Specifications for multi-omics M2 project

WRITTEN BY

THE PROJECT MANAGERS: VICTOR DEGUISE, ZAKARIA TOUGUI, FRANÇOIS VICTOR

FOR

THE CONTRACTING AUTHORITY REPRESENTED BY

MOURAD BEKHOUCHE

UNDER THE SUPERVISION OF

CÉCILE HILPERT

OCTOBER 14, 2022

Table of contents

1 Topic presentation

1.1 Context

The tooth is composed, from the outside to the inside, of the enamel, which is mineralized at 98%, the dentine deposited by the odontoblasts, mineralized at 70%, and, within the endodontic space, we can find the dental pulp (DP). The DP is responsible for the tooth's vitality, pain, immunity, repair mechanisms and regeneration. The main molecular constituents of the DP's connective tissue are type I and type III collagen as well as glycosaminoglycans (GAGs) and proteoglycans.

In case of infections, which can occur with caries disease, we usually observe a degradation of enamel and dentine that can lead to an inflammatory response, which in turn, leads to an increase in blood pressure in the DP, compression of blood vessels, cessation of blood flow and eventually DP necrosis. The intervention of a dental practitioner is required to remove the inflamed tissue, disinfect and fill the endodontic space with tight biomaterial.

However, the tooth is more likely to undergo fractures or reinfections, which are usually more severe than the previous one, due to the lack of an immune response. Re-infections need further treatments which affects the quality of life of patients. In order to prevent this, tissular engineering approaches have been proposed in order to design an antibacterial fibrin-based scaffold hydrogel that could enable DP regeneration within the endodontic space. However, the uncontrolled mechanisms of tissue regeneration lead to the formation of dysfunctional tissues (fibrotic or calcified) at preclinical stages. The lack of knowledge concerning the mechanisms of tissue regeneration is a critical scientific and clinic lock.

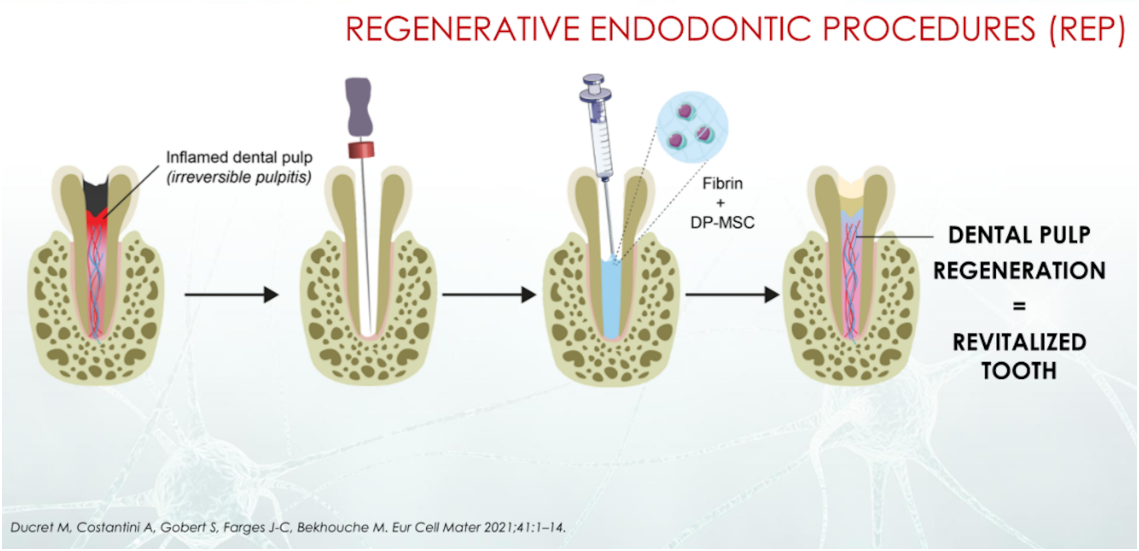

Figure 1: Regenerative endodontic procedures (REP)

This project investigates mechanisms of dental pulp tissue regeneration to identify the biomolecules and signaling pathways that are involved and that could be targeted to guide proper tissue regeneration including the time controlled replacement of fibrin hydrogel by a cell derived matrix similar to the dental pulp. To achieve this goal transcriptomics and proteomics approaches were performed. Human dental pulp mesenchymal stem cells (DP-MSCs) were sampled from human third molars extracted for orthodontic reasons from a population of males and females. DP-MSCs were then grown in 3 dimensional fibrin

scaffold hydrogel. Unbiased studies were performed to identify genes and proteins differentially expressed over time by MSCs in fibrin hydrogel. Two quantitative proteomics experiments were performed, each using MSCs from 3 different donors. Protein extracts obtained by mechanical disruption of the cellularized hydrogels were collected on days 0, 2 and 4 (D0, D2 D4) and the D2/D0, D4/D0 and D4/D2 ratios were calculated. Some proteins are difficult to identify by proteomics or could be easily revealed after proteolytic cleavages. A transcriptomic experiment (RNAseq) including cells from 6 donors under the same conditions was performed to complement the proteomic analysis and to assess whether the proteins identified by proteomics were regulated at the transcriptional level. The data from these experiments was used to extract the list of differentially abundant proteins and the list of differentially expressed transcripts for each time ratio D2/D0, D4/D0 and D4/D2.

During this project, we will dig into this multi-omics dataset in order to extract relevant information that could give insights on DP regeneration dynamics and kinetics within fibrin scaffold hydrogel, by enriching the dataset with biological annotations and by developing several integration strategies.

1.2 Project goals

The general objective of this project is to integrate these multi-omics data in order to analyze it in an exploratory fashion and thus highlight data-driven biological mechanisms and processes. The specific objectives are: (1) identifying significantly differentially expressed genes and differentially abundant proteins (2) associating and comparing expressions at the transcript and protein level (3) pathway enrichment analysis by querying an annotation database such as gene ontology, KEGG pathways, protein-protein interactions or even protein phosphorylation.

1.3 State of the art

RNA sequencing is a well-established next-generation sequencing technique to read the cell transcriptome[1]. Just like other NGS protocols, it involves fragmentation followed by PCR amplification and finally sequencing them into short uniformly sized reads. The mRNA molecules that are sequenced this way are “mature”, that is to say they have all been poly-adenylated and spliced. The analysis of the reads, obtained in FASTQ files, includes several steps, each of them involving some amount of quality control.

The software FastQC allows the check of the raw reads in order to discard low-quality reads and trim the remaining part. Then, mappers can be used to assign reads to the genome or transcriptome locus to which they most likely belong. To then quantify the expression, the standard measure is reads per kilobase of exon model per million reads (RPKM). Indeed, raw counts must be adjusted by transcript length, total number of reads, and sequencing biases. This can be done using softwares such as HTSeq-Count. These methods are however not adapted to differential expression, since their normalization relies on total counts. In this case, there are other softwares such as DESeq2 which are tailored to differential analysis. Literature shows[1] that the choice of method and package can have significant effects on the outcome of the analysis, so it is considered safer to treat the read data using more than one package.

Proteomics data are obtained through mass spectrometry. These methods generally involve subjecting a sample to ionization, either through electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In the case of MS/MS, peptides are fragmented twice and their mass-to-charge ratio of ions. Thanks to that measure, a set

of differentially abundant proteins is identified. In order to analyze this data, several approaches have been developed.

For transcriptomics data, log fold changes of differentially expressed genes are usually correlated pairwise in order to obtain a correlation matrix that can then be turned into an adjacency matrix to represent the information as weighted genes co-expression network [3].

Regarding proteomics data, a well-known approach is the so-called gene set enrichment analysis which consists in statistically test associations between and a given set of genes either pathways [7] or gene ontologies. It is also possible to retrieve a protein-protein interaction network.

Both transcriptomics and proteomics methods can be enriched using libraries such as Gene Ontology; this is known as functional annotation and allows identification of the most expressed genes. Once the omic data is enriched with biological/functional annotations, it becomes interesting to cluster genes by trying several methods such as k-means, or hierarchical clustering [2].

From these separate approaches, one can try to integrate the omics datasets or at least some intermediate results.

There has been a large amount of discussion over the nature of the correlation between mRNA transcription and protein translation. Clearly, the relationship between the two is not straightforward, and transcript levels are not sufficient to predict protein levels in many scenarios [5].
The most simple approach that we could think of, that will be tested as a first interpretation attempt, is the intersection or difference between the sets of differentially expressed and over/underabundant genes in both datasets [6]. Gene expression that correlates between the transcriptome and the proteome indicate a strong importance, while the difference between them allows observing post-transductional modifications in some of the genes of interest.
We also can try to integrate these datasets by using multivariate approaches such as principal component analysis or bayesian factor analysis [4].

1.4 Acceptance criteria

Complementarity of proteomic and transcriptomic results as well as signaling pathways.

2 Expression of the requirements

2.1 Functional requirements

- A Jupyter notebook with:
  - representation pathway analysis results in a clear and concise way
  - advanced management of inputs and outputs in order to enable the systematization of this kind of analyses
- A ReadMe file that gives detailed explanations on the analyses performed in the notebook

2.2 Non-functional requirements

If time allows, the design of a multi-omics data analysis pipeline will allow the identification of biomolecules to be targeted for the future development of intelligent biomaterials allowing the regeneration of functional tissues.

3 Constraints

3.1 Budgetary constraint

As the project is for the training of students and for the benefit of public research, no budget is allocated to it. It must be realized with free resources. In addition to our personal computers, we also have access to the pedagogical server of the bioinformatics Master of the Claude Bernard Lyon 1 University, hosted by the Laboratory of Biometry and Evolutionary Biology (LBBE), which allows us to benefit from a large storage capacity for the data and an important computing power to be able to run analyses.

3.2 Time constraint

This project is due on December 15th, 2022 and will be presented to the academic panel on December 16th, 2022.

4 Project planning

4.1 Sequence of events

The different tasks are planned and organized in the Gantt chart below.

PROJECT TITLE	Integration of multi-omics data for monitoring stem cell-based tissue regeneration	PROJECT MANAGER	Mourad BEKHOUCHE
ACADEMIC SUPERVISOR	Cécile Hilpert	PERIOD	from 03/10/2022 to 16/12/2022

Month				October								November										December								
day				3	4	5	6	7	10	11	12	13	14	31	1	2	3	4	7	8	9	10	11	28	29	30	1	2	5	6	7	8	15	16
TASK	**START**	**END**	**DURATION (in days)**	**WEEK 1**					**WEEK 2**						**WEEK 3**				**WEEK 4**					**WEEK 5**			**WEEK 6**							
1 project initiation, definition and planning																																	
1.1 Meeting with the project manager	03/10	03/10	1																														
1.2 Meeting with the academic supervisor	05/10	05/10	1																														
1.3 Writing project specifications	05/10	14/10	8																														
1.4 Project specifications handout	05/10	14/10	8																														
2 Project launch and execution																																	
2.1 Data recovery and pre-processing	04/10	04/10	1																														
2.2 Data integration	05/10	13/10	7																														
2.3 Identification and characterization	31/10	11/11	10																														
2.4 Associating and comparing expressions	31/10	11/11	10																														
2.5 Pathway enrichment analysis	31/10	11/11	10																														
2.6 Comparison with other programs	09/11	11/11	3																														
3 Project performance and control																																	
3.1 First results review	03/11	04/11	4																														
3.2 Checkpoint progress	10/10	08/12	11																														
3.3 Benchmarking with other programs	09/11	11/11	3																														
4 Project closing																																	
4.1 Final handout	15/12	15/12	1																														
4.2 Presentation	16/12	16/12	1																														
4.1 Final handout	15/12	15/12	1																														
4.2 Presentation	16/12	16/12	1																														

Figure 2: Project Gantt chart.

4.2 Quality-control framework

The final result's quality will be evaluated by how well the functions identified correlate between the types of data, as well as how much it explains the observations on the tissue. The assessment of the results obtained throughout the project will be done according to the enlightened knowledge of our supervisors as well as landmark results previously obtained such as GTEx data.

4.3 Documentation

A description of the methods used as part of the project such as scripts and the details of the workflow will be joined to the outcome of the project. The project code and data will be included in a github repository.

4.4 Responsibilities

4.4.1 Contracting authority

The data was produced by the Regeneration of Osteoarticular and Dental tissues research group, at the Tissue Biology and Therapeutic Engineering Laboratory (LBTI), hosted at the Institute for the Biology and Chemistry of Proteins. Our project was commissioned by Dr. Mourad Bekhouche, who is part of this team.

4.4.2 General contractors

This project is conducted by Victor Deguise, Zakaria Tougui, François Victor, Bioinformatics Master students at the Claude Bernard Lyon 1 University.

They are under the supervision of Cécile Hilpert working at the Institute for the Biology and Chemistry of Proteins in the Modeling Biological Macromolecules team (MOBI).

5 References

[1] *A survey of best practices for RNA-seq data analysis — Genome Biology — Full Text.* URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8 (visited on 10/14/2022).

[2] Liang Jin et al. "Integrative Analysis of Transcriptomic and Proteomic Profiling in Inflammatory Bowel Disease Colon Biopsies". In: *Inflammatory Bowel Diseases* 25.12 (Nov. 14, 2019), pp. 1906–1918. ISSN: 1078-0998. DOI: 10.1093/ibd/izz111. URL: https://doi.org/10.1093/ibd/izz111 (visited on 10/14/2022).

[3] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1 (Dec. 29, 2008), p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559. URL: https://doi.org/10.1186/1471-2105-9-559 (visited on 10/14/2022).

[4] *Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets — Molecular Systems Biology.* URL: https://www.embopress.org/doi/full/10.15252/msb.20178124 (visited on 10/14/2022).

[5] *On the Dependency of Cellular Protein Levels on mRNA Abundance - ScienceDirect.* URL: https://www.sciencedirect.com/science/article/pii/S0092867416302707 (visited on 10/14/2022).

[6] Indhupriya Subramanian et al. "Multi-omics Data Integration, Interpretation, and Its Application". In: *Bioinformatics and Biology Insights* 14 (Jan. 1, 2020), p. 1177932219899051. ISSN: 1177-9322. DOI: 10.1177/1177932219899051. URL: https://doi.org/10.1177/1177932219899051 (visited on 10/14/2022).

[7] Jiaogen Zhou et al. "Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment". In: *Frontiers in Genetics* 12 (2021). ISSN: 1664-8021. URL: https://www.frontiersin.org/articles/10.3389/fgene.2021.758131 (visited on 10/14/2022).