**Overview**

이 대회의 목표는 작은 분자들이 다양한 세포 유형에서 유전자 발현을 어떻게 변화시키는지 예측하는 것입니다.

**Description**

인간 생물학은 부분적으로 조직, 기관 및 시스템으로 조직된 인체의 약 37조 세포의 기능과 상호 작용 때문에 복잡할 수 있습니다. 그러나 최근 단일 세포 기술의 발전은 DNA, RNA 및 단백질 수준에서 세포와 조직의 기능에 대한 비할 데 없는 통찰력을 제공했습니다. 그러나 의약품을 개발하기 위해 단일 세포 방법을 활용하는 것은 화학적 동요와 세포 상태에 대한 하류 영향 사이의 인과 관계를 매핑하는 것이 필요합니다. 이러한 실험은 비용이 많이 들고 노동 집약적이며, 모든 세포와 조직이 높은 처리량의 전사체 선별에 적용할 수 있는 것은 아닙니다. 데이터 과학이 새로운 세포 유형의 화학적 동요를 정확하게 예측하는 데 도움이 될 수 있다면 새로운 의약품 개발을 가속화하고 확장할 수 있습니다.  
  
약물 섭동 예측을 위해 몇 가지 방법이 개발되었으며, 대부분은 오토인코더 아키텍처(Dr.VAE, scGEN 및 ChemCPA)의 변형입니다. 그러나 이러한 방법은 다양한 셀 유형을 가진 데이터 세트가 얼마나 잘 일반화하는지 결정할 수 있는 적절한 벤치마킹이 부족합니다. 사용 가능한 가장 큰 훈련 데이터 세트는 NIH 자금 지원 연결 맵(CMAP)으로, 1.3M 이상의 소분자 섭동 측정으로 구성됩니다. 그러나 CMAP에는 모든 유전자의 5% 미만인 978개 유전자에 대한 관찰만 포함되어 있습니다. 또한 CMAP 데이터는 거의 전적으로 암 세포주의 측정으로 구성되어 있으며, 이는 인간 생물학을 정확하게 표현하지 않을 수 있습니다.  
  
단세포 분석 분야의 오픈 문제는 단세포 데이터 과학의 혁신을 이끄는 것을 목표로 하는 비영리 과학 협력체입니다. 이들은 질병의 세포 특징을 연구하고 변경함으로써 의약품을 개발하는 최초의 치료제 회사인 Cellarity와 함께 이 대회를 개최하기 위해 협력하고 있습니다.  
  
비록 모든 세포의 모든 섭동을 측정하는 것은 불가능하지만, 우리는 조합의 부분집합을 측정하고 나머지를 추론하는 것이 가능하다고 가정합니다. 오늘날 우리는 이 목표로부터 멀리 떨어져 있지만, 우리는 이 경쟁이 중요한 개념의 증거가 되기를 희망합니다.  
  
새로운 세포 유형의 화학적 섭동을 정확하게 예측하는 데 도움을 주는 여러분의 연구는 발견을 가속화하고 질병을 치료하거나 치료할 수 있는 새로운 약을 개발하는 것을 가능하게 할 수 있습니다.

**Evaluation**

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

**Submission File**

평가 집합의 각 **id**에 대해 나머지 열에 명명된 18,211개의 유전자 각각에 대한 값을 예측해야 합니다. 각 **id**는 **id\_map.csv** 파일에서 식별할 수 있는 **cell\_type / sm\_name 쌍**에 해당합니다.  
  
제출물에는 머리글이 포함되어야 하며 다음과 같은 형식이 있어야 합니다:

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**Dataset Description**

이 경쟁을 위해 인간 말초혈액 단핵 세포 (PBMC)에서 새로운 단일 세포 섭동 데이터 세트를 설계하고 생성했습니다. 우리는 통합 네트워크 기반 세포 서명 라이브러리 (PMID: 29195078) 데이터 세트에서 144개의 화합물을 선택하고 24시간 치료 후 단일 세포 유전자 발현 프로필을 측정했습니다. 실험은 건강한 인간 기증자 3명에게서 반복되었고, 화합물은 CD34+ 조혈 줄기 세포 (공개되지 않은 데이터)에서 관찰된 다양한 전사 서명을 기반으로 선택되었습니다. 세포가 공개에 대한 사전 동의를 받으면 상업적으로 이용할 수 있고 PBMC는 세포 유형의 주석에 대한 확립된 마커를 가진 여러 성숙 세포 유형 (T 세포, B 세포, 골수 세포 및 NK 세포 포함)을 포함하는 1차 질병 관련 조직이기 때문에 인간 PBMC에서 이 실험을 수행했습니다. 이 데이터 세트를 보완하기 위해 10x 멀티옴 검사를 사용하여 공동 scRNA 및 단일 세포 크로마틴 접근성 측정을 통해 기준선에서 각 기증자의 세포를 측정했습니다. 우리는 각 기증자 및 세포 유형에 대한 풍부한 다중 원자 데이터를 기준선에서 추가하면 차이 생물학적 맥락에서 섭동 반응을 나타내는 특정 유전자의 민감성을 설명하는 생물학적 선행 연구를 수립하는 데 도움이 될 것으로 기대합니다.

(말초혈액단핵세포(PBMC)는 림프구 (T세포, B세포, NK세포), 단핵구, 수지상세포 (Dendritic cell), 줄기세포, 전구세포등으로 구성되어 있습니다. 이 세포는 면역체계의 중요한 구성요소이며, 체액 및 세포매개 면역 반응에서 중요한 역할을 합니다.)

**Technical details about the experiment**

데이터 세트를 이해하려면 처리 효과를 측정하는 데 사용되는 플레이트의 설계를 아는 것이 중요합니다. 기증자의 PBMC를 해동하고 96개의 웰 플레이트에 도금했습니다. 플레이트의 두 열은 양성 대조군(dabrfenib 및 belinostat) 전용이고 한 열은 음성 대조군(DMSO) 전용입니다. 양성 대조군은 전사에 큰 영향을 미치는 경향이 있기 때문에 선택되었으며, 음성 대조군은 본 연구에 사용되는 화합물의 용매로 사용됩니다. 플레이트의 나머지 웰은 72개의 화합물 각각에 할당됩니다. 전체 데이터 세트는 총 6개의 플레이트에 대해 기증자당 2개의 다른 화합물 플레이트로 구성됩니다.

텍스트, 스크린샷, 번호, 도표이(가) 표시된 사진

자동 생성된 설명

참고로, 각 웰은 PBMC를 포함하고 있으며, PBMC는 다양한 세포 유형의 집합체입니다. 여기에는 T 세포, B 세포, NK 세포, 대식세포 및 Monocytes와 같은 Myeloid 세포가 포함됩니다. scRNA에서 측정된 유전자 발현 데이터를 기반으로 각 세포를 계산적으로 세포 유형에 할당할 수 있습니다. 참고로, 우리는 단지 웰 당 ~350개의 세포만을 측정하기 때문에 화합물이 일부 세포 유형에 독성 영향을 미칠 수 있기 때문에 모든 웰에서 모든 세포 유형을 항상 관찰하지는 않습니다.  
  
이 실험에서 원시 데이터에 영향을 미치는 또 다른 기술적 변수는 판의 각 행에 있는 각 우물의 화학적 태그를 지정한 다음 각 행의 모든 표본을 단일 풀로 묶어 순서를 정하는 것입니다. 이것을 셀 다중화라고 하는데, 이에 대한 자세한 내용은 10x 유전체학 웹사이트 셀 다중화란 무엇인가? 당신이 알아야 할 것은 이것이 판의 각 행에 있는 모든 우물을 연결하는 기술적 편향을 만든다는 것입니다. 판의 각 행에 두 개의 양의 대조군과 하나의 음의 대조군을 포함하는 한 가지 목적은 우리가 미분 식을 계산할 때 이 소음원을 설명할 수 있도록 하는 것입니다.

**How we calculate differential expression**

이 경쟁 설정에서 참가자들은 실험 동요가 전사의 모든 유전자(이 데이터 세트의 18211 유전자)의 발현 수준에 미치는 영향을 추정할 수 있는 차등 발현 (DE) 모델링 작업을 수행합니다. 우리는 먼저 각 표본에 있는 특정 유형의 각 세포에 있는 원시 유전자 발현 카운트를 평균함으로써 각 화합물의 영향을 추정하는데, 이를 단일 세포 문헌에서 의사 벌크라고 합니다. 그런 다음 림마를 사용하여 선형 모델을 의사 벌크 카운트 데이터에 적합시키고 라이브러리(행), 플레이트 및 공여체를 기술 공변량으로, 실험 공변량으로 화합물을 포함합니다. 여기서 의사 벌크란 실험의 각 우물에 대해 주어진 유형의 모든 세포에 대한 원시 카운트를 합산했다는 것을 의미합니다.  
  
이 프로세스에 대한 다이어그램은 다음과 같습니다:

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

**What is differential expression**  
생물정보학 및 전산 생물학 백과사전에서: 흔히 DG 또는 DGE 분석으로 약칭되는 차등 유전자 표현은 전사체 내 유전자 전사체의 풍부한 차이에 대한 분석 및 해석을 말합니다(Conesa et al., 2016). 2개의 샘플 세트 간에 다른 유전자 목록은 종종 RNA-seq 데이터 분석 도구에 의해 제공되거나 데이터 세트의 통계적 테스트에 의해 수동으로 생성될 수 있습니다. 검사할 유전자의 수가 많기 때문에(예: 인간 게놈에서 20,000개 이상), 본페로니 보정과 같은 다중 검사 보정이 보통 적용됩니다.  
자세한 내용은 여기서부터 시작하는 것이 좋습니다 : [Single-cell best practices - Differential gene expression analysis](https://www.sc-best-practices.org/conditions/differential_gene_expression.html).

이 모델의 출력은 유전자 발현의 추정된 접힘 변화와 주어진 유전자의 발현이 복합 실험 변수에 의존적인 다중 시험 수정 p-값입니다. 미분 발현 시험의 세계에는 앞으로 가야 할 긴 토끼 구멍이 있습니다. 우리는 scRNA 데이터를 수집하는 데이터 생성 과정에 대한 완전한 기계론적 모델을 가지고 있지 않으며, 많은 그룹들은 성가신 변수나 기술적 잡음을 설명하는 가장 좋은 방법에 대해 의견을 달리합니다. 우리는 림마가 우리의 시험에서 잘 수행되기 때문에 림마를 선택했습니다.  
  
참고로 원시 카운트 데이터와 모든 표본에서 계산된 차분식 분석을 공개하면 검정 집합에서 개인 정보 유출이 발생할 수 있습니다. 이를 방지하기 위해 차분식 모형을 두 번 적합시킵니다. 훈련 데이터를 생성하기 위해 훈련 집합의 표본에만 DE 모형을 적합시킵니다. 비공개 및 공개 검정 데이터를 생성하기 위해 실험의 모든 표본에 DE 모형을 적합시킵니다. 이를 통해 검정 데이터를 비공개로 유지하고 검정 데이터가 가장 정확한지 확인할 수 있습니다.

**Data Split**

Your task is to predict differential expression values for Myeloid and B cells for a majority of compounds!

**Train :**

1. All Compounds in T, NK cells
2. 15 Compounds + positive and negative controls in B and myeloid cells

**Public Test :**

1. 50 randomly selected compounds in B and myeloid cells

**Private Test :**

1. 79 randomly selected compounds in B and myeloid cells

**Note that there is no additional test data beyond the indicated cell\_type / sm\_name pairs.** The input to your model will be a tuple of cell\_type and sm\_name and the output of your model will be predicted signed -log10(p-values) for all 18211 genes.

**Details텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 폰트, 스크린샷, 대수학이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명**