Mouse-Geneformerの学習に使用するrawデータ収集

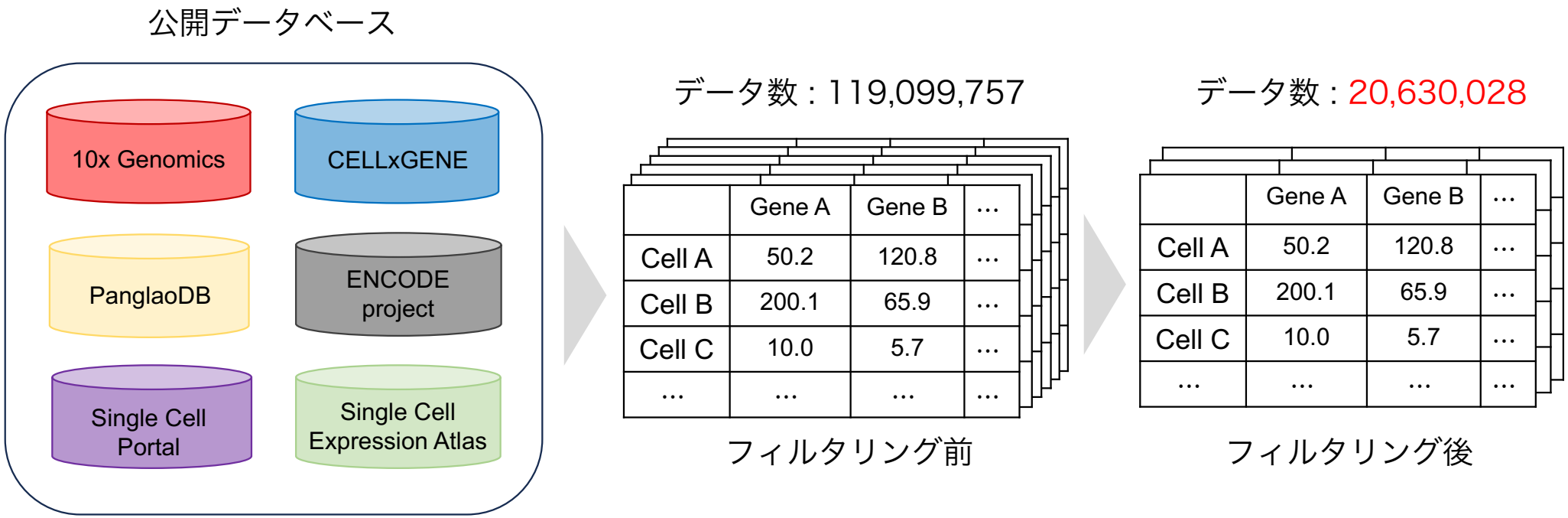
中部大学 工学研究科 ロボット理工学専攻 藤吉研究室 修士2年生 西尾 優希

http://mprg.jp

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

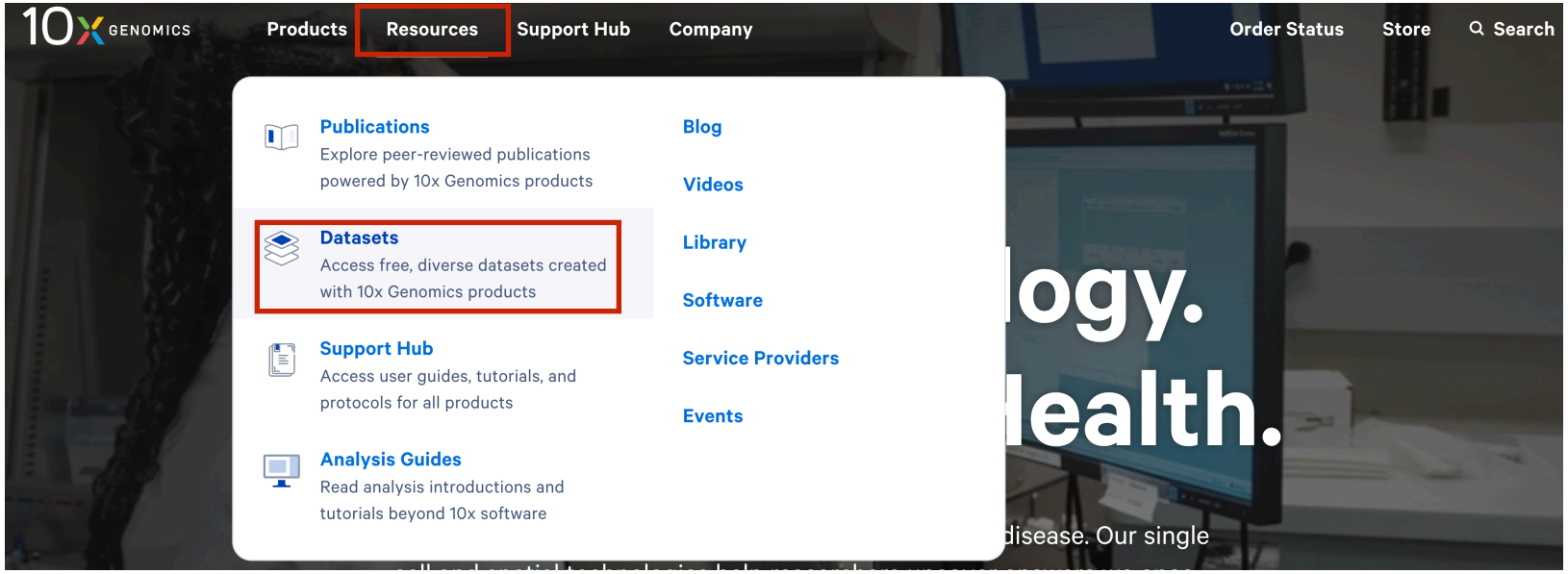
Mouse-Geneformerの学習データ

- Mouse-Genecorpus-20M
  - 複数種類の臓器から収集した単一細胞データセット
    - 10x, cellxgene等の公開データベースから構成



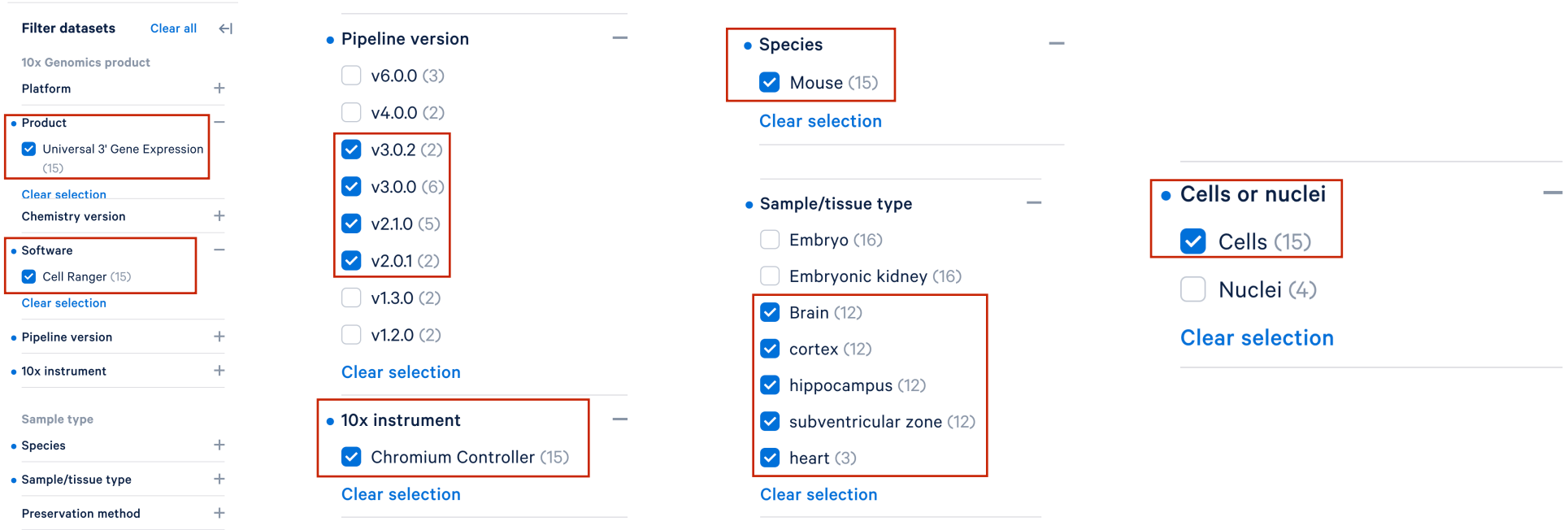
具体的なデータ収集方法

- raw_dataの収集
  - raw_data一覧
  - 今回は10x genomicsを例に説明
    - Resources → Datasetsの順にクリック



具体的なデータ収集方法

- 左側のタブからMouse-Genecorpus-20Mと同様の設定



具体的なデータ収集方法

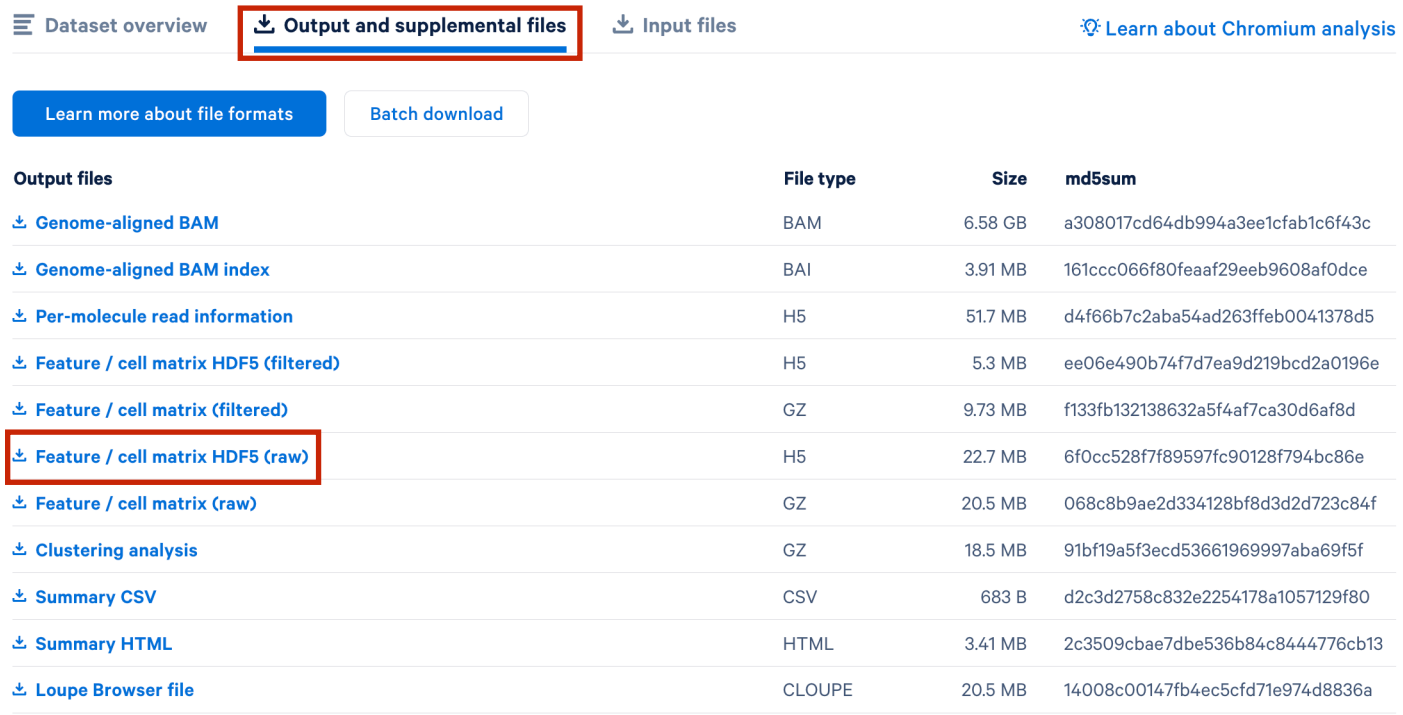
- Datasets (Showing 15 datasets) となっていればOK
  - 複数の結果が出るが，今回は例として 1k Brain Cells from an E18 Mouse v2 を使用

Datasets (Showing 15 datasets)	Product	Species	Sample type	Cells or nuclei	Preservation
5k Cells from a combined cortex, hippocampus and subventricular zone of an E18 mouse (v3 chemistry)	Universal 3' Gene Expression v3.1	Mouse	brain, cortex, subventricular zone, hippocampus	Cells	N/A
5k Cells from a combined cortex, hippocampus and subventricular zone of an E18 mouse (Next GEM)	Universal 3' Gene Expression v3.1	Mouse	brain, cortex, subventricular zone, hippocampus	Cells	N/A
10k Heart Cells from an E18 mouse (v3 chemistry)	Universal 3' Gene Expression v3	Mouse	heart	Cells	N/A
10k Brain Cells from an E18 Mouse (v3 chemistry)	Universal 3' Gene Expression v3	Mouse	brain, subventricular zone, cortex, hippocampus	Cells	N/A
1k Heart Cells from an E18 mouse (v3 chemistry)	Universal 3' Gene Expression v3	Mouse	heart	Cells	N/A
1k Heart Cells from an E18 mouse (v2 chemistry)	Universal 3' Gene Expression v3	Mouse	heart	Cells	N/A
1k Brain Cells from an E18 Mouse (v3 chemistry)	Universal 3' Gene Expression v3	Mouse	brain, cortex, hippocampus, subventricular zone	Cells	N/A
1k Brain Cells from an E18 Mouse (v2 chemistry)	Universal 3' Gene Expression v3	Mouse	brain, cortex, hippocampus, subventricular zone	Cells	N/A

ここをクリック →

具体的なデータ収集方法

- 該当項目をクリックし，Output and.. を表示
  - Feature / cell matrix HDF5 (raw) を選択



Output files	File type	Size	md5sum
Genome-aligned BAM	BAM	6.58 GB	a308017cd64db994a3ee1cfab1c6f43c
Genome-aligned BAM index	BAI	3.91 MB	161ccc066f80feaaf29eeb9608af0dce
Per-molecule read information	H5	51.7 MB	d4f66b7c2aba54ad263ffeb0041378d5
Feature / cell matrix HDF5 (filtered)	H5	5.3 MB	ee06e490b74f7d7ea9d219bcd2a0196e
Feature / cell matrix (filtered)	GZ	9.73 MB	f133fb132138632a5f4af7ca30d6af8d
Feature / cell matrix HDF5 (raw)	H5	22.7 MB	6f0cc528f7f89597fc90128f794bc86e
Feature / cell matrix (raw)	GZ	20.5 MB	068c8b9ae2d334128bf8d3d2d723c84f
Clustering analysis	GZ	18.5 MB	91bf19a5f3ecd53661969997aba69f5f
Summary CSV	CSV	683 B	d2c3d2758c832e2254178a1057129f80
Summary HTML	HTML	3.41 MB	2c3509cbae7dbe536b84c8444776cb13
Loupe Browser file	CLOUPE	20.5 MB	14008c00147fb4ec5cfd71e974d8836a

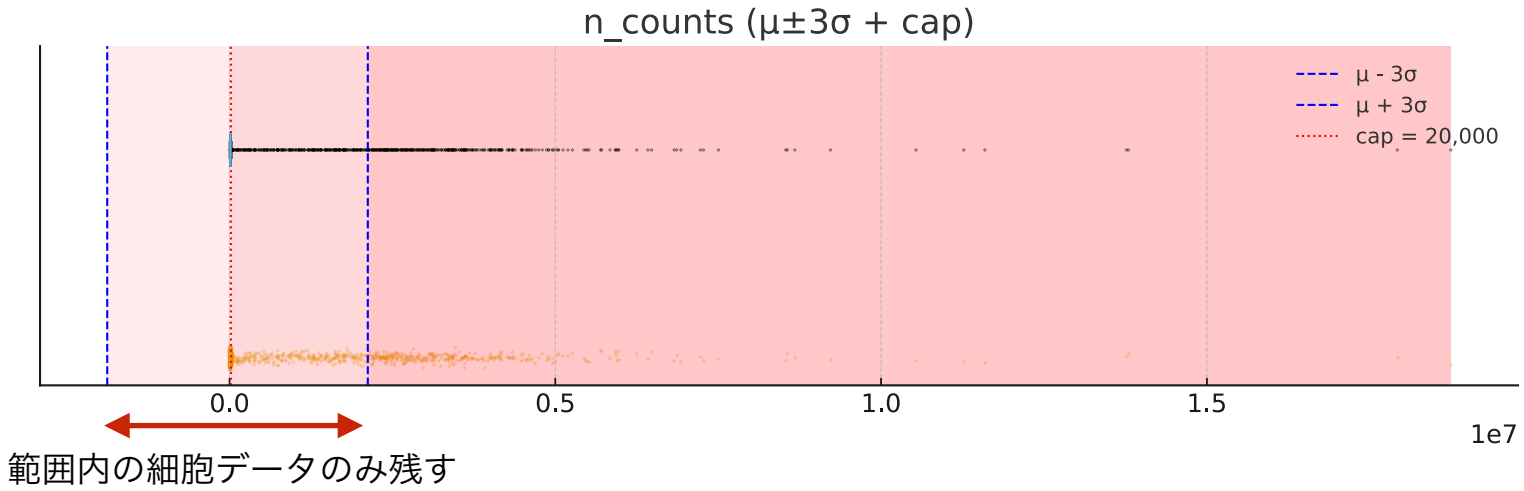
Jupyter notebookによる実習

- raw_dataをフィルタリング&学習データ作成

- data_processing.pyを実行
  - raw_dataとしてh5形式（大規模データ向き）のデータを入力
  - arrow形式（GPUや分散学習に強い）へ変換

収集データのフィルタリング

- Mouse-Genecorpus-20Mでは4つのフィルタリングを実行
  1. 各細胞において，遺伝子発現量が平均値から±3σの範囲外のものを除去
  2. ミトコンドリア遺伝子の発現量に対して平均値から±3σの範囲外のものを除去
  3. 1細胞の発現遺伝子が7未満の細胞を除去
  4. 総発現数が20000を超えるものを除去
     - 異常細胞，空細胞等のノイズを除去



まとめ

- raw_dataは10xやCELLxGENE等の公開データから収集
  - フィルタリング，外れ値除去等で品質を向上

- 具体的な方法
  - 10x : Feature / cell matrix HDF5 (raw)をフィルタリング
  - 保存に向いたデータ形式からGeneformerに向いた形式に変換