PC 분석 과정

태그	Work

보고서

제목: PC(췌장암) vs HC(건강 대조군) 비교를 위한 Microbiome 및 Pathway, Orthology 데이터에서의 바이오마커 발굴 및 상관관계 분석

목적:

본 연구의 주요 목표는 Microbiome genus 데이터에서 췌장암(PC)과 건강 대조군(HC)을 구별하는 안정적이고 강력한 바이오마커를 발굴하는 것입니다. 이를 위해 통계 및 머신러닝 기반의 특성 선택 기법을 활용하여, 높은 일관성과 생물학적 중요성을 지닌 바이오마커를 검출하고자 합니다. 특히, 특정 미생물(예: *Desulfovibrio*)의 양적 변화가 대사 경로(pathway) 및 유전자 orthology에 미치는 영향을 평가함으로써, 암 환자와 건강 대조군 간 미생물군 양의 변화가 경로와 orthology 수준에서 나타나는 분명한 차별성을 규명하고자 합니다. 암환자의 경우, 특정 미생물군의 증감이 대사 경로의 활성화 또는 비활성화(up/down regulated)로 이어지고, 나아가 해당 경로를 구성하는 유전자 orthology에까지 영향을 미칠 가능성을 평가하며, 이를 통해 암환자 미생물군 프로파일의 고유한 특성을 제공하는지 확인하고자 합니다.

프로젝트 단계:

1. **Taxonomy Genus 데이터에서의 통계분석 및 특징 선택**
2. **머신러닝 모델 비교 및 선택**
3. **Genus와 Pathway 데이터 간 상관관계 분석**
4. **Orthology 분석과 최종 보고서에 통합**

Report

Title: Biomarker Discovery and Correlation Analysis between Microbiome, Pathway, and Orthology Data for PC (Pancreatic Cancer) vs HC (Healthy Control) Comparison

Objective:

The primary goal of this study is to discover stable and robust biomarkers to distinguish pancreatic cancer (PC) from healthy controls (HC) using microbiome genus data. By utilizing statistical and machine learning-based feature selection techniques, we aim to identify biomarkers with high consistency and biological significance. Specifically, we will evaluate the impact of quantitative changes in specific microbes (e.g., *Desulfovibrio*) on metabolic pathways and gene orthologies. This assessment aims to uncover distinct differences at the pathway and orthology levels in microbiome profiles between cancer patients and healthy controls. In cancer patients, the increase or decrease in certain microbial taxa may lead to up- or down-regulation of metabolic pathways and subsequently affect the orthologies of genes within these pathways, potentially offering unique microbiome profiles for cancer patients.

Project Stages:

1. **Statistical Analysis and Feature Selection from Taxonomy Genus Data**
2. **Model Comparison and Selection for Machine Learning**
3. **Correlation Analysis between Genus and Pathway Data**
4. **Orthology Analysis and Integration into the Final Report**

2. 분석 과정(Analysis Process)

1단계: Taxonomy 통계 분석(Stage 1: Taxonomy Statistical Analysis)

방법: 기존 지훈 T가 진행했던 통계 분석 결과 확인 후 사용.

File name: Summary_JH.xlsx

Method: Reviewing the previous statistical analysis conducted by Jihun T and using it as a foundation.

File name: Summary_JH.xlsx

결과:

PC와 HC 간에 유의미하게 상향 또는 하향 조절된 Genus가 식별되었으며, 이는 후속 경로 및 orthology 평가의 기초 자료로 활용되었습니다.

Results:

Significantly up- or down-regulated genera between PC and HC groups were identified, providing a basis for subsequent pathway and orthology evaluation.

2단계: 머신러닝 모델 적용 및 선택

목적:

PC (Pancreatic Cancer)와 HC (Healthy Control) 간의 차이를 분석하여 특정 Genus가 암 발생에 미치는 영향을 이해하기 위해 Machine Learning 모델을 이용한 주요 feature 식별.

Stage 2: Machine Learning Model Application and Selection

Objective:

To identify key features that impact cancer occurrence using machine learning models, by analyzing the differences between PC (Pancreatic Cancer) and HC (Healthy Control) groups.

방법:

ML 모델을 사용하여 반복적 Feature Selection (RFE)을 수행하고, 모델의 Permutation Importance와 SHAP(Shapley Additive Explanations) 값으로 각 feature의 중요도를 해석.

Method:

Machine learning models were used to perform Recursive Feature Elimination (RFE) and analyze feature importance through Permutation Importance and SHAP (Shapley Additive Explanations) values.

과정:

1. **데이터 준비 및 전처리**:
    - PC와 HC 데이터를 로드하여 ‘group_1’을 label로 지정.
    - 모든 feature는 float 타입으로 변환, label은 int 타입으로 변환.
    - study_no을 index로 설정하여 discovery,valdiation 구분 가능.
2. **Cross-Validation 기반 모델 학습 및 평가**:
    - Stratified K-Fold를 이용하여 5-겹 교차검증을 수행.
    - 모델 옵션으로 XGBoost, Random Forest, CatBoost, Gradient Boosting, LightGBM 중 선택 가능하며, 각 모델의 성능을 비교하여 최적 모델을 도출.

- 각 fold에서 RFE로 주요 feature를 선별하여 모델을 재학습하고, Accuracy와 FNR을 평가.

3. **Permutation Importance와 SHAP 해석**:
   - 모델의 Permutation Importance를 계산하여 학습 데이터와 검증 데이터 각각에서 중요한 feature를 확인하고 비교.
   - SHAP 값을 이용하여 각 feature가 모델 예측에 미치는 영향을 시각화하여 해석의 용이성을 높임.
4. **결과 기록 및 주요 feature 비교**:
   - 교차 검증을 통해 선정된 feature의 빈도수를 기록하고, 훈련과 검증에서 공통적으로 중요한 feature를 식별.
   - 교차 검증의 평균 Accuracy와 FNR을 산출하여 모델의 성능을 종합적으로 평가.
5. **결과 저장**:
   - 각 fold의 feature ImportancePermutation Importance 결과를 엑셀 파일로 저장.
   - SHAP 해석을 포함한 주요 feature 분석 결과(AUC, Feature Importance, confusion Matrix)를 시각화하여 PNG 파일로 저장.

Process:

1. **Data Preparation and Preprocessing**:
   - Load PC and HC data, setting ‘group_1’ as the label.
   - Convert all features to float type and labels to integer type.
   - Set `study_no` as the index to distinguish discovery and validation sets.
2. **Cross-Validation Based Model Training and Evaluation**:
   - Perform 5-fold cross-validation using Stratified K-Fold.
   - Available models include XGBoost, Random Forest, CatBoost, Gradient Boosting, and LightGBM. Each model’s performance was compared to select the best model.
   - Important features were selected in each fold using RFE, and model performance was evaluated using Accuracy and FNR.
3. **Interpretation of Permutation Importance and SHAP**:

- Permutation Importance was computed for both training and validation sets to identify and compare important features.
- SHAP values were used to visualize the impact of each feature on model predictions for easier interpretation.

4. **Recording and Comparing Key Features**:
    - Feature selection frequency was recorded across cross-validation, and common important features between training and validation were identified.
    - The model's overall performance was assessed by calculating the mean Accuracy and FNR across folds.
5. **Saving Results**:
    - Feature importance from each fold and permutation importance results were saved as Excel files.
    - Important feature analysis results (AUC, Feature Importance, Confusion Matrix) with SHAP interpretation were saved as PNG images

결과

- 교차 검증을 통해 도출된 주요 feature는 cancer group을 구분하는 데 유의미한 기여를 하며, train과 validation set 간 공통 feature도 높은 비율로 확인됨.
- Permutation Importance와 SHAP 분석을 통해 PC와 HC 간 차이를 유도하는 핵심 microbiome genus feature를 확인할 수 있었음.
- 최종적으로 도출된 평균 Accuracy은 88.86%, FNR 은 42.34% AUC값은 0.9355 정도이다

Results:

- Key features identified through cross-validation significantly contributed to distinguishing the cancer group. A high percentage of common features were found across train and validation sets.
- Through Permutation Importance and SHAP analysis, we confirmed the key microbiome genus features driving the differences between PC and HC groups.
- The final average Accuracy was 88.86%, FNR was 42.34%, and the AUC value was approximately 0.9355.

방법:

- **Recursive Feature Elimination (RFE)**: XGBoost, CatBoost, LightGBM, Random Forest, Gradient Boosting Machine 모델과 결합하여 일관성 있는 중요한 특징을 선별
- **Bootstrap Resampling**: 다양한 데이터 샘플링을 통해 선정된 특징이 모델에서 일관되게 중요한 역할을 하는지 확인
- Cross- Valdiation with label discovery and validation set.

Method:

- **Recursive Feature Elimination (RFE)**: Combined with XGBoost, CatBoost, LightGBM, Random Forest, and Gradient Boosting Machine models to select consistent, important features.
- **Bootstrap Resampling**: Assessed whether the selected features consistently played a crucial role across resampled data.
- **Cross-Validation with discovery and validation label sets**

베이스라인 모델 및 최종 선택:

RFE로 인해 학습 시간이 증가하여, 먼저 하이퍼파라미터 튜닝 없이 베이스라인 모델로 성능을 비교하였습니다. 최종적으로 XGBoost가 우수한 성능을 보여 최종 선택되었습니다.

선정된 공통 특징:

Discovery 및 Validation 데이터셋에서 **Permutation Feature Importance**를 통해 선별된 공통 특징들은 다음과 같습니다:

Baseline Model and Final Selection:

Without hyperparameter tuning, baseline models were compared due to the increased training time from RFE. XGBoost demonstrated the best performance and was selected as the final model.

Common Selected Features:

Common features identified through **Permutation Feature Importance** across discovery and validation datasets are as follows:

- **공통 특징 Genus**:(Common Genus Features)
  - `Desulfovibrio`, `Fretibacterium`, `Lactobacillus`, `Leuconostoc`, `Olsenella`, `Parvimonas`, `Pseudomonas`, `Ralstonia`, `Simonsiella`

- **Upregulated Genera**:
  - `Desulfovibrio`, `Fretibacterium`, `Lactobacillus`, `Leuconostoc`, `Olsenella`, `Parvimonas`, `Ralstonia`
- **Downregulated Genera**:
  - `Pseudomonas`, `Simonsiella`

이렇게 분류된 Genus는 다음 단계에서 Pathway 및 Orthology와의 상관관계 분석에 사용됩니다.

These classified genera will be used in the next step for correlation analysis with pathways and orthologies.

3단계: Pathway 통계 분석(Stage 3: Pathway Statistical Analysis)

방법:

- **Wilcoxon Signed-Rank Test**: 그룹 간 통계적 차이를 확인하여 유의미한 Pathway 탐색
- **Fold Change (FC) 계산**: 상대적 양의 변화를 평가하여 상향 또는 하향 조절 여부 판단
- **Logistic Regression with FDR adjusted**: 로지스틱 회귀 분석을 통해 각 Pathway 유의성을 평가하여 바이오마커 후보 선정

목적: Pathway에서 PC와 HC 그룹 간 차별화된 Pathway 선별

Method:

- **Wilcoxon Signed-Rank Test**: Used to identify significant pathways by comparing group differences.
- **Fold Change (FC) Calculation**: Evaluated the relative amount of change to assess up- or down-regulation.
- **Logistic Regression with FDR adjustment**: Logistic regression was conducted to evaluate pathway significance and identify biomarker candidates.

Objective: Identify differentiated pathways between PC and HC groups.

결과: (Results:)

Pathway Regulation	KEGG Code	Pathway Names
Downregulated	ko00472	D-Arginine and D-ornithine metabolism
Downregulated	ko05231	Choline metabolism in cancer

Downregulated	ko00430	Taurine and hypotaurine metabolism
Downregulated	ko00984	Prodigiosin biosynthesis
Downregulated	ko04150	D-Glutamine and D-glutamate metabolism
Downregulated	ko03320	Retinol metabolism
Downregulated	ko00333	Ether lipid metabolism
Downregulated	ko00471	Steroid degradation
Downregulated	ko04750	mTOR signaling pathway
Downregulated	ko00830	Inflammatory mediator regulation of TRP channels
Downregulated	ko00944	Flavone and flavonol biosynthesis
Downregulated	ko00404	Notch signaling pathway
Downregulated	ko04330	PPAR signaling pathway
Downregulated	ko00565	Staurosporine biosynthesis
Downregulated	ko04975	Fat digestion and absorption
Upregulated	ko04024	cAMP signaling pathway
Upregulated	ko04977	Tight junction
Upregulated	ko00601	PI3K-Akt signaling pathway
Upregulated	ko04530	Protein processing in endoplasmic reticulum
Upregulated	ko04151	Biosynthesis of siderophore group nonribosomal peptides
Upregulated	ko04141	Arginine biosynthesis
Upregulated	ko01053	Ubiquitin mediated proteolysis
Upregulated	ko00220	Vitamin digestion and absorption
Upregulated	ko04120	Glycosphingolipid biosynthesis - lacto and neolacto series

→ William T에게 전달

Up-regulated pathway(9개)

Pathway_ID	pathway_kegg_no
‹o04024	cAMP signaling pathway
‹o04977	Vitamin digestion and absorption
‹o00601	Glycosphingolipid biosynthesis - lacto and neolacto series
‹o04530	Tight junction
‹o04151	PI3K-Akt signaling pathway
‹o04141	Protein processing in endoplasmic reticulum
‹o01053	Biosynthesis of siderophore group nonribosomal peptides
‹o00220	Arginine biosynthesis
‹o04120	Ubiquitin mediated proteolysis

Down-regulated pathway(15개)

Pathway_ID	pathway_kegg_no
ko00944	Flavone and flavonol biosynthesis
ko04330	Notch signaling pathway
ko00472	D-Arginine and D-ornithine metabolism
ko04975	Fat digestion and absorption
ko00404	Staurosporine biosynthesis
ko00565	Ether lipid metabolism
ko05231	Choline metabolism in cancer
ko00430	Taurine and hypotaurine metabolism
ko00984	Steroid degradation
ko04150	mTOR signaling pathway
ko03320	PPAR signaling pathway
ko00333	Prodigiosin biosynthesis
ko00471	D-Glutamine and D-glutamate metabolism
ko04750	Inflammatory mediator regulation of TRP channels
ko00830	Retinol metabolism

4단계: Genus 데이터와의 상관관계 분석(Stage 4: Correlation Analysis with Genus Data)

목적:

특정 Genus와 영향을 미치는 대사 경로를 파악하기 위해, 상향 및 하향 조절된 Pathway를 식별하고 해당 Pathway와 Genus 간의 상관관계를 분석하였습니다.

과정:

1. **Spearman 상관분석**: Genus와 Pathway 데이터 간의 유의미한 상관관계를 찾기 위해 Spearman 상관계수를 사용하여 분석. 상관계수가 0.2 이상이고 p-value가 0.05 미만인 경우를 최종적으로 선별

97개의 조합 확인(위의 모든 pathway는 상관성을 가지고 있음)

Objective:

To identify the pathways influenced by specific genera, we conducted a correlation analysis to detect pathways that are up- or down-regulated, then assessed the correlation between these pathways and genera.

Process:

1. **Spearman Correlation Analysis**: Conducted to detect meaningful correlations between genus and pathway data. Final selection was based on correlations greater than 0.2 with a p-value less than 0.05.

97 combinations were identified, and all pathways displayed correlations.

File name: correlation_analysis_results.xlsx

File name: correlation_analysis_results.xlsx

Taxonomy	Pathway	pathway_kegg_no	Correlation	p-value	p_value_adjusted	Method	Regulation Type
Desulfovibrio	ko04024	cAMP signaling pathway	2.27E-01	5.80E-07	5.80E-07	Spearman	Upregulated
Desulfovibrio	ko04530	Tight junction	2.25E-01	6.83E-07	6.83E-07	Spearman	Upregulated
Desulfovibrio	ko04151	PI3K-Akt signaling pathway	2.13E-01	2.63E-06	2.63E-06	Spearman	Upregulated
Desulfovibrio	ko04141	Protein processing in endoplasmic reticulum	2.01E-01	9.42E-06	9.42E-06	Spearman	Upregulated
Fretibacterium	ko04024	cAMP signaling pathway	3.94E-01	3.99E-19	3.99E-19	Spearman	Upregulated
Fretibacterium	ko04977	Vitamin digestion and absorption	3.06E-01	9.04E-12	9.04E-12	Spearman	Upregulated
Fretibacterium	ko00601	Glycosphingolipid biosynthesis - lacto and neolacto series	2.26E-01	6.39E-07	6.39E-07	Spearman	Upregulated
Fretibacterium	ko04530	Tight junction	3.88E-01	1.64E-18	1.64E-18	Spearman	Upregulated
Fretibacterium	ko04151	PI3K-Akt signaling pathway	3.22E-01	6.32E-13	6.32E-13	Spearman	Upregulated
Fretibacterium	ko04141	Protein processing in endoplasmic reticulum	3.03E-01	1.36E-11	1.36E-11	Spearman	Upregulated
Fretibacterium	ko01053	Biosynthesis of siderophore group nonribosomal peptides	2.86E-01	2.17E-10	2.17E-10	Spearman	Upregulated
Fretibacterium	ko00220	Arginine biosynthesis	3.45E-01	9.06E-15	9.06E-15	Spearman	Upregulated
Fretibacterium	ko04120	Ubiquitin mediated proteolysis	2.44E-01	7.43E-08	7.43E-08	Spearman	Upregulated
Lactobacillus	ko04024	cAMP signaling pathway	3.20E-01	8.23E-13	8.23E-13	Spearman	Upregulated
Lactobacillus	ko04977	Vitamin digestion and absorption	3.01E-01	1.91E-11	1.91E-11	Spearman	Upregulated
Lactobacillus	ko00601	Glycosphingolipid biosynthesis - lacto and neolacto series	2.39E-01	1.34E-07	1.34E-07	Spearman	Upregulated
Lactobacillus	ko04530	Tight junction	2.19E-01	1.43E-06	1.43E-06	Spearman	Upregulated
Lactobacillus	ko04141	Protein processing in endoplasmic reticulum	2.16E-01	1.99E-06	1.99E-06	Spearman	Upregulated
Lactobacillus	ko01053	Biosynthesis of siderophore group nonribosomal peptides	2.51E-01	2.75E-08	2.75E-08	Spearman	Upregulated
Leuconostoc	ko04024	cAMP signaling pathway	2.58E-01	1.14E-08	1.14E-08	Spearman	Upregulated
Leuconostoc	ko04977	Vitamin digestion and absorption	2.79E-01	5.76E-10	5.76E-10	Spearman	Upregulated
Leuconostoc	ko00601	Glycosphingolipid biosynthesis - lacto and neolacto series	2.66E-01	3.53E-09	3.53E-09	Spearman	Upregulated
Leuconostoc	ko04530	Tight junction	2.41E-01	1.04E-07	1.04E-07	Spearman	Upregulated
Leuconostoc	ko04151	PI3K-Akt signaling pathway	2.31E-01	3.34E-07	3.34E-07	Spearman	Upregulated
Leuconostoc	ko04141	Protein processing in endoplasmic reticulum	2.57E-01	1.27E-08	1.27E-08	Spearman	Upregulated
Leuconostoc	ko01053	Biosynthesis of siderophore group nonribosomal peptides	2.47E-01	4.55E-08	4.55E-08	Spearman	Upregulated
Leuconostoc	ko00220	Arginine biosynthesis	2.46E-01	5.14E-08	5.14E-08	Spearman	Upregulated
Olsenella	ko04024	cAMP signaling pathway	2.65E-01	4.05E-09	4.05E-09	Spearman	Upregulated
Olsenella	ko00601	Glycosphingolipid biosynthesis - lacto and neolacto series	2.51E-01	2.88E-08	2.88E-08	Spearman	Upregulated
Olsenella	ko04530	Tight junction	2.22E-01	1.02E-06	1.02E-06	Spearman	Upregulated
Olsenella	ko01053	Biosynthesis of siderophore group nonribosomal peptides	2.42E-01	9.55E-08	9.55E-08	Spearman	Upregulated

결과 (ReusIts)

예시(postivie, negative 가장 상관관계 높은 tax, pathway 선정)

Taxonomy	Pathway	Correlation	p-value adjusted
Parvimonas	Prodigiosin biosynthesis	-0.395	2.98E-19
Parvimonas	Biosynthesis of siderophore group nonribosomal peptides	0.47	1.63E27

5단계: Orthology 분석

목적:

식별된 Genus-Pathway 관계를 검증하기 위해 orthology 분석.

방법

기존 pathway 분석과 동일하게 진행

과정:

1. **Orthology 데이터에 대한 그룹 비교 분석**: PC와 HC 간 유의미한 차이를 보이는 Ortholog를 식별하기 위해 통계 검정을 수행 (statistical_analysis_results_with_labels_Orthology.xlsx)
2. 위에서 선정되었던 Pathway의 KEGG에서 pathway map을 통해 연결된 ortholog 병합.(orthology_analysis_results.xlsx)

결과

저희가 가지고 있는 Pathway에 contained 되어있는 orthology unique(722) 855 개

기존 Orthology label list에서 확인이 가능했던 orthology 개수는 unique(683) 813개,

그렇지 않은 orthology의 개수는 42개 그러나 아래에 보이다시피 label이 없는 데이터는 대부분 값이 0으로 확인

	Pathway_ID	Orthology_kegg_no	Orthology_Name	Control_mean	ncer_me	FC_value	ilcoxon_v	g2FC_val	Reg_p_va	Reg_p_fd
811	ko00220	K22477	N-acetylglutamate synthase	1.50072395833333E-05	1.735E-05	1.15613	0.14236	0.209304	3.765E-10	3.85E-09
812	ko00220	K22478	bifunctional N-acetylglutamate synthase/kinase	6.90104166666667E-09	7.609E-09	1.102543	0.042661	0.140835	0.999972	1
813	ko04150	K08267		2.60416666666667E-11	0	0	0.628278		1	1
814	ko04151	K06485		0	0		1			
815	ko04150	K20403		1.30208333333333E-10	6.522E-10	5.008696	0.271473	2.324435	0.999998	1
816	ko04151	K06584		0	0		1			
817	ko00430	K18966		0	0		1			
818	ko04141	K01367		0	0		1			
819	ko04530	K08020		0	0		1			
820	ko04530	K12751		0	0		1			
821	ko00565	K13519		0	0		1			
822	ko04750	K03898		0	0		1			
823	ko00404	K19887		0	0		1			
824	ko00430	K19699		0	0		1			
825	ko04141	K14006		0	0		1			
826	ko04330	K20995		0	0		1			
827	ko04330	K21635		0	0		1			
828	ko04141	K04554		0	0		1			
829	ko04120	K04554		0	0		1			
830	ko04120	K05632		0	0		1			
831	ko04330	K05632		0	0		1			
832	ko04024	K18435		0	4.348E-10		0.04158		0.999999	1
833	ko04530	K04416		0	0		1			
834	ko04120	K04416		0	0		1			
835	ko00220	K13427		0	0		1			
836	ko05231	K06515		0	0		1			
837	ko04977	K01435		0	0		1			
838	ko04120	K04678		0	0		1			
839	ko04151	K03259		0	0		1			
840	ko04150	K03259		0	0		1			
841	ko04141	K08860		0	0		1			
842	ko04330	K20994		0	0		1			
843	ko04151	K06583		0	0		1			
844	ko00333	K21793		0	0		1			
845	ko04141	K10088		0	0		1			
846	ko04151	K09554		1.5625E-10	0	0	0.628278		1	1
847	ko03320	K08525		0	0		1			
848	ko04151	K10159		0	0		1			
849	ko04750	K04960		0	0		1			
850	ko04530	K09183		0	0		1			
851	ko04150	K08271		0	0		1			
852	ko04151	K10605		0	0		1			
853	ko04120	K10605		0	0		1			
854	ko00565	K22387		0	4.348E-10		0.04158		0.999999	1

이후 orthology값 자체가 0인 경우를 제거하여 unique(683) 789개의 orthology

통계적 유의성 검정을 진행한 결과 유의한 결과로 나온 Orthology 목록

결과:

Downregulated orthologies(4개)

File Name: down_regulated_ortho.xlsx

- Taurine dioxygenase
- Beta-1,4-galactosyltransferase 1
- DnaJ homolog subfamily C member 3
- DNA damage-binding protein 1

Upregulated Orthologies(14개)

File Name: up_regulated_ortho.xlsx

- Dimethylaniline monooxygenase (N-oxide forming)
- Arginase
- Alpha-N-acetyl-neuraminate alpha-2,8-sialyltransferase (sialyltransferase 8A)
- Sulfoacetaldehyde acetyltransferase
- Glutamate receptor ionotropic, NMDA 1

    - Tuberous sclerosis 2
- Serine/threonine-protein kinase 11
- Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-3
- Cytoskeleton-associated protein 4
- 5-Epimerase

위의 orthology들이 어떤 pathway에 속하는지 확인하기 위해 pathway_ID와 매핑

pathway_kegg_no	Pathway_ID	Orthology_kegg_no	Orthology_Name	Control_mean	Cancer_m
Taurine and hypotaurine metabolism	ko00430	K00485	dimethylaniline monooxygenase (N-oxide forming)	1.58643E-06	3.93337E
Arginine biosynthesis	ko00220	K01476	arginase	9.02911E-06	1.91448E
Glycosphingolipid biosynthesis - lacto and neolacto series	ko00601	K03371	alpha-N-acetyl-neuraminate alpha-2,8-sialyltransferase (sialyltransferase 8A)	5.51419E-06	1.11666E
Taurine and hypotaurine metabolism	ko00430	K03852	sulfoacetaldehyde acetyltransferase	3.67253E-06	7.72348E
cAMP signaling pathway	ko04024	K05208	glutamate receptor ionotropic, NMDA 1	6.86878E-06	1.40782E
PI3K-Akt signaling pathway	ko04151	K07207	tuberous sclerosis 2	6.38419E-06	1.52618E
Choline metabolism in cancer	ko05231	K07207	tuberous sclerosis 2	6.38419E-06	1.52618E
mTOR signaling pathway	ko04150	K07207	tuberous sclerosis 2	6.38419E-06	1.52618E
Tight junction	ko04530	K07298	serine/threonine-protein kinase 11	1.3374E-06	2.79946E
PI3K-Akt signaling pathway	ko04151	K07298	serine/threonine-protein kinase 11	1.3374E-06	2.79946E
mTOR signaling pathway	ko04150	K07298	serine/threonine-protein kinase 11	1.3374E-06	2.79946E
PI3K-Akt signaling pathway	ko04151	K07825	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-3	8.14297E-07	2.80087E
Protein processing in endoplasmic reticulum	ko04141	K13999	cytoskeleton-associated protein 4	3.46409E-06	1.04224E
Staurosporine biosynthesis	ko00404	K16438	5-epimerase	3.48617E-06	7.16011E
Taurine and hypotaurine metabolism	ko00430	K03119	taurine dioxygenase	2.38448E-06	6.75761E
Glycosphingolipid biosynthesis - lacto and neolacto series	ko00601	K07966	beta-1,4-galactosyltransferase 1	4.8125E-06	1.77848E
Protein processing in endoplasmic reticulum	ko04141	K09523	DnaJ homolog subfamily C member 3	3.10208E-05	0.000014
Ubiquitin mediated proteolysis	ko04120	K10610	DNA damage-binding protein 1	4.99734E-06	1.56348E

6단계: Orthology와 genus 간 상관관계 분석

과정:

Spearman 상관분석: 위에서 통계적으로 유의미하다고 판단된 Orthology들의 genus 데이터 간의 유의미한 상관관계를 찾기 위해 Spearman 상관계수를 사용하여 분석.

상관계수가 0.2 이상이고 p-value가 0.05 미만인 경우를 최종적으로 선별

결과

up regulated된 genus와의 주요 상관성이 있는 orthology

File Name : Correlation_analysis_up_genus_orthology.xlsx

	Taxonomy	Orthology	Orthology_Correlation	Orthology_p_value	Orthology_p_value_adjusted	Orthology_Correlation_Method	Orthology_Regulation_Type
70	Parvimonas	K00485	0.722189155	6.66897E-78	6.53559E-76	Spearman	Upregulated
16	Fretibacterium	K03371	0.656057598	6.51141E-60	3.19059E-58	Spearman	Upregulated
94	Ralstonia	K03119	0.606136493	4.43686E-49	1.44938E-47	Spearman	Downregulated
18	Fretibacterium	K05208	0.542644955	8.43836E-38	2.0674E-36	Spearman	Upregulated
72	Parvimonas	K03371	0.494270238	1.08511E-30	2.12682E-29	Spearman	Upregulated
15	Fretibacterium	K01476	0.483422375	3.01614E-29	4.92636E-28	Spearman	Upregulated
74	Parvimonas	K05208	0.473590667	5.55039E-28	7.77054E-27	Spearman	Upregulated
90	Ralstonia	K07298	0.43777765	1.04704E-23	1.28262E-22	Spearman	Upregulated
14	Fretibacterium	K00485	0.425209539	2.53629E-22	2.76174E-21	Spearman	Upregulated
71	Parvimonas	K01476	0.416131628	2.33636E-21	2.28963E-20	Spearman	Upregulated
56	Olsenella	K00485	0.370324387	6.41975E-17	5.71942E-16	Spearman	Upregulated
40	Lactobacillus	K09523	-0.368789552	8.80039E-17	7.18699E-16	Spearman	Downregulated
20	Fretibacterium	K07298	0.353184169	1.97805E-15	1.49115E-14	Spearman	Upregulated
39	Lactobacillus	K07966	-0.351233733	2.88409E-15	2.01886E-14	Spearman	Downregulated
58	Olsenella	K03371	0.343685186	1.21117E-14	7.91296E-14	Spearman	Upregulated
36	Lactobacillus	K13999	0.317228123	1.37285E-12	8.40874E-12	Spearman	Upregulated
2	Desulfovibrio	K03371	0.311931829	3.35199E-12	1.93232E-11	Spearman	Upregulated

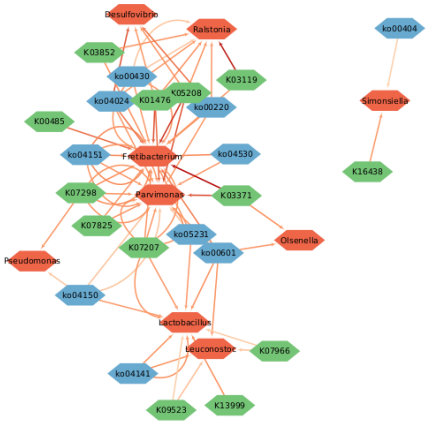
down regulated 된 genus와의 주요 상관성이 있는 orthology

	Taxonomy	Orthology	Orthology_Correlation	Orthology_p_value	Orthology_p_value_adjusted	Orthology_Correlation_Method	Orthology_Regulation_Type
6	Pseudomonas	K07298	0.251945176	2.51133E-08	1.75793E-07	Spearman	Upregulated
10	Pseudomonas	K03119	0.687884747	5.73047E-68	1.60453E-66	Spearman	Downregulated
23	Simonsiella	K16438	0.207118514	5.2035E-06	2.91396E-05	Spearman	Upregulated
25	Simonsiella	K07966	0.312152871	3.23055E-12	3.01518E-11	Spearman	Downregulated
26	Simonsiella	K09523	0.356041705	1.13304E-15	1.58625E-14	Spearman	Downregulated

All combined in Summary(analysis_output)

11.4~11.8

Using Cytoscape for Network Analysis



1. 분석 목표

각 미생물 속(`Genus`)이 특정 **Pathway** 및 그 Pathway 내의 **Orthology**에 미치는 영향을 분석하여, **암 발생 여부**에 따라 이들이 어떻게 변하는지 파악하는 것을 목표로 합니다.

---

2. 분석 개요

- **SEM의 원리**: 선형 회귀의 기본 원리를 확장하여 다중 종속 변수를 동시에 모델링함으로써 유전자 경로 분석에서 다층적 인과관계를 추정
- **다중 종속 변수 모델링**: 한 변수(`Genus`)가 여러 종속 변수(`Pathway`, `Cancer`)에 동시에 영향을 미칠 수 있음
- **잠재 변수 포함**: 관측되지 않는 잠재 변수(`Pathway`)를 포함하여 여러 관측 변수(`Orthology`)로 설명 가능
- **직접 및 간접 효과 분석**: 각 변수 간의 직접적, 간접적 인과관계를 모델링해 유전자 경로와 암 발생의 전체적인 영향을 심층적으로 분석

---

3. SEM 모델 설정 개요

- **잠재변수 정의**:
  - **Genus**: 독립변수로 설정
  - **Pathway**: 잠재변수로 설정, 여러 Orthology로 측정
  - **Orthology**: Pathway의 하위 개념으로 구성
  - **암 발생 여부**: Pathway와 Orthology에 미치는 영향을 평가하기 위한 조절변수로 설정
- **SEM 모델 구조**:
  - **Genus ➔ Pathway ➔ Orthology**: 각 Genus가 특정 Pathway에 영향을 주고, 그 Pathway가 하위 Orthology에 영향을 미친다고 가정
  - **암 발생 여부의 조절 효과**: 암 발생 여부에 따라 Pathway와 Orthology가 어떻게 변하는지 확인

---

4. 분석 데이터 구성

분석에 활용할 데이터는 다음과 같습니다.

- `Genus` abundance
- `Pathway` activity
- `Orthology` 수치
- 암 발생 여부 (`Cancer`)

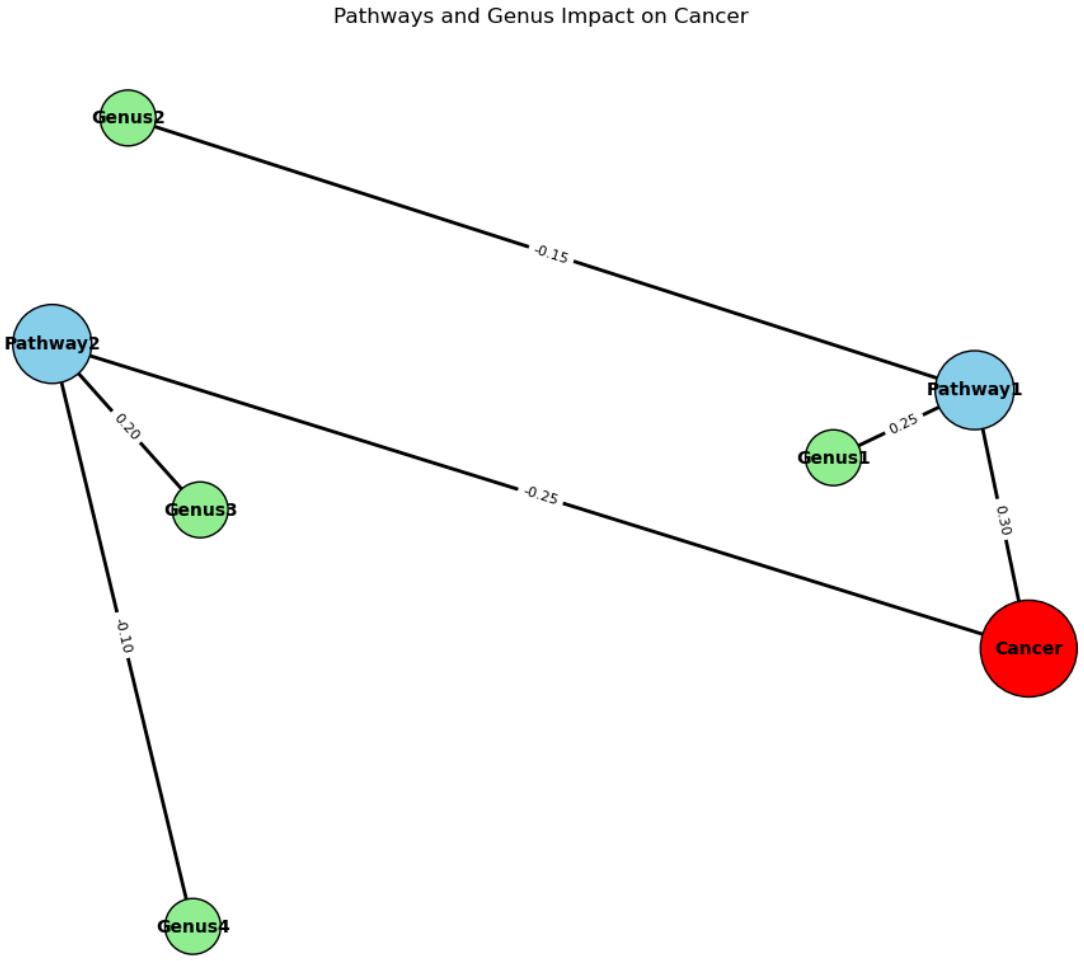
5. 분석 내용

- **Genus가 Pathway에 미치는 영향**:
  - 각 Genus의 abundance가 특정 Pathway 활성에 얼마나 기여하는지 확인
  - Pathway의 회귀 계수가 유의미하다면, 해당 Genus가 Pathway에 중요한 영향을 미친다는 의미로 해석 가능
- **Pathway가 Orthology에 미치는 영향**:
  - Pathway가 하위의 Orthology에 미치는 영향을 파악해, Pathway 활성화가 특정 Orthology 변화에 미치는 영향 확인
- **암 발생 여부의 조절 효과**:
  - `Cancer` 발생 여부가 Pathway와 Orthology 간 관계에 미치는 영향을 분석
  - `Cancer`가 유의미한 영향을 미친다면, 암 발생 여부에 따라 특정 Pathway나 Orthology에 차이가 있음을 시사

---

6. SEM과 기존 회귀 분석의 차이점

- **다중 종속 변수 모델링**: 기존 회귀 분석은 하나의 종속 변수만 다루지만, SEM은 여러 종속 변수 간 인과 관계를 동시에 추정하여 `Genus`가 `Pathway`와 `Cancer`에 미치는 직접적, 간접적 영향을 분석
- **잠재 변수 사용**: SEM은 관측되지 않은 잠재 변수(예: `Pathway`)를 모델에 포함하여 여러 관측 변수(`Orthology`)를 통해 설명 가능, 반면 기존 회귀 분석은 잠재 변수를 직접적으로 모델링하지 않음
- **직접 및 간접 효과 파악**: 기존 회귀 분석은 변수 간의 직접적 관계만을 고려하는 반면, SEM은 `Genus → Pathway → Cancer`와 같은 간접 경로도 분석하여 중간 단계 변수가 최종 변수에 미치는 간접적 영향까지 파악
- **인과 구조 설정**: SEM은 변수 간 인과적 경로 구조를 명시적으로 설정하여 인과 관계를 직관적으로 파악 가능, 반면 회귀 분석은 단순한 예측 관계만 설정하여 인과관계 해석에 제한적


Pathways and Genus Impact on Cancer
Genus2
-0.15
Pathway2
Pathway1
0.20
0.25
Genus1
Genus3
-0.25
0.30
-0.10
Cancer
Genus4