날짜 : 2025-08-01
이름 : 김현지
LLM : GPT4 based perplexity pro

PRACTICE OF EDA

1. 단백질 발현 차이 분석 - 해석

각각의 그래프를 해석하고, 질환별 발현 차이를 보이는 단백질이 무엇인지 정리합니다.

(a~d) 질환별 단백질 발현 차이

NUMBER OF SITES 의 해석 : 여러 기관에서 다수 발견됐다면 재현성이 높은 바이오 마커임.

- 좌측 상단의 극값에 가까운 점들은 높은 유의성과 큰 감소 효과를 의미.
- 우측 상단의 극값에 가까운 점들은 높은 유의성과 큰 증가 효과를 의미.

- **(a) AD, 알츠하이머**
    - 발현이 감소하는 단백질 : VAT1, GPD1, ARPC2, PA2G4 etc.
    - 발현이 증가하는 단백질 : APHE, SPC25, APOB, GDF2, LRRN1.
    - 발현이 증가하는 단백질 중, 특히 ACHE 는 확실한 극값이며, 많은 연구기관에서 발견되어 재현성이 높은 바이오마커로 보임.
    -
- **(b) PD, 파킨스**
    - 발현이 감소하는 단백질 : PRSS8, BAGE3, NPS, PRL, HEXB etc.
      위 현상이 나타나는 연구기관이 전부 2 개여서 특정 site 에서 나타난 결과로 대규모 재현성은 확신할 수 없으나, 유망한 후보 단백질로 보임.
    - 발현이 증가하는 단백질 : SUMF1, PSMC5, DDX1 , VSIR, PRR15
    - 발현이 증가하는 단백질 중, 특히 SUMF1 은 확실한 극값이며, 4-5 개의 연구기관에서 발견되어 재현성이 높은 바이오마커로 보임.
- **(c) FTD, 치매**
    - **특정 site 에서 나타난 결과로 대규모 재현성은 확신할 수 없음.**
    - 발현이 감소하는 단백질 : NPTXR, APLP1, HS6ST3, C1QL3, SEZ6L,NPTXR etc.
    - 발현이 감소하는 단백질 중, 특히 NPTXR 은 확실한 극값이며, APLP1 도 가까움.
      위 현상이 나타나는 연구기관이 전부 2 개여서 유망한 후보 단백질로 보임.
    - 발현이 증가하는 단백질 : STC1, CDSN, PI3, COL6A3, COL28A1
    - 발현이 증가하는 단백질은 y 값이 특별하게 높다고 판단하기 어려워 우연일 가능성도 배제할 수 없음.
- (d) ALS, 루게릭
    - 발현이 감소하는 단백질: ART3, ANTXR2, RGMA, CRTAC1 etc.

- 발현이 증가하는 단백질 : PDLIM3, CA3, APOBEC2, TNNI2, MYLPF, TNNT2, MYBPC1, MYOM2, ACTN2, KLHL41, HSPB6 etc.
- 루게릭에서 작용하는 단백질은 1 개 site 에서만 나타나 재현성이나 임상적 신뢰성은 제한적임.
- 주요 바이오마커 후보 : PDLIM3, CA3, APOBEC2 (증가), ART3, ANTXR2 (감소).
- 특히 PDLIM3, ART3 가 극단적으로 작용하여 후속 실험 대상으로 선정할 법함.

(e~h) 주요 단백질 기능(경로) 분석

질환별 주요 경로는 무엇인지 찾고, 단백질 기능을 질병과 어떻게 해석할 수 있는지 정리합니다.

- **프로세스** :
  전체 단백질(7K SOMASCAN 패널) 중, AD 에서 유의미하게 변화한(증가/감소) 단백질만 추려서, REACTOME PATHWAY 별로 얼마나 모여 있는지, 통계적으로 유의미한 경로가 무엇인지 DOT PLOT 으로 시각화
- Gene Ratio ↑ : 해당 경로 대부분이 영향을 받았음을 의미함.
- p-value ↓ (color) : 통계적으로 매우 유의미함.
- Dot Size ↑ : 많은 단백질이 해당 경로에서 변화한 것 → 시스템 수준의 영향 반영

(e) 알츠하이머 :

- 포도당 대사(Glucose metabolism), 해당과정(Glycolysis)
  → AD 에서 가장 강력하고 재현성 높은 대사 장애 패턴을 나타냄.
- 세포 수송 및 분해 경로 :
  → 단백질 축적, 노폐물 처리 실패 같은 AD 의 병리와 연결됨
- AD 환자에게 진한 점(유의성 높음), 큰 점(많은 단백질), 오른쪽 위치(gene ratio 큼)가 이 경로들에 몰려있음.
- 성장인자 신호 약화
  → 세포 정비 능력 저하 및 진행성 신경 손상과 연결

- 포도당 대사, 해당과정 등 주요 대사경로가 AD 에서 시스템적으로 붕괴되었으며,
  이는 다수 단백질의 대규모 변화와 통계적 엄밀함이 모두 만족되어 실제 환자의 새로운 진단/치료 타깃으로 중요하게 작용 가능할 것임.

(f) 파킨슨병(PD) :

- 베시클(소포체) 매개 운반(Vesicle-mediated transport), 막 이동(Membrane trafficking)

–PD 에서 유의 단백질이 가장 많이 포함(점이 크고 오른쪽, 진한색).
- 신경세포의 신경전달물질 전달/소포 이동 이상이 파킨슨병 핵심 병리임을 재확인.

- Rho GTPase 신호전달(Signaling by Rho GTPases, Miro GTPases 등)
  - 세포 골격 및 신경 돌기의 성장·재생 관련.
  - 신경 구조 및 세포 간 연결 유지에 필수.
- 혈소판 활성화, RAB 단백질, 미토콘드리아 단백질 분해 등
  - 에너지 및 세포내 물질 운반/분해의 이상이 포착됨.

- 파킨슨병은 신경 소포/막 이동, 신경세포 내 운반체계의 구조적 붕괴가 질병의 핵심임을 데이터가 뒷받침.

(g) 전측두엽 치매(FTD) :

- 단백질 번역 후 인산화(Post-translational protein phosphorylation)
  - FTD 에서 가장 유의한 경로 (gene ratio 높음, 점 큼/진함).
  - 단백질 기능·구조 변화, 신호전달 조절 장애가 FTD 병리에 깊이 관련.
- (Insulin-like growth factor) 신호 조절
  - 세포 성장/생존, 신경보호 관련됨.
  - FTD 에서 신경세포 생존·성장 신호 전달이 저하되어 있음.
- 두 경로 모두 대상 단백질 수가 적어 점은 작지만, 색이 진해서(FDR 보정 p 값이 작음) 실제 변화 신뢰도는 높음.
  → FTD 에서는 신호전달 단백질의 다양한 '변형/조절' 경로가 장애의 핵심임.

(h)루게릭병(ALS) :

- 근수축(Muscle contraction), 횡문근(가로무늬 근육) 수축(Striated muscle contraction)
  - ALS 환자에서 유의 단백질이 집중된 경로(점 큼, gene ratio 높음, 진한색).
  - 운동뉴런 퇴화로 근육 기능 상실이 직접적 병리임을 반영.
- 콜라겐 합성 및 수정(Collagen biosynthesis & modifying enzymes, Collagen chain trimerization)
  - 근육·결합조직의 변화도 동반됨.
  - ALS 특이적 근육–결합조직 병리에 연결.
- 큰 점(단백질 많음), 진한색(p-adj 작음), gene ratio 도 높음.
  → ALS 에서는 근육수축 및 구조성분 경로가 시스템적으로 가장 크게 영향,
  임상증상(근육 위축/약화)과도 직결.

(i~j) 임상적 질병 심각도 예측

질환별 주요 경로는 무엇인지 찾고, 특정된 단백질 기능을 질병과 어떻게 해석할 수 있는지 정리합니다.

- **i : 단백질 시그니처 분포**
  - x-axis : CDR, Clinical Dementia Rating, 질병 심각도 → 점수가 높을수록 질병이 심한 상태
  - y-axis : Multivariate protein signature → LASSO 모델로 도출한 단백질 조합의 점수
  - 해석 :
    - Train set 에서 r = 0.68, Test set 에서 r = 0.58
    → 상관계수(r-value)가 높을수록 단백질 시그니처 조합이 실제 임상적 중증도(CDR score)와 잘 일치함을 의미함.
    - 분포적 특징은 CDR 이 높아질수록(y 축) protein signature 점수가 일관되게 증가함.
    - Violin plot 은 각 CDR 별 점수 분포와 중심값이 뚜렷하게 상향 이동하는 것을 볼 수 있음.
    - 실용적 의미
      1. 혈액 기반 단백질 시그니처만으로도 질병 심각도를 꽤 정확하게 예측 가능
      2. 훈련, 테스트 세트 모두에서 유사하게 높은 설명력을 보임(모델 일반화 우수)

- **j : 단백질 시그니처와 CDR 의 관계**
  - 접근 방식 : LASSO 통해 질병 심각도(CDR)를 예측 할 수 있는 단백질 조합 찾음
  - color : CDR 점수 (0.5 ~ 3)
  - r-value : 질병 내에서 CDR 과 단백질 시그니처 간의 상관계수
  - plot : Violin plot 은 데이터 분포와 중심 경향을 동시에 확인할 수 있음
  - 해석 :
    1. FTD 는 상관계수가 0.85 로 가장 높아, 단백질 시그니처가 임상 중증도(CDR)와 매우 밀접하게 일치하여 단백질 변화가 임상 증상 심각도를 가장 잘 반영함.
    2. PD 는 상관계수가 0.70, AD 는 상관계수가 0.55 로 유의한 양의 상관을 보이며, 단백질 시그니처 점수가 높아질수록 CDR(인지기능 저하)도 뚜렷하게 악화함.
    3. CDR 이 높을수록(병이 심할수록) protein signature 값이 일정하게 증가하는 경향을 보임.
    4. 분포: 각 CDR 점수 그룹마다 단백질 시그니처의 중심값이 일관되게 상향되며, 질병 진행을 시계열로 따라갈 수 있음을 시사함.
  - LASSO 기반 단백질 조합 시그니처는 임상 치매 척도(CDR)와 높은 상관관계를 보이며, 특히 FTD 에서 그 설명력이 가장 탁월하다.

2. 장기별 노화 패턴 비교

질환과 장기 노화 연관성

질환별 노화를 가속 또는 저하의 관계를 보이는 장기가 무엇인지 찾고, 해당 장기과 병과의 관계를 어떻게 해석할 수 있는지 정리합니다.

- color

- red : 나이 차이 (pred - actual)가 클수록 질병과 양의 연관성
- blue : 나이 차이가 작을수록 질병과 음의 연관성

-
- AD, 알츠하이머
    - 뇌(Brain), 인지뇌(Cognition brain), 동맥(Artery), 간(Liver), 장(Intestine)에서 노화가 가속됨.
    - 인지 저하와 뇌기능 장애가 핵심이며, 뇌와 혈관, 대사 장기의 노화가 특히 빨라짐.
    - 이로 인해 신경 퇴행과 더불어 전신 대사와 혈관 건강 악화가 함께 작동함을 의미함.
- FTD, 치매
    - 뇌(Brain), 인지뇌, 근육(Muscle), 동맥, 간, 신장, 췌장, 장 등 여러 장기의 노화가 가속됨.
    - FTD 에서는 뇌 뿐만 아니라 근육, 동맥, 신장, 대사장기 등 전신적으로 노화가 가속.
    - 즉, 신경계뿐 아니라 다양한 장기의 노화가 복합적으로 진행되어 환자 임상상이 다양할 수 있음을 시사함.
- ALS, 루게릭
    - 근육(Muscle) 의 노화가 가속되고, 폐(Lung)는 느린 노화와 관련.
    - 파킨슨병에서는 근육 노화가 가장 두드러지게 가속되어, 주요 증상인 운동장애와 직결됨.
    - 반면, 폐는 오히려 노화가 감소(혹은 속도가 늦음)하는 특징이 있어, 주요 병태가 근육과 운동신경계에 집중됨을 강조함.

3. 바이오마커 효과 분석

- 핵심 질문 : APOE ε4 보유 여부가 단백질 발현을 어떻게 바꾸고, 그 변화가 알츠하이머(AD) 특이적인가, 아니면 모든 사람에게 공통적인가?
- (참고) APOE, Apolipoprotein E
    - 지질(지방) 대사에 관여하는 단백질 생성
    - 3 가지 대립형질(allele)를 가지는데 ε4 는 특히 알츠하이머의 위험을 증가시키는 대립형질로 알려져 있음.

연관 단백질 분석

APOE ε4 보유 시 가장 크게 영향 있는 단백질은 무엇인지 정리합니다. (2 개 이상)

- **APOE ε4 보유 시 단백질 발현 변화 분석**
    - APOE ε4 는 유전자형의 한 형태로, 알츠하이머병의 위험을 높임.
    - x 축 (Standardized beta) :
      음수(-): ε4 보유자가 해당 단백질을 덜 가지고 있음을 의미 (발현 감소)
      양수(+): ε4 보유자가 해당 단백질을 더 많이 가지고 있음을 의미 (발현 증가)

- **APOE ε4 와 가장 크게 연관된 단백질**
    - ε4 보유 시 감소 단백질(→ 좌측에 위치):
      NEFL, TBCA, S100A13, FOXO1, BCDIN3D, ARL2, TP53I11 등
      NEFL(Neurofilament light), TBCA(Tubulin folding cofactor A), S100A13 : 신경세포 구조 및 기능과 밀접히 관련됨.
      FOXO1: 세포 노화/사멸과 관련된 전사인자

    - ε4 보유 시 증가 단백질(→ 우측에 위치):
      LRRN1, SPC25, CTF1 등
      LRRN1(Leucine rich repeat neuronal protein 1), SPC25(Spindle pole body component 25) : 세포 구조 및 세포분열, 신경계 발달과 관련됨.
      CTF1(Cardiotrophin 1): 세포 신호, 염증 등 연관.
- 이 변화가 AD 특이적인가, 아니면 공통적인가?
    - 그래프 해석상, APOE ε4 보유와 관련된 단백질 변화는 AD 환자에만 국한되지 않고, 파킨슨병, FTD, ALS 등 주요 신경퇴행성 질환에서도 반복적으로 확인됨.
    - 즉, APOE ε4 에 의해 유도되는 단백질 발현 변화는 'AD 특이성' 보다는 모든 신경퇴행성 질환에서 공통적으로 나타나는 현상.
    - 공통 패턴은 면역 경로/염증 반응, 신경세포 구조, 세포 대사 등 '광범위한 시스템 변동' 으로 해석 가능.

APOE ε4 보유자는 NEFL, TBCA, S100A13(감소), LRRN1, SPC25(증가) 등 단백질 발현 변화가 크며, 이 변화는 알츠하이머뿐만 아니라 공통적으로 주요 신경퇴행성 질환에서 관찰된다.

즉, APOE ε4 의 영향은 전반적인 신경계 노화·손상 환경과 관련된 시스템적 특성을 띤다 할 수 있다.

단백질 RNA 단일세포 연관성 분석

1) **SPC 25 단백질은 어떤 뇌세포와 연관성이 높은지 확인합니다.**
    - SPC2 의 발현이 가장 뚜렷하게 높은 세포 : ' 희돌기아교세포 전구체, OPC'.
    - OPC 에서 SPC25 RNA 발현이 가장 강하게 나타나며, 일반 Oligodendrocyte(성숙 희돌기아교세포)나 다른 신경세포에는 낮거나 거의 발현되지 않음.

2) **해당 뇌세포가 어떤 역할을 하는지 조사하여 정리합니다.**
   - OPC 은 중추신경계에서 미엘린을 합성하는 성숙 Oligodendrocyte(희돌기아교세포)로 분화하는 전구체.
   - 이 세포들은 신경축삭 주위에 미엘린(절연막)을 형성하여 신경전달 효율을 극대화하고, 축삭의 생존 및 회복을 지원하는 중요한 역할을 하며, 뇌의 항상성과 회복, 손상 시 재생에 관여함.

3) **조사한 내용을 토대로 세포가 손상되면 어떤 현상이 발생하여 AD(알츠하이머)와 연관성이 높다고 해석할 수 있는지 정리합니다.**

- OPC 가 손상되거나 SPC25 단백질 기능의 교란이 있는 경우, 희돌기아교세포의 분화 및 증식이 저해되어 미엘린 재생이 어렵거나 느려짐.
- 이로 인해 신경축삭의 탈수초화(demyelination, 미엘린 손실)가 촉진되고, 신경 신호전달 속도 저하, 신경계 염증 반응 증가, 신경세포 퇴화가 유발될 수 있음.
- 실제로 알츠하이머(AD)에서는 OPC 및 Oligodendrocyte 의 미엘린 기능 저하와 이로 인한 신경망 붕괴, 인지기능 저하가 주요 병태로 밝혀져 있음.
- SPC25 가 높은 OPC 가 손상될 경우 AD 에서 관찰되는 신경전달 저하, 신경망 구조 붕괴, 인지기능 감퇴와 같은 현상으로 직접 연결될 수 있음.