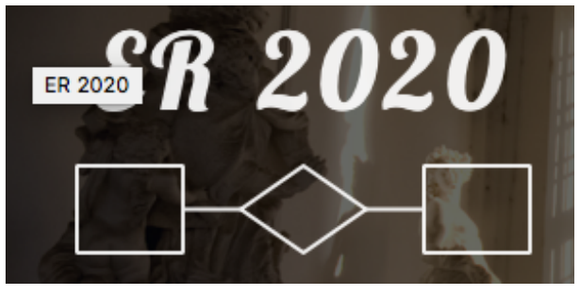
Using BioPAX-Parser (BiP) to annotate lists of biological entities with pathway data

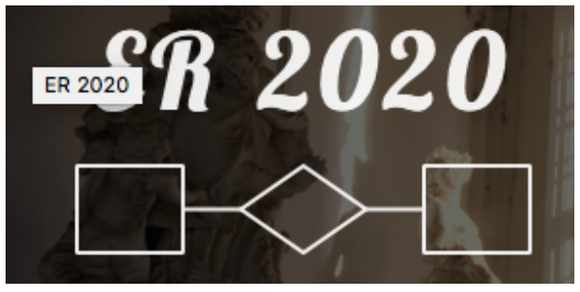
Giuseppe Agapito, Mario Cannataro
Magna Graecia University of Catanzaro
Italy

The 1st International Workshop on Conceptual Modeling for Life Sciences (CMLS 2020), in conjunction with the 39th International Conference on Conceptual Modeling (ER 2020).
November 03-06, 2020
Vienna, Austria
Virtually Conference



Outline

- Biological Pathway definition
- Pathway Databases
- Pathway Data Format
- Pathway Enrichment Analysis
- BiP
- Case Study
- Conclusion



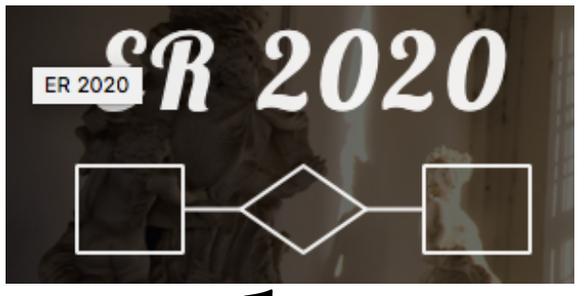
What is a Biological Pathway?

- Biological pathways represent the biological reactions and interaction networks in a cell.

- Biological Pathway can be grouped in three categories:
  - Metabolic, Regulatory and Transduction Pathways.

- It is now becoming increasingly evident that the connection between pathways and diseases is fashioned at multiple interconnected levels.

- When something goes wrong in a pathway, the result can be a disease such as cancer or diabetes.



Pathway Databases

- **KEGG** (Kyoto Encyclopedia of Genes and Genomes) is a collection of 19 databases, including genomic, chemical, pathway, and phenotypic information belonging to several organisms, including human. KEGG provides significant coverage for the human with 7, 217 annotated proteins.

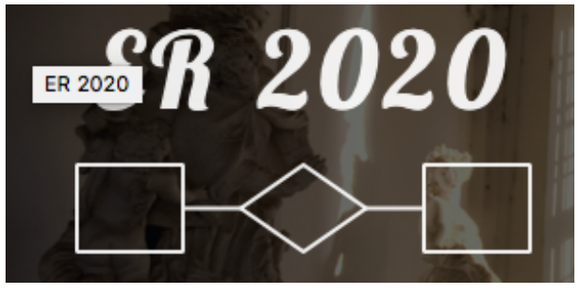
- **MetaCyc** is a curated database of experimentally elucidated metabolic pathways from all domains of life, providing information about pathways, enzymes, reactions, and metabolites. It contains 2, 801 pathways from 3,123 different organisms.

- **Panther** (Protein Analysis Through Evolutionary Relationships Classification System) database includes protein sequencing, evolutionary information, metabolic and signaling pathways information. In the current version, it stores pathways from several organisms, including human, for a total of 177 pathways.

- **PathwayCommons** is a collection of public pathway databases, providing an access point for a collection of public databases. PathwayCommons allows users to download data in PSI-MI and BioPAX formats. In the current version integrate information from 22 databases, and it provides information about 5,772 Pathways and 2,424,055 Interactions.

- **Reactome** is an open source pathway database, it contains the whole known pathways coming from 22 different organisms including the human. Pathway can be download in different formats comprising SBML, BioPAX and other graphical formats. Reactome includes 2,423 pathways for Homo sapiens and annotates 10,923 proteins.

- **WikiPathways** is an open, collaborative platform dedicated to the curation of biological pathways. WikiPathways is a new model of pathway databases that improves and complements ongoing efforts, such as KEGG, Reactome and PathwayCommons. WikiPathways provides coverage for 6,233 annotated proteins.



Pathway Data Formats

- The most used formats to represent Pathway data are:
  - The BioPAX Data Exchange format (BIOPAX): The aim of BIOPAX is to define a unified framework for sharing pathway information.
  - The Proteomics Standards Initiative Molecular Interaction XML format (PSI MI): The aim of the PSIMI is to develop standards for data representation in proteomics to facilitate data comparison, exchange and verification.
  - The Systems Biology Markup Language (SBML): The aim of SBML is to model biochemical reaction networks, including cell signaling, metabolic pathways and gene regulation.
  - Plain Text Files

BioPAX Format

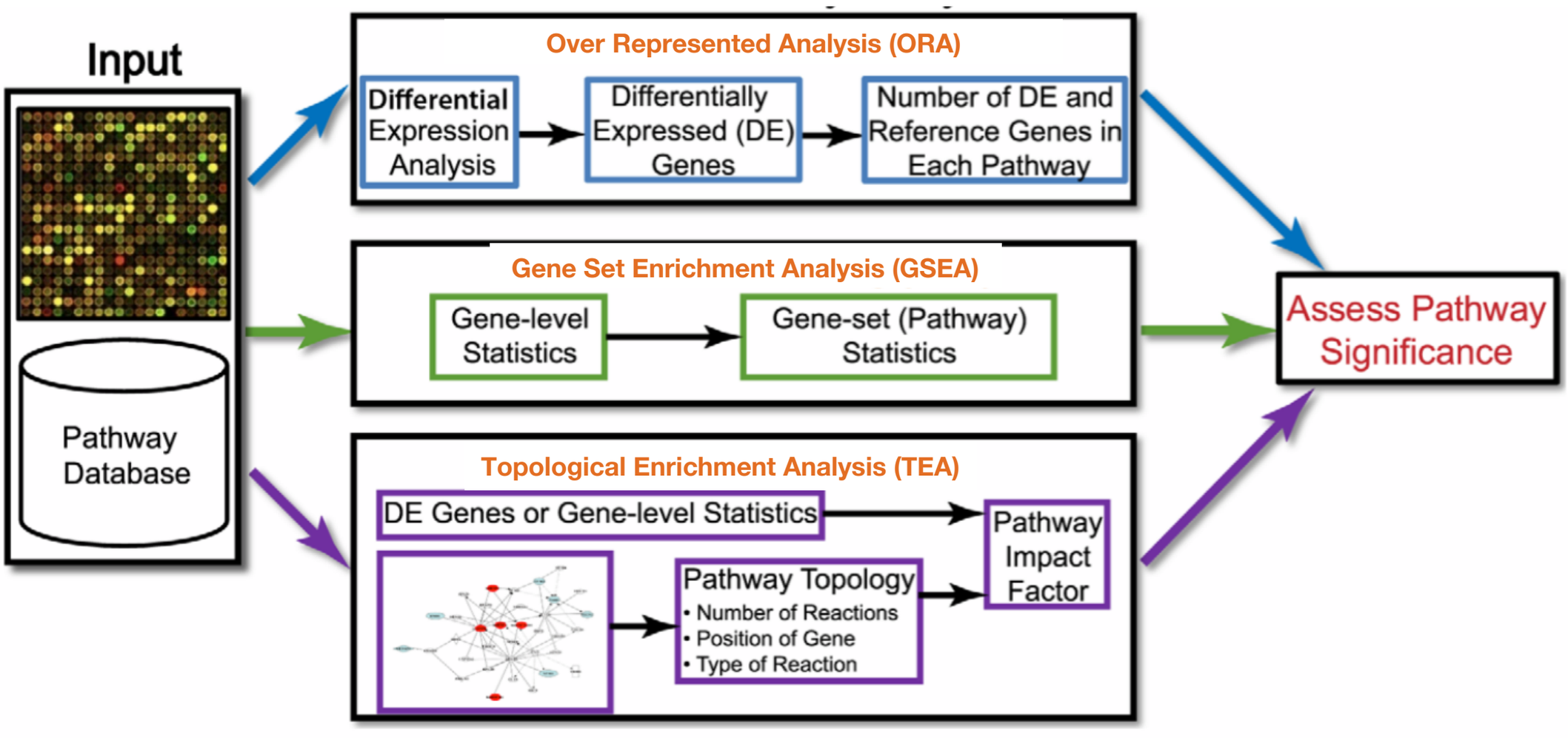
- The high level of heterogeneity among available pathway data, makes automatic data analysis ineffective.

- An attempt to reduce the high level of heterogeneity is adopted by the BIOPAX consortium.

- BIOPAX -Biological Pathway Exchange-, is a meta language defined in OWL DL and is represented in the RDF/XML format.

- BioPAX aims is to facilitate the integration and exchange of data maintained in biological pathway databases, by means of a unique way to represent data.

Pathway Enrichment Analysis (PEA)

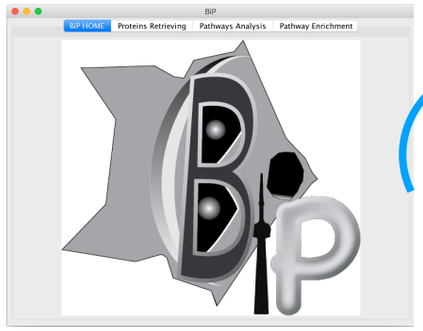


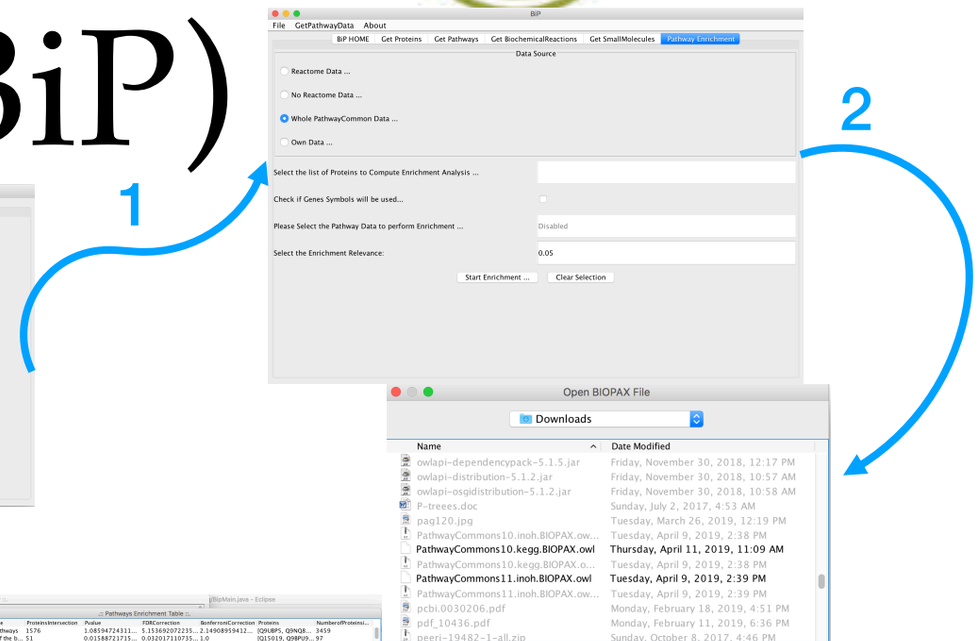
BioPAX-PARSER (BiP)

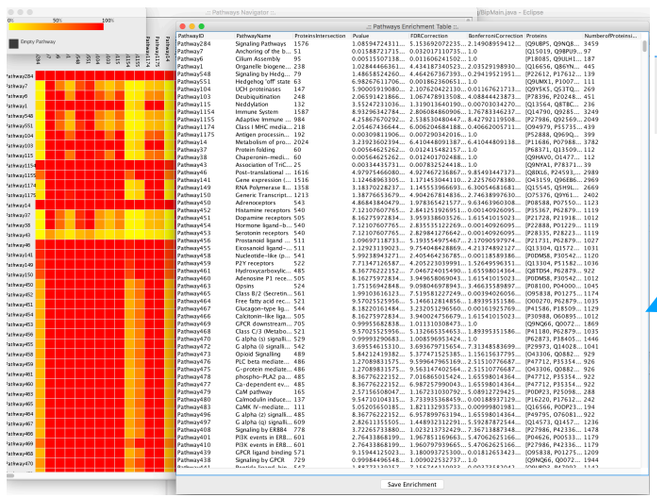
- BioPAX-Parser (BiP) is a novel software tool implemented in Java able to compute PEA from a list of genes or proteins, performing enrichment by using Hypergeometric test, along with the False Discovery Rate (FDR) and Bonferroni correctors, to correct the p-value from errors due to multiple statistical tests

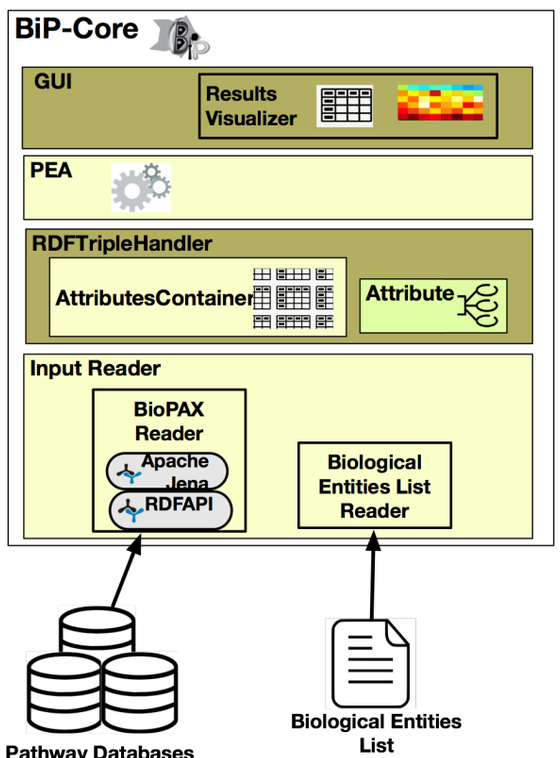
- BiP simply and effectively manages information contained in several online pathway databases based on BioPAX format.

- The current version of BiP can retrieve information from each database compliant with the BioPAX format.









Case Study

- We compared the PEA results of BiP with respect to the pathDIP software tool.

- pathDIP provides core pathways from major curated pathway databases, allowing users to perform pathway enrichment analysis

- Conversely from pathDIP, BiP, in addition to PEA, allows to retrieve information enclosed in pathways represented using the BioPAX format

- We downloaded the Endometrial cancer mutated genes list that contains about 7, 443 mutated genes.

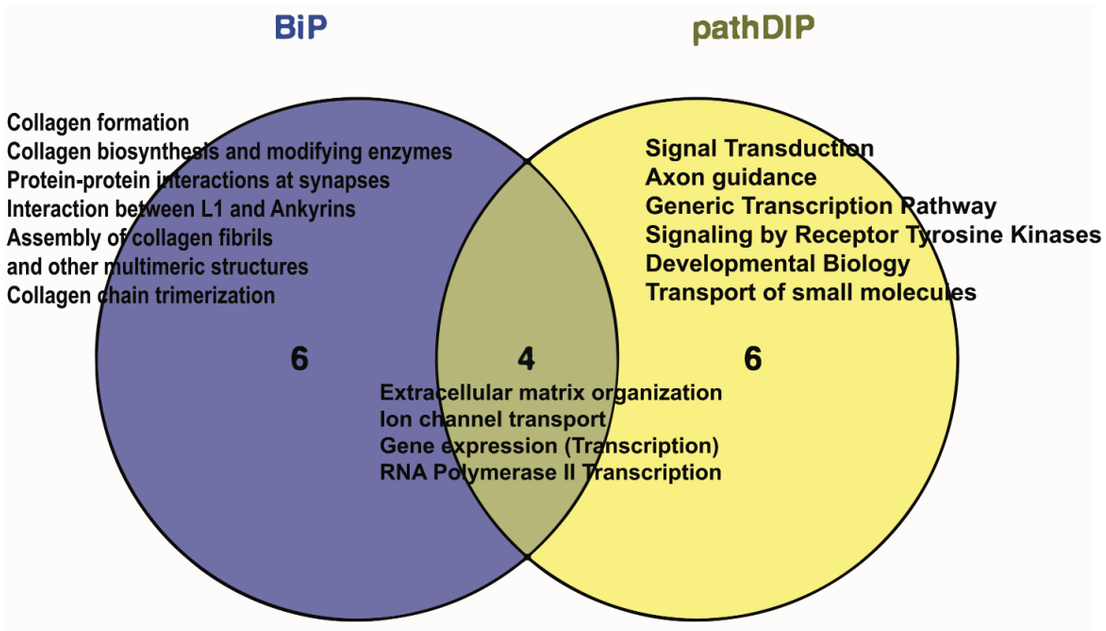
- To perform pathway enrichment we used the pathway data available in KEGG and Reactome databases.

Reactome Enrichment Results Comparison

- BiP enriched 266 Reactome pathways, whereas pathDIP enriched 225 Reactome pathways

- The comparison between the first 10 enriched pathways using Reactome and the endometrial genes list and considering p-value lower than 0.05 provided an PE overlap of 40%.

- The enriched pathways are related to the *cell activities* such as, ***Extracellular matrix organization***, ***Ion channel transport***, ***Gene expression (Transcription)***, and ***RNA Polymerase II Transcription pathways*** are well known to contribute in endometrial cancer.

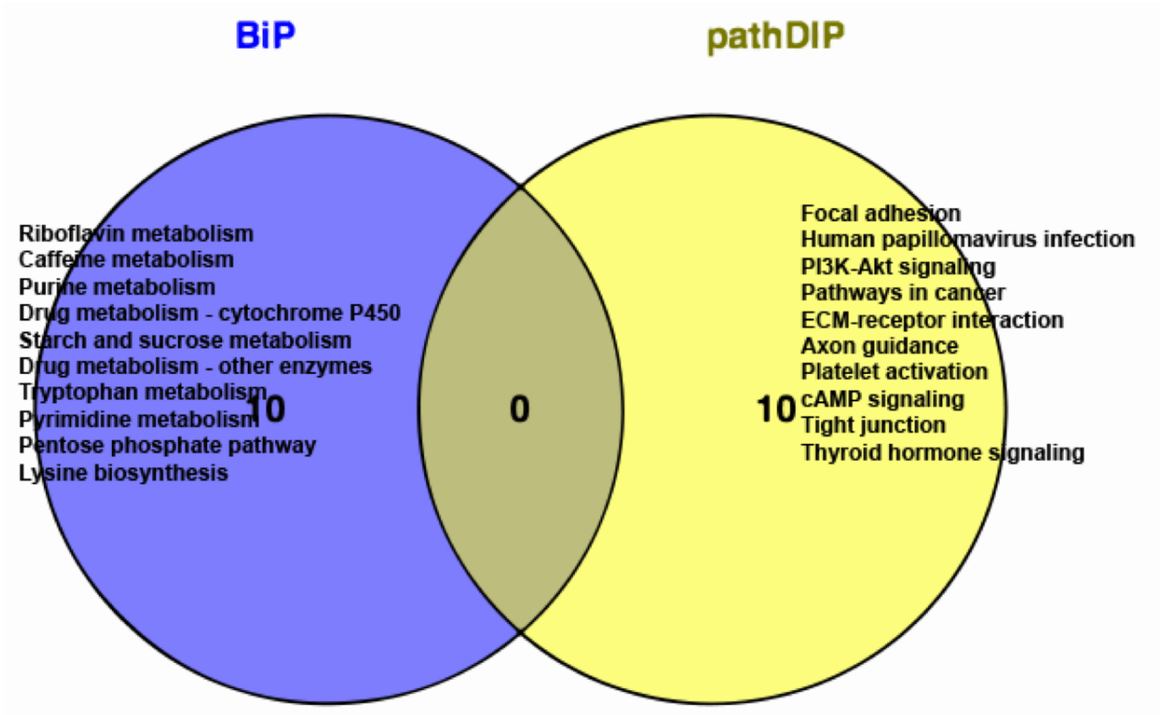


KEGG Enrichment Results Comparison

- BiP enriched 73 KEGG pathways, whereas pathDIP enriched 232 KEGG pathways.

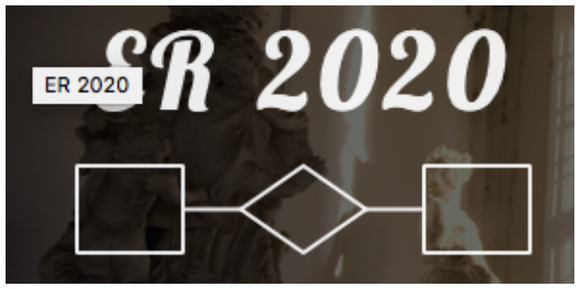
- The enrichment results obtained by using KEGG database does not provide any overlap between the two tools.

- The non-overlap between enriched pathways is due to the use of different versions of KEGG pathway database.



Conclusion and Future Work

- BiP is a Java application with which users can simply and quickly perform PEA, as well as retrieve information from BioPAX files
- The main advantage of BiP in pathway enrichment is the possibility to perform enrichment from different databases compatible with the BioPAX format, to produce more informative pathway enrichment results.
- Future work will regard the possibility to extend BiP in order to be able to deal with other pathway representation format such as XML, SBML, and GMTL.


ER 2020
ER 2020


CMLS
2020


UMG
ΒΡΕΤΤΙΩΝ
Dubium sapientiae initium

“Thank You”