s) and possible roles in pathways using MetaboAnalyst Pathway Analysis module. Report key steps and results of your analysis and explain your findings.*

***Notes:*** *KEGG pathways for zebrafish are available in the MetaboAnalyst pathway library.*

1. **Integrated multi-omics interpretation of molecular contributions to positional memory**

A lot more omics data are accumulating, showing the regulatory complexity of biological phenotypes. This regulatory complexity needs systems-level approaches to understand it. There are a series of layers in the omics field that are closely connected to each other as described in the “central dogma”. Therefore, we have to interpret each single omics layer and also to systematically integrate multi-omics layers to get a full picture of the regulatory landscape of the biological phenotype. While each particular omics approach has an underlying biological network for applying appropriate systematic analysis, a full regulatory landscape can only be obtained when multi-omics data are combined and incorporated within adequate networks.

The “Joint Pathway Analysis” module in MetaboAnalyst has been designed to support multi-omics analysis and interpretation (i.e., metabolomics, transcriptomics or proteomics). In particular, it allows users to map significant genes/proteins (identified from gene expression or proteomics studies) together with significant metabolites (identified from metabolomics studies) to metabolic pathways for functional enrichment analysis and pathway topology analysis. The working assumption behind this module is that by integrating evidence based on changes in both gene/protein expression and metabolite concentrations, one is more likely to pinpoint the pathways involved in the underlying biological processes.

***Task 4***

*Perform Joint Pathway Analysis on the RNA-seq/LFQ proteomics/metabolomics data produced in this project to seek integrated molecular clue(s) of positional memory at the systems level. Report key results and explain your findings. Detailed introduction to the usage of the Joint Pathway Analysis module can be found on the MetaboAnalyst website.*

1. <http://geneontology.org/> [↑](#footnote-ref-1)
2. <https://www.kegg.jp/kegg/pathway.html> [↑](#footnote-ref-2)