BIO-EXPRESS 2.0
USER MANUAL

바이오 익스프레스 2.0 사용자 매뉴얼

BIO-EXPRESS 파이프라인

kobic 국가생명연구자원정보센터
Korea Bioinformation Center

CONTENTS

BIO-EXPRESS 2.0 USER MANUAL

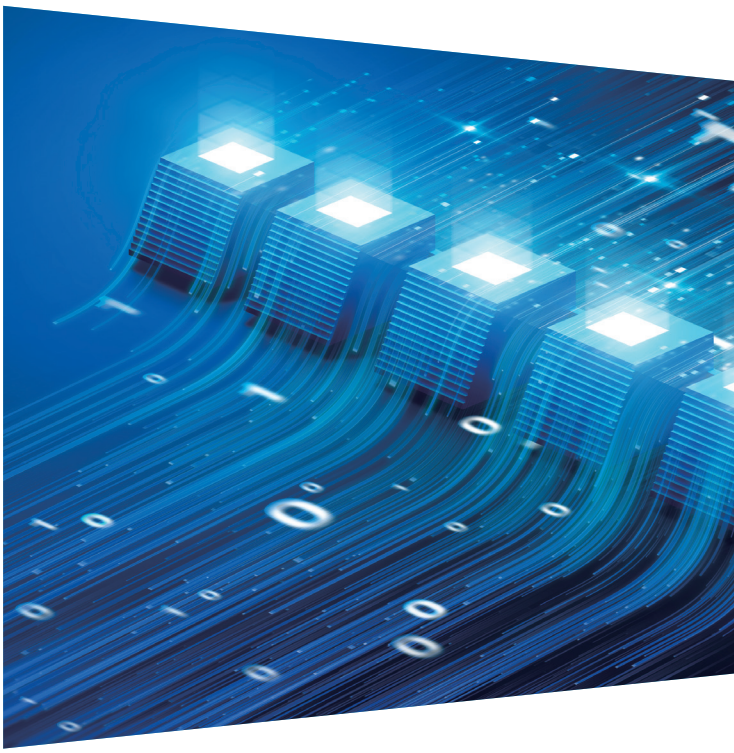
바이오 익스프레스 2.0 사용자 매뉴얼

Bio-Express 파이프라인

BIO-EXPRESS 2.0

국가생명연구자원정보센터 (KOBIC)는

대규모 분석 인프라 또는
고도화된 분석 기술이 필요한
연구자를 위해 Bio-Express 유전체
빅데이터 분석 클라우드 서비스를
제공합니다.



BIO EXPRESS

Bio-Express 서비스는 동적 컨테이너 기반 자동화된 워크플로우 분석 플랫폼과 고속 데이터 전송 서비스를 통해 과학 분야의 빅데이터 분석을 가능하게 하는 클라우드 기반 통합 데이터 분석 서비스입니다.
Bio-Express는 코딩을 모르는 실무자부터 데이터 전문가에 이르기까지 간편한 데이터 분석이 가능한 다양한 클라우드 기반 데이터 분석 서비스와 고속 분석을 위한 빅데이터 플랫폼 기반 인프라 서비스를 제공합니다.

Cloud-Based Open Integrated Analytics Systems

클라우드 기반 개방형 통합 분석 시스템

CLOSHA

Bio-Express 워크벤치는 분석 알고리즘 구성 요소를 이용해 빅데이터 분석 워크플로우를 설계하고 분석을 고속으로 실행하는 플랫폼입니다. 사용자는 원하는 분석 파이프라인 또는 알고리즘의 세트 값을 정의하여 파이프라인을 재구성함으로써 간단하고 신속하게 분석할 수 있습니다. 동적 컨테이너 기반 사용자 코드 실행을 지원하여 자유도가 높은 서비스를 지원합니다.

GBOX

다양한 과학 분야에서 빅데이터의 전송을 위해 KOBIC은 최대 전송 속도, 절감된 파일 전송 비용, 양방향 데이터 전송, 체계적 파일 관리, 신속한 동기화 및 빅 데이터 저장소 백업을 제공합니다.

1.1

유전체 분석 파이프라인 소개

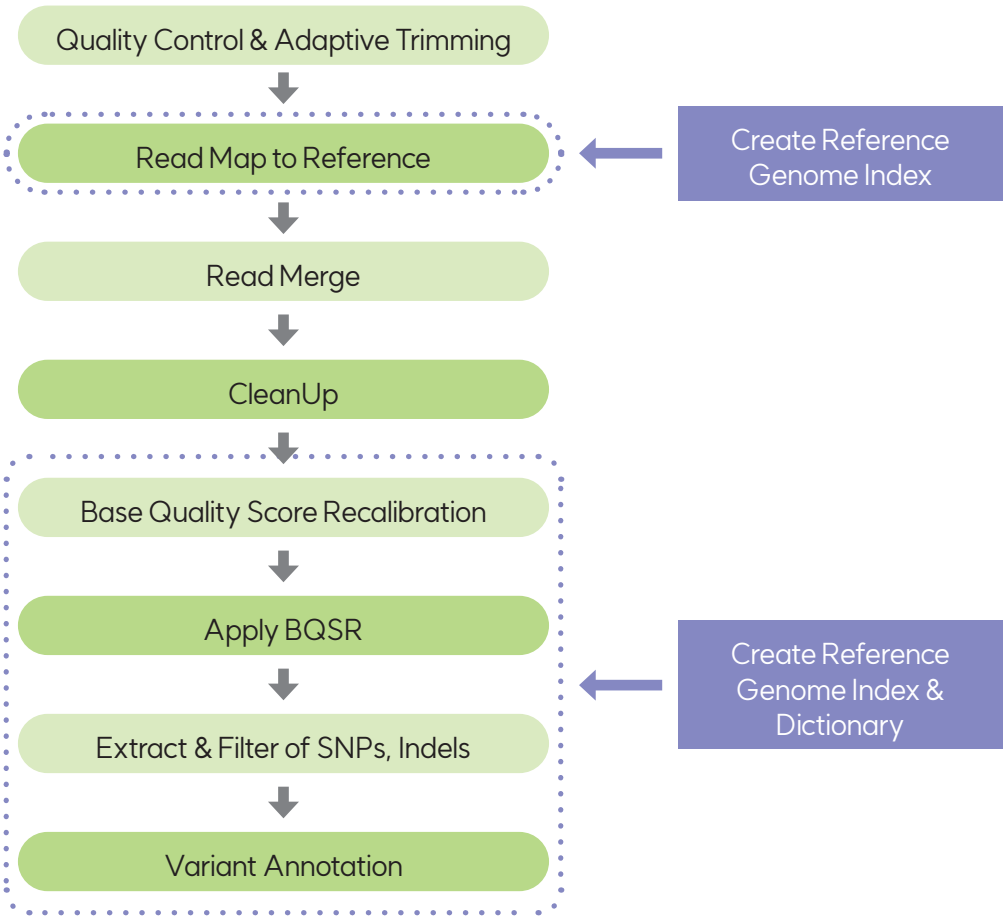
1.1.1 Whole Genome Sequencing Pipeline

● **개요**

바이오 익스프레스에서 제공하는 전체 게놈 시퀀싱(WGS) 파이프라인은 WGS 데이터를 처리하기 위한 모듈식 파이프라인입니다. 이 파이프라인은 FASTQ 파일을 입력으로 사용하고 GATK 파이프라인을 기반으로 Haplotype 호출 결과와 주석 및 시각화를 제공합니다.

● **파이프라인 요약**

파이프라인 모식도



기능 설명

1) FASTQ 형식의 잘 보정된 기본 오류 추정치가 있는 원시 read 데이터가 참조 게놈에 매핑됩니다. BWA 매핑 응용 프로그램은 읽기를 인간 게놈 참조에 매핑하는 데 사용되어 30개 염기 시드에서 두 개의 불일치를 허용하고 기술 독립적인 SAM/BAM 참조 파일 형식을 생성합니다.

2) GATK로 중복 조각을 표시 및 제거하고, 매핑 품질을 평가하고 저품질의 매핑된 read를 필터링하고 모든 mate-pair 정보가 일치하는지 확인하기 위해 pair read 정보를 평가합니다.

3) 그런 다음 로컬 재정렬을 통해 초기 정렬을 세분화하고 의심스러운 영역을 식별합니다. 이 정보를 다른 기술 공변량 및 알려진 변형 사이트와 함께 공변량으로 사용하여 GATK 기본 품질 점수 재조정(BQSR)이 수행됩니다.

4) 재보정 및 재정렬된 BAM 파일을 사용하여 일배체형의 로컬 재조립을 통해 생식계열 SNP 및 INDEL을 호출합니다. 마지막으로 BAM, CRAM 또는 VCF 형식의 시퀀싱 데이터에서 관련성을 신속하게 평가하는 도구인 somalier를 제공합니다.

사용 프로그램

fastp, Cutadapt, BWA, GATK4, somalier

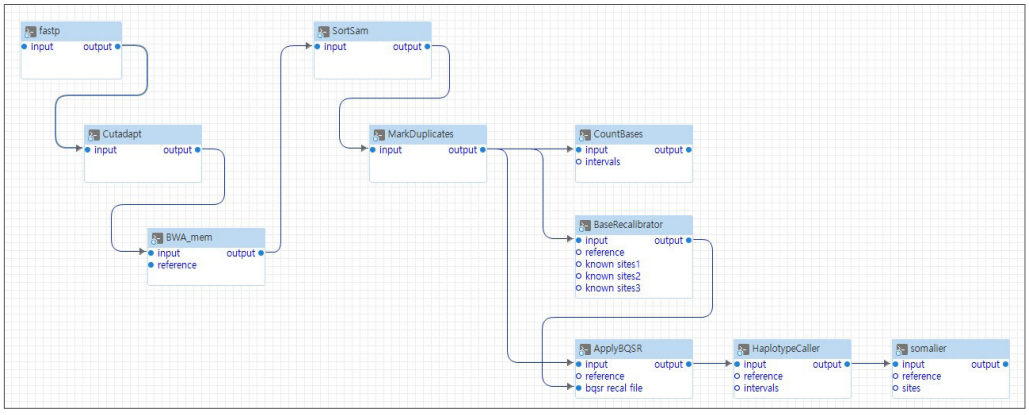
입력 파일

- FASTQ File
- BWA Index File
- Reference FASTA File
- Reference & Resource VCF File

최종 결과 파일

Variant annotate file of Somatic SNP & InDel

● 등록 파이프라인



단계별 설명

Quality Check – Quality Control & Adaptive Trimming

Fastp

- **버전 및 라이선스** : 0. 20.1 / MIT License
- **설명**

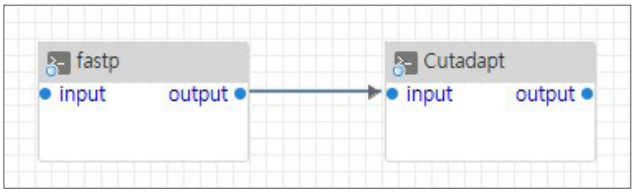
  Fastq 파일에 대한 all-in-one preprocessing을 수행하여 HTML 기반 보고서를 만들어주는 프로그램
- **입력 인자**

  [input] : input – Raw data로 pair-ended 된 fastq.gz 파일이 있는 경로 (path of *.fastq.gz)

  [output] : output – preprocessing의 결과 보고서가 생성될 경로 (path of *.html, *.json, *.fastq.gz)

  [option 1] : length required – 설정한 값보다 긴 read는 폐기 (Integer)

  [option 2] : average qual – 설정한 값보다 품질이 낮은 read는 폐기 (Integer)
- **파이프라인 연결**

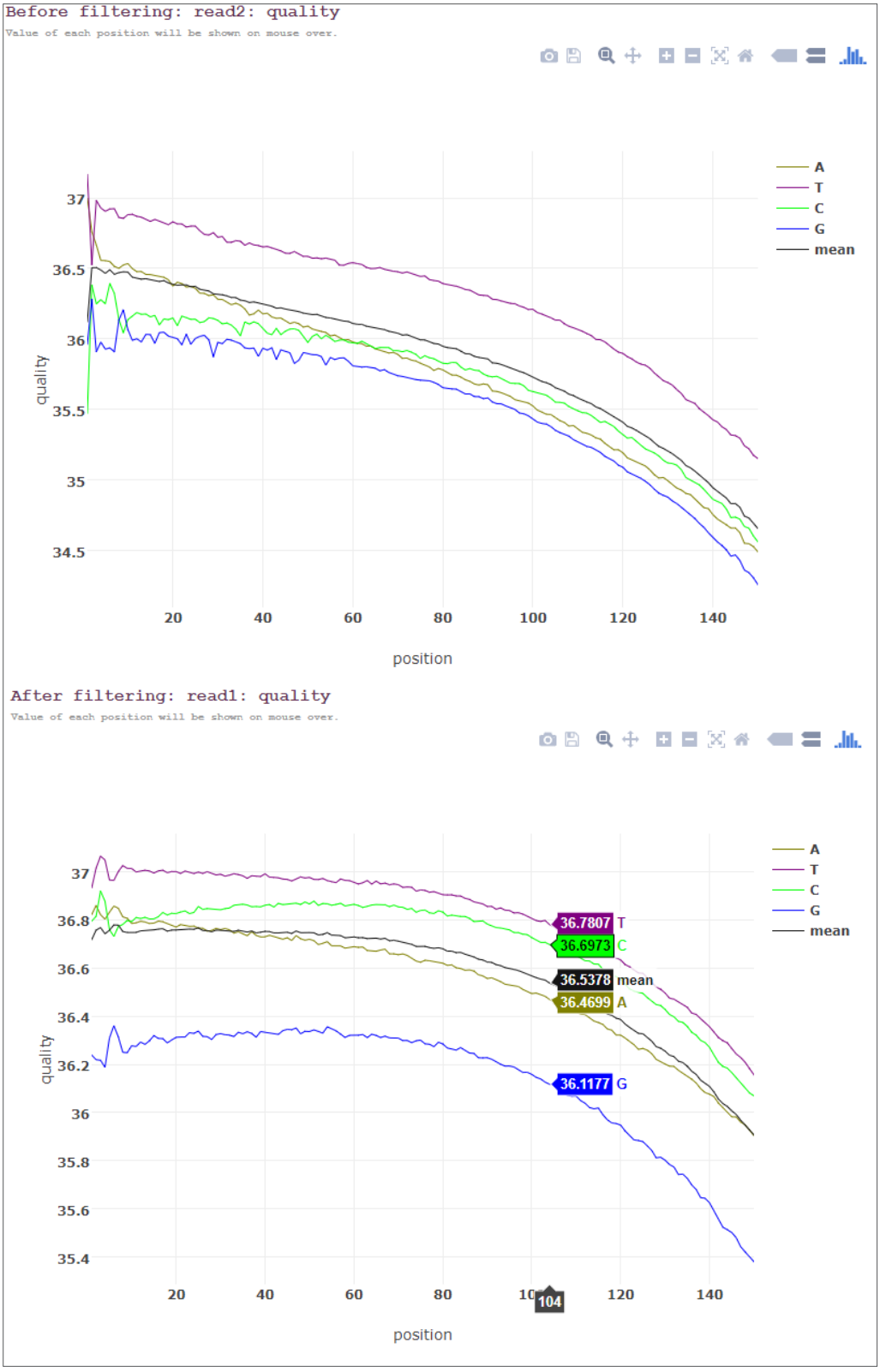


- **결과 파일**

fastp report

Summary

General	
fastp version:	0.21.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (150 cycles + 150 cycles)
mean length before filtering:	150bp, 150bp
mean length after filtering:	149bp, 149bp
duplication rate:	0.387935%
Insert size peak:	269
Before filtering	
total reads:	765.356996 M
total bases:	114.803549 G
Q20 bases:	112.930448 G (98.368429%)
Q30 bases:	108.378770 G (94.403676%)
GC content:	40.268192%
After filtering	
total reads:	765.356890 M
total bases:	114.478725 G
Q20 bases:	112.623980 G (98.379834%)
Q30 bases:	108.090678 G (94.419882%)
GC content:	40.264220%
Filtering result	
reads passed filters:	765.356890 M (99.999986%)
reads corrected:	6.895119 M (0.900902%)
bases corrected:	9.816617 M (0.008551%)
reads with low quality:	0 (0.000000%)
reads with too many N:	106 (0.000014%)
reads too short:	0 (0.000000%)


Before filtering: read2: quality
Value of each position will be shown on mouse over.
A
T
C
G
mean
quality
position
After filtering: read1: quality
Value of each position will be shown on mouse over.
36.7807 T
36.6973 C
36.5378 mean
36.4699 A
36.1177 G
104

Trimming – Quality Control & Adaptive Trimming

Cutadapt

- **버전 및 라이선스** : 3.4 / MIT License
- **설명**

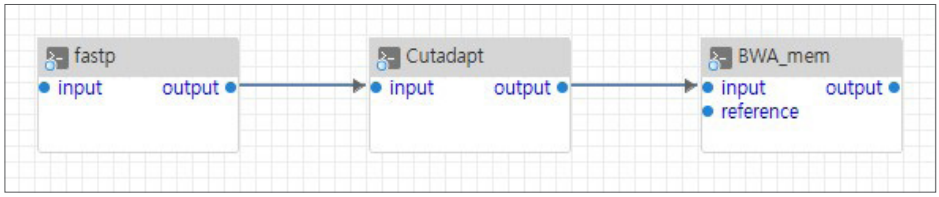
  high-throughput sequencing reads으로부터 adapter sequences, primers, poly-A tails, unwanted sequence 를 찾아 제거하여 정제된 fastq 파일을 생성하는 프로그램
- **입력 인자**

  [input] : input – Fastp에 의해 preprocessing된 fastq.gz 파일이 있는 경로

  [output] : output – Adaptive Trimming의 결과인 fastq 파일이 생성될 경로

  [option 1] : minimum length – 설정한 값보다 짧은 read는 폐기 (Integer)

  [option 2] : pair filter – R1및 R2에 대한 필터를 pair에 대한 단일 결정으로 결합하는방법 (any, both, first)
- **파이프라인 연결**



- **결과 파일**

```
@A00939:131:HTWK7DSXY:3:1101:1054:1000 1:N:0:CAACAATG+CCGTGAAG
ACGTAAAATCAAAGCAGCAGGGAGGGCTCTGCTCCCCAAAAAACTAGAAGTGATCCCAGGTTTTCCCCC
TCACCCTCCACTAACTGCAAGGACAGGTGGAAAATGCATGCTGAGGAAGAAGGTTTGGGGAACAGCCAC
CAGTCTCTACCAT
+
FFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF:FFFFFFFF,FF:FFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFF,FFFFFFF:,FFFFFFF::FFFFF,FFFF:,FFFFFFFFFFFFFFFFFFFFFFFFF:FF
@A00939:131:HTWK7DSXY:3:1101:1127:1000 1:N:0:CAACAATG+CCGTGAAG
AATAGGGATAATTATAGCTACCTTGCAGGGTTGTTGTAAAAATTAGATAAATACAATGTGTTCTATAAATGG
GGACTATATGTAATGAATATTATTTTCATGACTAAATCTCCATTGGAGCTTCACAGCCATTAAGTACTAGTA
GTAATAC
+
FFFF:FFFFFFFFFFFFFFF:FFFFFFFFFF,FFFFFFFFFF:FFFFFFFFFFFF::FFFFFFFFFFFFFFFFFF:
FFFFFFFFF:FFFFFFF:FFFFFFFFFFFF:F:,FF:FFFF,FFFFFFFFF,,F:FF,:FF,FFFFF
@A00939:131:HTWK7DSXY:3:1101:1235:1000 1:N:0:CAACAATG+CCGTGAAG
GGCGCCATGGAGCAGGGGGCGGCGCTCATCGGGGAGGCTCGGGCCGCACAGGAGTCCACCGAGGG
GGTGGGAGGCTCAGGCACGGCGGGCTGCAGGTCCCGAGCCCTGCCCCACGGGAAGGCGGCTAAGGC
CCCGGCGAGAAATCGAGCGC
+
FFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFF,FFFFFFFFFFF
```

Alignment-Read Map to Reference

BWA mem

- **버전 및 라이선스** : 0.7.17 / GNU General Public License version 3.0
- **설명**

  70bp~10mbp query sequences를 최대 완전 일치로 seeding한 다음 정렬하여 SAM 파일로 제공합니다.
- **입력 인자**

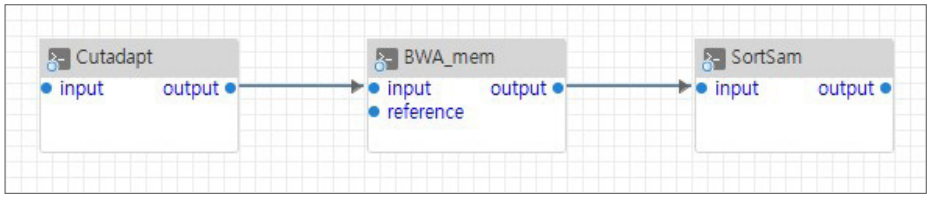
  [input 1] : input – Cutadapt에 의해 Adaptive Trimming 된 FASTQ 파일이 있는 경로

  [input 2] : reference – BWA index에 의해 index가 생성된 Reference FASTQ 파일의 index가 생성된 경로

  [output] : Read Map to Reference 된 SAM 파일이 생성될 경로

  [option 1] : threads – Thread 수 (Integer)

  [option 2] : verbose level – 로그 출력 수준 설정 (Integer)
- **파이프라인 연결**



- **결과 파일**

```
@HD	VN:1.0	GO:none	SO:coordinate
@SQ	SN:HLA-DRB1*16:02:01	LN:11005
@RG	ID:/bioex/output/Cutadapt	PL:Nova6000S4	LB:Truseq_PCRFREE	SM:8888888888
@PG	ID:bwa	PN:bwa	VN:0.7.15-r1140 CL:/BiO/program/bwa/current/bwa mem -M -R
@RG₩tID:/bioex/output/Cutadapt₩tPL:Nova6000S4₩tLB:Truseq_PCRFREE₩tSM:8888888888 -v 1 -t 1
/bioex/example/index/BWA/hg38/references_hg38_v0_Homo_sapiens_assembly38.fasta
/bioex/output/Cutadapt/8888888888-x1_1.fastq /bioex/output/Cutadapt/8888888888-x1_2.fastq
① A00939:131:HTWK7DSXY:3:1101:1054:1000 ② 83 ③ chr10 ④ 14120593 ⑤ 60 ⑥ 151M ⑦ =
⑧ 14120149 ⑨ -595
⑩ ATGGTAGAGACTGGTGGCTGTTCCCCAAACCTTCTTCCTCAGCATGCATTTTCCACCTGTCCTTGCAGTTAGTGGAGGGTGAGG
GGGAAAACCTGGGATCACTTCTAGTTTTTTGGGGAGCAGAGCCCTCCCTGCTGCTTTGATTTTACGT
FF:FFFFFFFFFFFFFFFFFFFFFFFF,:FFFF,FFFFF::FFFFFFF,:FFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF,FFFFFFFF:FFF
FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFF NM:i:2 MD:Z:39G18T92 AS:i:141 XS:i:20
RG:Z:/bioex/output/Cutadapt ⑪
```

Alignment-Read Merge

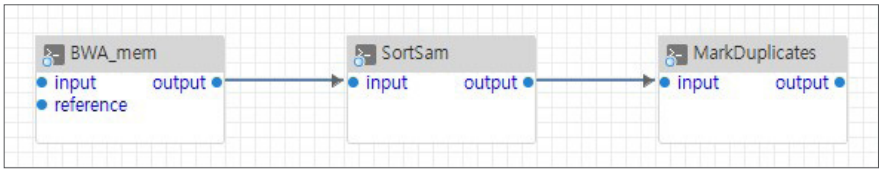
GATK SortSam

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명** : SAM 또는 BAM 파일을 coordinate, query name 또는 SAM 레코드의 다른 속성별로 정렬합니다.
- **입력 인자**

  [input] : input – BWA mem에 의해 생성된 SAM 파일이 있는 경로

  [output] : output – 속성별로 정렬된 BAM 파일이 생성될 경로

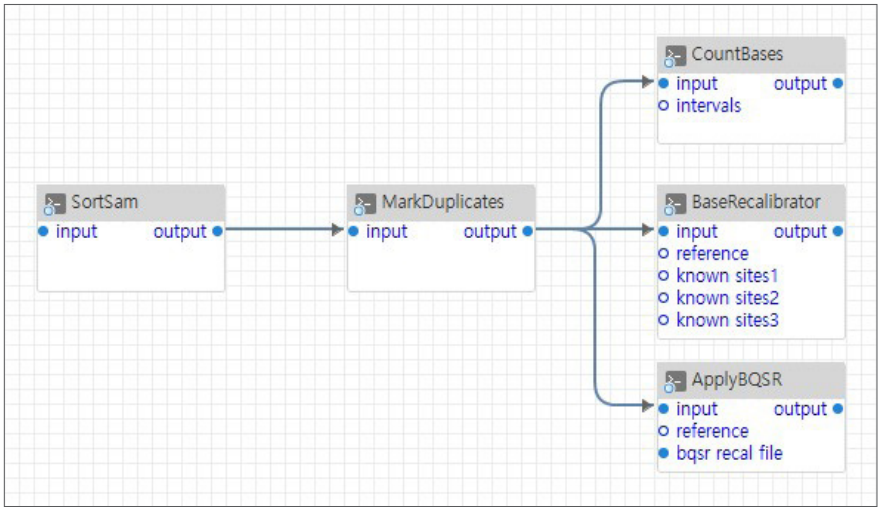
  [option] : sort order – 정렬 순서 설정 (queryname, coordinate, duplicate)
- **파이프라인 연결**



Alignment-CleanUp

GATK MarkDuplicates

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명**
  입력 파일의 다섯 가지 주요 위치에 있는 sequence를 비교하여 중복의 read를 찾아 태그를 지정합니다.
- **입력 인자**
  [input] : input – SortSam 에 의해 생성된 BAM 파일이 있는 경로
  [output] : output – 중복의 read를 찾아 태그가 지정된 BAM 파일이 생성될 경로
  [option] : remove sequencing duplicates – 중복제거 설정 (True, False)
- **파이프라인 연결**



Alignment-Base Count

GATK CountBases

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명**

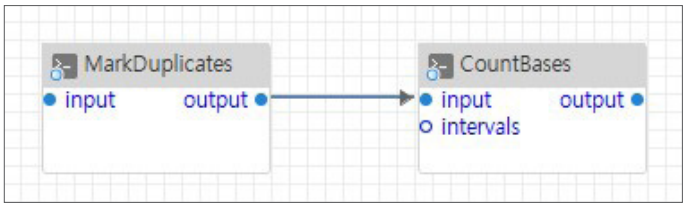
  SAM/BAM/CRAM 파일에 있는 total number of bases를 표준출력으로 count합니다.
- **입력 인자**

  [input] : input – MarkDuplicates 에 의해 생성된 BAM 파일이 있는 경로

  [input 2] : intervals – Genomic Interval 파일이 있는 경로

  [output] : output – BAM 파일의 전체 bases 수가 출력된 COUNT 파일이 생성될 경로

  [option] : read filter – 분석 전에 적용할 filter 설정
- **파이프라인 연결**



Alignment-Base Quality Score Recalibration

GATK BaseRecalibrator

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명**

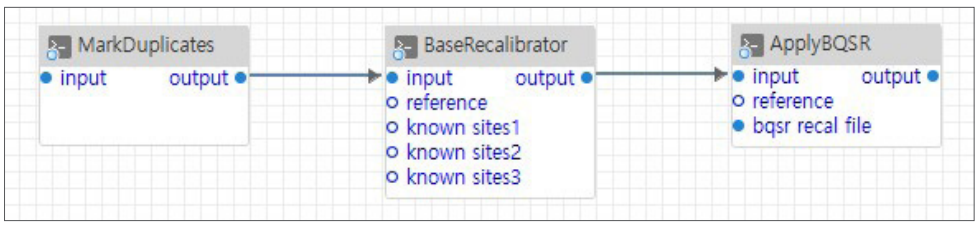
  read group, reported quality score, machine cycle, nucleotide context 등의 다양한 covariates 를 기반으로 base quality score 를 재교정하여 TABLE을 생성합니다.
- **입력 인자**

  [input 1] : input – MarkDuplicates에 의해 생성된 BAM 파일이 있는 경로

  [input 2] : reference – Reference sequence FASTA 파일이 있는 경로

  [input 3~5] : known sites 1~3 – known polymorphic sites인 VCF 파일이 있는 경로

  [output] : output – Base quality score를 재 교정하여 생성된 TABLE 파일이 생성될 경로
- **파이프라인 연결**



- **결과 파일**

```
#:GATKTable:Arguments:Recalibration argument collection values used in
this run
Argument                    Value
binary_tag_name             null
covariate                   ReadGroupCovariate,QualityScoreCovariate,Cont
extCovariate,CycleCovariate
default_platform            null
deletions_default_quality   45
force_platform              null
indels_context_size         3
insertions_default_quality  45
low_quality_tail            2
maximum_cycle_value         500
mismatches_context_size     2
mismatches_default_quality  -1
no_standard_covs            false
quantizing_levels           16
recalibration_report        null
run_without_dbsnp           false
solid_nocall_strategy       THROW_EXCEPTION
solid_recal_mode            SET_Q_ZERO

#:GATKTable:3:94:%d:%d:%d:;
#:GATKTable:Quantized:Quality quantization map
QualityScore  Count     QuantizedScore
           0         0               6
           1         0               6
           2         0               6
           3         0               6
           4         0               6
           5         0               6
           6      3784               6
```

Alignment-Apply BQSR

GATK ApplyBQSR

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명**

  BaseRecalibrator 도구로 생성된 재 교정 테이블을 기반으로 입력 read의 기본 품질을 재교정하여 BAM 또는 CRAM 파일을 생성합니다.
- **입력 인자**

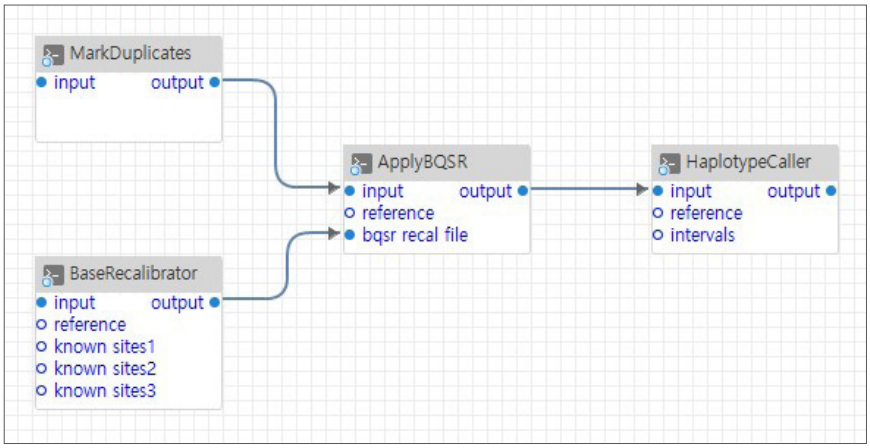
  [input 1] : input – MarkDuplicates 에 의해 생성된 BAM 파일이 있는 경로

  [input 2] : reference – Reference sequence FASTA 파일이 있는 경로

  [input 3] : bqsr recal file – BaseRecalibrator 에 의해 TABLE 파일이 생성된 경로

[output] : output – BaseRecalibrator 에 의해 생성된 TABLE 파일 기반으로 재교정된 BAM 파일이 생성될 경로

- **파이프라인 연결**



Alignment-Extract&Filter of SNPs, Indels

GATK HaplotypeCaller

- **버전 및 라이선스** : GATK 4.2.6.1 / MIT License
- **설명**

  active region 에 있는 haplotypes의 SNP와 indels를 call하여 read를 reassemble 한 결과로 VCF 파일을 생성하고 이를 GVCF.GZ 로 압축합니다.
- **입력 인자**

  [input 1] : input – ApplyBQSR 에 의해 생성된 BAM 파일이 있는 경로

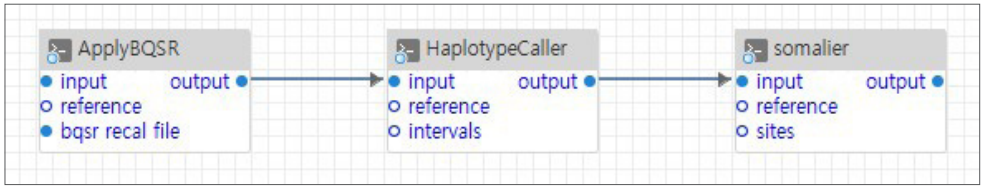
  [input 2] : reference – Reference sequence FASTA 파일이 있는 경로

  [input 3] : intervals – Genomic Intervals list 파일이 있는 경로

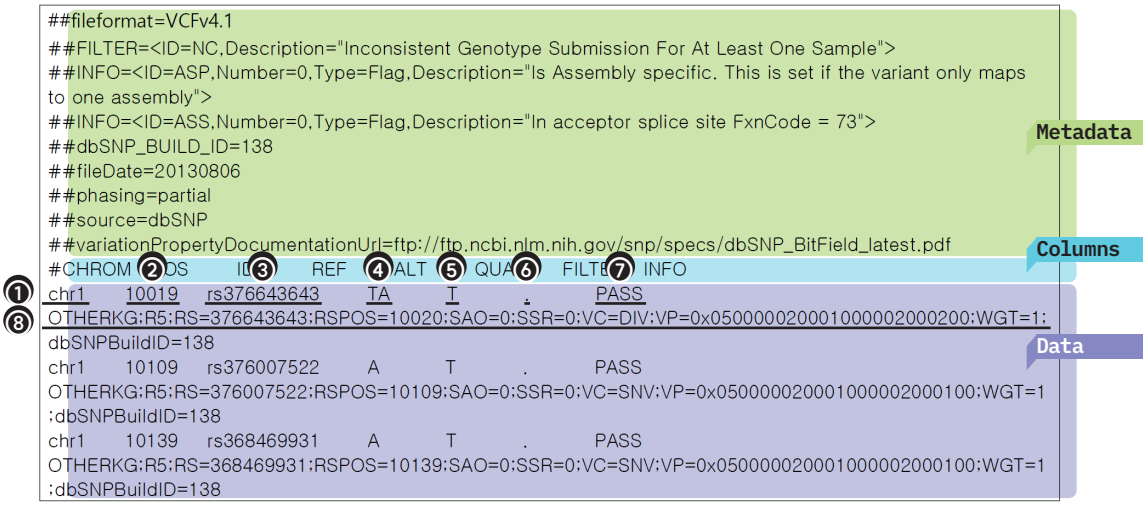
  [output] : output – HaplotypeCaller 를 통해 GVCF.GZ 파일이 생성될 경로

  [option 1] : emit reference confidence – 기준 신뢰도 점수 출력 모드

  [option 2] : output variant index – VCF 인덱스 생성 설정 (True, False)
- **파이프라인 연결**



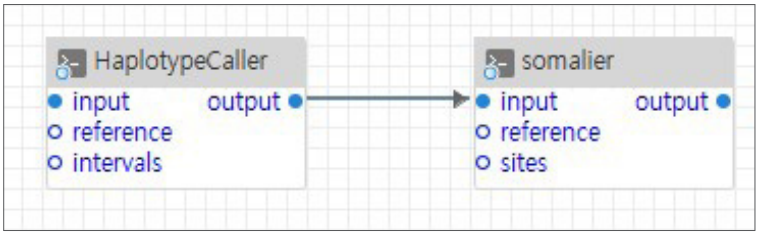
- 결과 파일


##fileformat=VCFv4.1
##FILTER=<ID=NC,Description="Inconsistent Genotype Submission For At Least One Sample">
##INFO=<ID=ASP,Number=0,Type=Flag,Description="Is Assembly specific. This is set if the variant only maps to one assembly">
##INFO=<ID=ASS,Number=0,Type=Flag,Description="In acceptor splice site FxnCode = 73">
##dbSNP_BUILD_ID=138
##fileDate=20130806
##phasing=partial
##source=dbSNP
##variationPropertyDocumentationUrl=ftp://ftp.ncbi.nlm.nih.gov/snp/specs/dbSNP_BitField_latest.pdf
#CHROM ② DS ID ③ REF ④ ALT ⑤ QUA ⑥ FILTE ⑦ INFO
① chr1 10019 rs376643643 TA T . PASS
⑧ OTHERKG;R5;RS=376643643;RSPOS=10020;SAO=0;SSR=0;VC=DIV;VP=0x050000020001000002000200;WGT=1;
dbSNPBuildID=138
chr1 10109 rs376007522 A T . PASS
OTHERKG;R5;RS=376007522;RSPOS=10109;SAO=0;SSR=0;VC=SNV;VP=0x050000020001000002000100;WGT=1
;dbSNPBuildID=138
chr1 10139 rs368469931 A T . PASS
OTHERKG;R5;RS=368469931;RSPOS=10139;SAO=0;SSR=0;VC=SNV;VP=0x050000020001000002000100;WGT=1
;dbSNPBuildID=138
Metadata
Columns
Data

Alignment-Variant Annotation

somalier

- **버전 및 라이선스** : 0.2.12/ MIT License
- **설명**
  BAM/CRAM/VCF로부터 informative sites를 추출하고, sequencing에 따른 relatedness을 평가해주는 프로그램
- **입력 인자**
  [input 1] : input – HaplotypeCaller를 통해 생성된 GVCF.GZ 파일이 있는 경로
  [input 2] : reference – Reference sequence FASTA 파일이 있는 경로
  [input 3] : sites – 추출할 variant의 sites인 VCF 파일이 있는 경로
  [output] : output – somalier를 통해 SOMALIER 파일이 생성될 경로
- **파이프라인 연결**



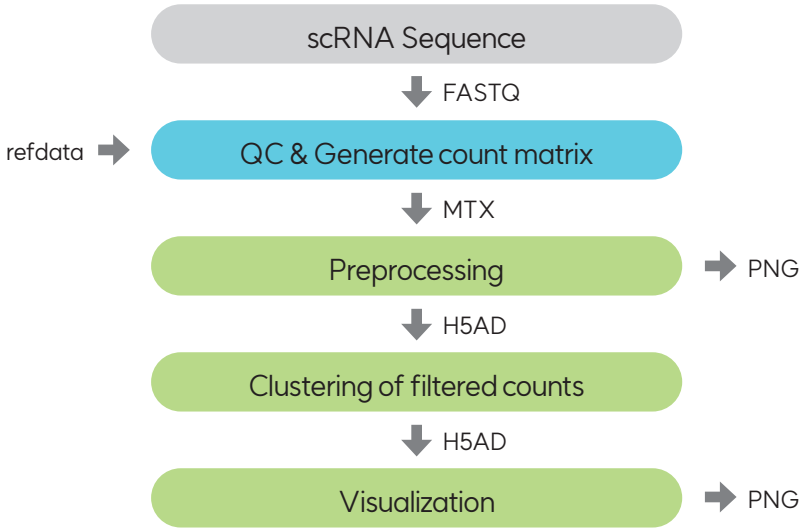
1.1.2 Single Cell RNA Sequencing Pipeline

● **개요**

Bio-express의 Single-cell-RNA-Sequencing-Pipeline은 Scanpy 파이프라인을 사용하여 단일 세포 유전자 발현 데이터를 분석하기 위한 확장 가능한 툴킷입니다. 여기에는 전처리, 시각화, 클러스터링, 미분 표현 테스트 방법이 포함됩니다. Python 기반 구현은 백만 개 이상의 셀로 구성된 데이터 세트를 효율적으로 처리합니다. 주석이 달린 데이터 매트릭스를 처리하기 위한 일반 클래스인 ANNDATA를 제공합니다.

● **파이프라인 요약**

파이프라인 모식도



기능 설명

1) 교란 변수를 회귀분석하고, 정규화하고, 매우 가변적인 유전자를 식별합니다.

2) TSNE 및 다양한 그래프(Fruchterman-Reingold) 시각화 bulk expression과 비교하여 얻은 세포 유형 주석을 보여줍니다.

3) Louvain 알고리즘을 사용하여 셀을 클러스터링하고 플로팅합니다. 또한 다양한 알고리즘을 사용한 클러스터링을 지원합니다.

4) 클러스터에서 차별적으로 발현된 유전자의 순위를 매기고 세포 라벨과 일치하는 마커 유전자를 식별합니다.

사용 프로그램

cellranger, scanpy

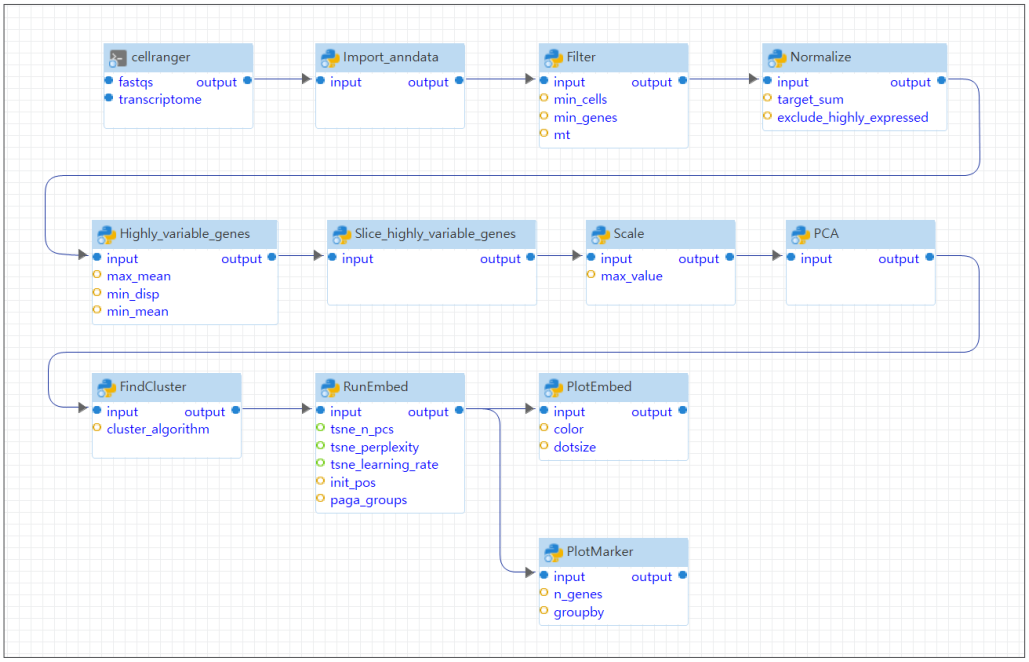
입력 파일

- Single-Cell RNA FASTQ File
- Transcriptome Reference File

최종 결과 파일

다양한 시각화 파일 (.png)

등록 파이프라인



● 단계별 설명

Cellranger – QC & Generate count matrix

Cellranger

- **버전 및 라이선스** : Cell Ranger 6.1.2
- **설명**

  CellRanger는 10X genomics에서 생성되는 Chromium scRNA의 염기서열 데이터를 처리하여 reads align, feature-barcode matrices, clustering을 수행하는 파이프라인입니다. 해당 bash shell은 'cellranger count'만 수행합니다. 이 단계에서는 사용자의 fastq 파일을 가져와서 **1)**reference에 alignment, **2)**filtering, **3)**barcode 카운팅, **4)** UMI 카운팅등을 수행하고 해당 결과와 매트릭스 형태의 결과파일을 제공합니다.

- **입력 인자**

  [input 1] : input – cellranger mkfastq 또는 fastq 파일을 포함하는 폴더에 의해 생성된 fastq_path 폴더의 경로

  [input 2] : transcriptome – Reference file이 있는 디렉토리,
  사용 가능한 reference : Pre-built refence [1], Custom reference build [2]

  [output] : output – 출력 파일을 저장할 상위 경로

  [option 1] : sample – cellranger mkfastq에 제공된 샘플 시트에 지정된 샘플 이름
  (샘플 이름에 허용되는 문자 : 문자, 숫자, 하이픈 및 밑줄)

  [option 2] : exprect cells – 복구된 세포의 예상 수

- **결과 파일**

  [web_summary.html] : 분석 수행한 내용에 대한 요약된 측정 정보

  [cloupe.cloupe] : Loupe Browser에서 사용되는 시각화 파일, 클러스터링 결과에 대한 시각화 정보 제공

  [filtered_feature_bc_matrix] : **1)**feature(유전자), **2)**barcode(각 세포), **3)**matrix(barcode별 feature에 매핑된[UMI] read 개수에 대한 count matrix) 정보. 사용자는 이 정보를 다른 분석 패키지에 이용하는 인풋으로 사용가능(이상치 세포 필터링, 차원 축소, 유전자 발현 정규화 등)

  [filtered_feature_bc_matrix.h5] : filtered_feature_bc_matrix을 HDF5 형식으로 변환한 파일

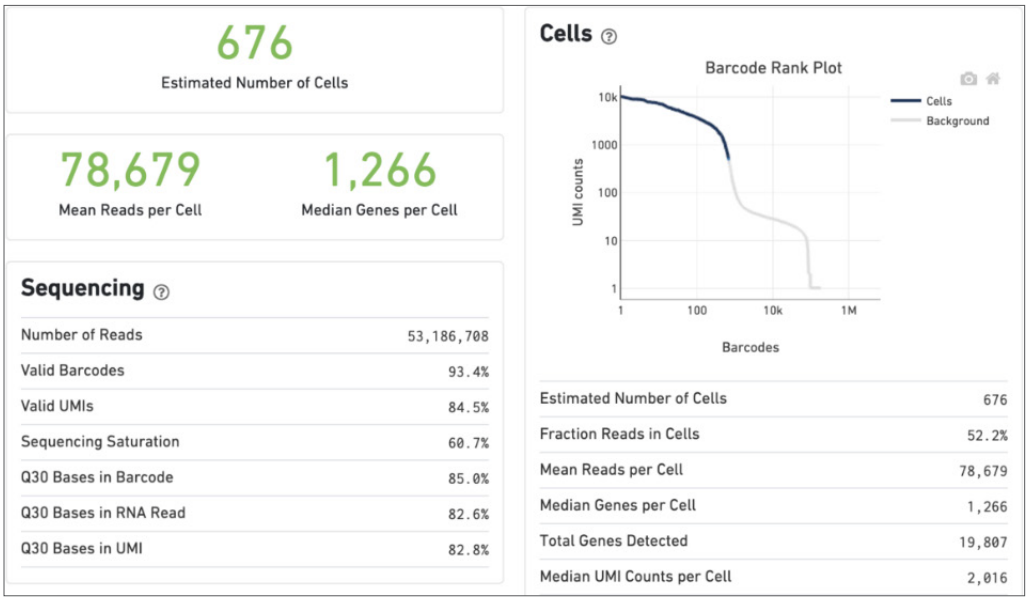
  [raw_feature_bc_matrix] : filtered_feature_bc_matrix와 달리 분석과정에서 QC를 통해 정상 단일 세포로 판정된 barcode뿐 아니라 모든 barcode에 대한 feature, barcode, count matrix에 대한 정보

  [그 외 상세한 결과 파일 정보]

  https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/output/gex-outputs

- **결과 파일 활용**

  [web_summary.html]



Estimated Number of Cells : cellranger를 통해 식별된 세포 수. 예상값 : 500-10,000

Mean Reads per Cell : 시퀀싱된 reads의 총 개수를 예상 세포 수로 나눈 값.

Median Genes per Cell : 모든 세포들(barcode)에서 검출된 유전자의 중앙값

Barcode Rank Plot : cellranger count를 통해 각 세포(barcode)당 유전자(UMI)수를 복원하여 UMI count수로 정렬하여 순서대로 찍은 plot.

[주요 수치에 대한 설명[4]]

1. 제조사에서 권장하는 reads의 수는 일반적으로 최소 30,000-70,000, 샘플당 세포의 수는 최소 500,000 개 이며, 10x Chromium의 한 칸(GEM well)에 loading하는 권장 세포수는 최대 10,000 개 입니다.
2. 라이브러리 제작 시 loading 한 세포수와 [Estimated Number of Cells] 수치와 비교하여 세포의 퀄리티를 확인 할 수 있습니다. 일반적으로 라이브러리 제작시 10x Chromium Chip의 한 칸(GEM well)에 loading한 세포 수가 10,000개면 약 5,700개의 세포를 식별 가능합니다. [Estimated Number of Cells] 수치가 낮다면 loading한 세포 중 죽거나 깨진 세포가 많음을 의미하며, 이런 경우 라이브러리 제작 시 Digitonin 시약을 처리하여 죽은 세포를 제거하는 것을 권장합니다. 이 수치가 낮으면서 Mean reads per cell, Median Genes per Cell 수치가 비정상적으로 높다면 하나의 barcoded bead에 하나의 세포가 들어간 게 아니라 여러 세포가 들어간 multipet인 경우이므로 다운스트림 분석전에 multipet 세포 제거 과정 수행 또는 추가로 library를 제작하는 것을 권장합니다.

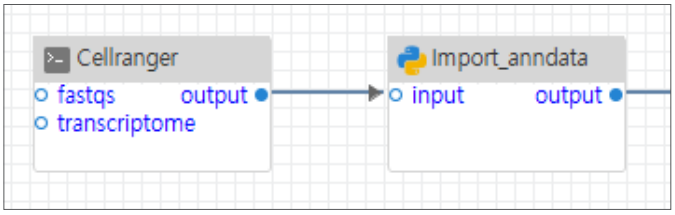
3. Barcode Rank Plot은 각 세포마다 UMI 개수 측정을 통해 예측된 세포들의 상태를 파악할 수 있습니다. 그래프에서 급격히 꺾이는 점을 기준으로 정상세포인지 실제 세포가 들어가지 않은 empty droplet인지 판단하는 cut-off를 정하게 되어 Cells와 Background로 구분합니다. 따라서 이상적인 plot은 cliff and knee라는 불리는 급격히 하락되는 구간이 명확히 나타납니다.
   절벽구간이 두개로 나타나는 경우는 이질적인 세포 집단이 있는 경우이며, 절벽 구간이 없이 완만한 곡선을 이루고 있다면 샘플 품질이 낮거나 라이브러리 제작 과정의 실패로 예상됩니다. 절벽구간이 관찰되나 x축의 barcode수가 일반적인 범위 보다 작다면 droplet 제작 과정에서 샘플 막힘 또는 부정확한 cell counting이 예상되므로 라이브러리 제작을 다시 수행하는 것을 권장합니다.

4. **그 외 살펴볼 수치**

A. Reads Mapped Confidently to Transcriptome : 몇 퍼센트의 reads가 unique하게 transcriptome reference에 제대로 매핑됐는지 확인. Ideal > 30%

B. Fraction Reads in Cells : 유효한 barcode에서 유래한 read 중 해당 세포의 transcriptome에 매핑된 비율. Ideal > 70%

- **파이프라인 연결**



이 모듈의 결과파일 중 filtered_feature_bc_matrix 폴더 경로가 다음 모듈인 Import_anndata의 input으로 사용됩니다.
Filtered_feature_bc_matrix 폴더에는 feature(유전자), barcode(각 세포), matrix(barcode-feature에 대응하는 read count matrix) 정보 파일이 있으며 Import_anndata 모듈을 통해 다음 과정인 차원축소, 클러스터링, 유전자 발현 정규화를 위한 h5ad 형태로 변환 됩니다.

Data load – Preprocessing

Import_anndata

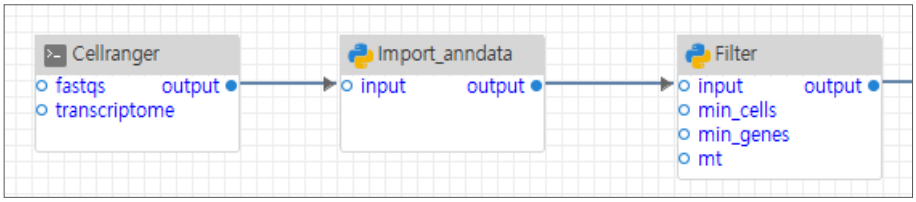
- **버전 및 라이선스** : scanpy 1.9.1, anndata 0.8.0
- **설명**

Ccount matrix를 scanpy에서 사용할 수 있도록 annotated sparse matrix(anndata)로 변경합니다.

- **입력 인자**

  [input] : input – sc-RNA count 매트릭스가 존재하는 경로.
  해당 디렉토리는 outs/filtered_feature_bc_matrix/ 디렉토리를 포함해야하며, 해당 디렉토리 내에는 반드시 다음 3개 파일이 존재해야 합니다.
  (matrix.mtx.gz / features.tsv.gz / barcodes.tsv.gz)
  [output] : output – .H5AD파일을 저장할 경로

- **파이프라인 연결**



Filtering data – Preprocessing

Filter

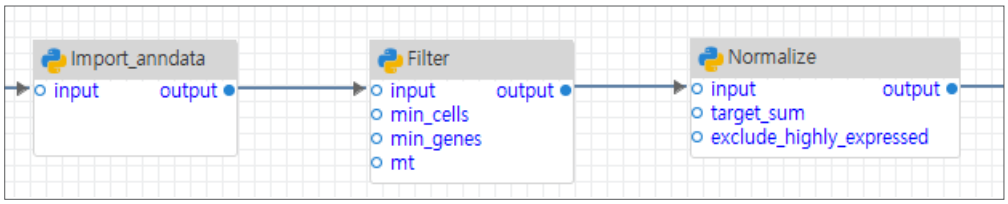
- **버전 및 라이선스** : scanpy 1.9.1, anndata 0.8.0
- **설명**

  cell과 gene 필터링, qc metrics 계산 후 2개의 scatter plot과 3개의 violin plot을 생성합니다.

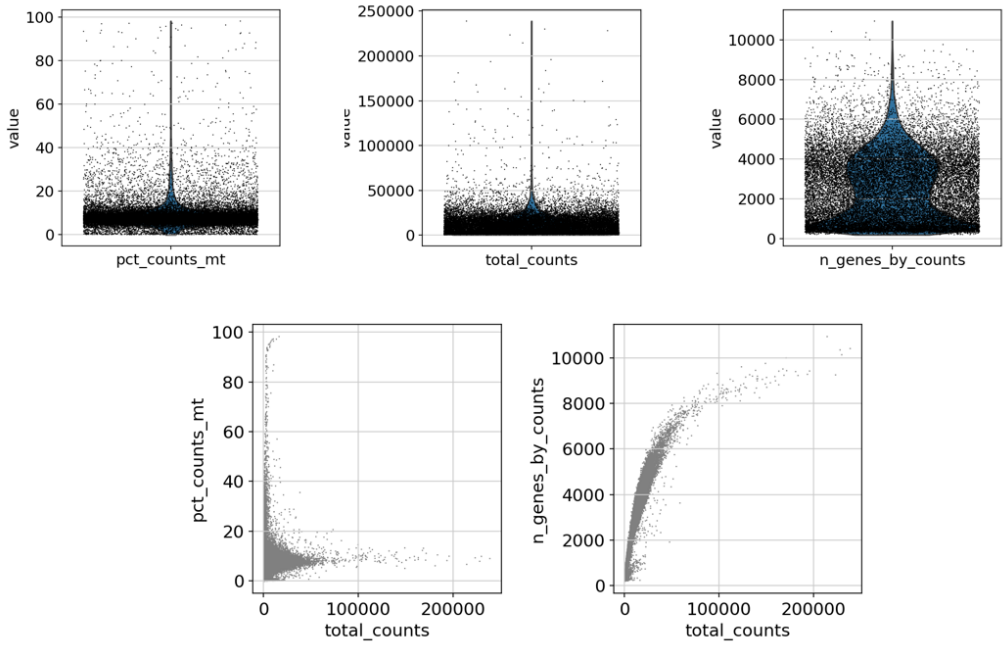
- **입력 인자**

  [input 1] : input – Raw 데이터 anndata가 존재하는 경로
  [input 2] : min_cells – 유전자를 필터링 할 때 필요한 최소 세포 수
  [input 3] : min_genes – 세포를 필터링할 때 필요한 최소 유전자 수
  [input 4] : mt – 제거할 미토콘드리아 유전자 심볼(Lun, McCarthy & Marioni, 2017). MT- 는 인간의 미토콘드리아를 의미
  [output] : output – 필터링된 데이터의 시각화 파일을 저장할 경로
  [option 1] : log1p – log1p 변환 주석을 건너뛰려면 False로 설정
  [option 2] : inplace – anndata의 .obs, .var에서 계산된 지표 배치 여부
  [option 3] : jitter – 바이올린 시각화 그래프에 적용할 지터의 양

- **파이프라인 연결**



- **결과 파일**



- **결과 파일 활용**

  이 plot들은 분석한 데이터셋에서 미토콘드리아의 분포 또는 세포 이중체 또는 다중체 (doublets or multiplets)에서 생산된 비정상적으로 높은 유전자 수를 다양한 plot으로 제공합니다. 따라서 각 plot에서 미토콘드리아의 높은 비율은 세포의 손상으로 인해 품질이 낮은 세포가 많다는 것을 의미합니다. Citing from "Simple Single Cell" workflows (Lun, McCarthy & Marioni, 2017). 이 결과를 참고하여 미토콘드리아 유전자 발현이 너무 많거나 total count(세포당 생산된 total read count)대비 미토콘드리아에서 유래된 read count 비율이 너무 높은 세포를 제거 할 수 있습니다. (제거 기준은 세포마다 다릅니다. 일반적인 세포는 세포내 유전자 중 미토콘드리아 유전자 비율이 10% 내외지만 간 세포와 같이 미토콘드리아가 많이 생산되는 세포는 세포가 손상되어 미토콘드리아 비율이 높은 게 아니기 때문에 그 기준을 더 높게 잡아야 합니다.)

Normalization-Preprocessing

Normalize

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  전체 유전자에 대한 총합으로 정규화를 진행하여 모든 세포가 같은 총합을 갖도록 합니다.
- **입력 인자**

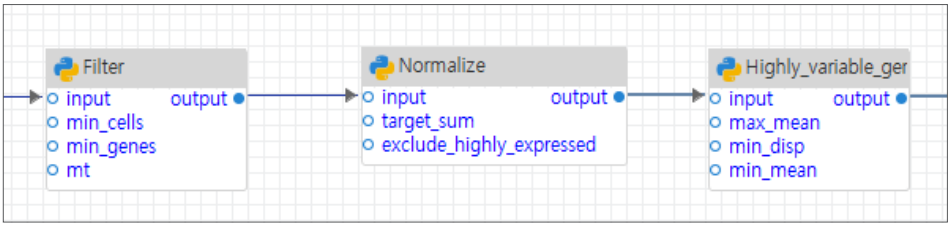
  [input 1] : input – Raw 데이터 anndata가 존재하는 경로

  [input 2] : target_sum – 정규화 후의 총 합 결정

  [input 3] : exclude_highly_expressed – True인 경우, 높게 발현된 유전자는 각 셀에 대한 정규화 계산에서 제외

  [output] : output – 필터링된 데이터를 저장할 경로

  [option 1] : log1p – True인 경우, 데이터 행렬 로그화
- **파이프라인 연결**



Find highly variable genes – Preprocessing

Highly_variable_genes

- **버전 및 라이선스** : scanpy 1.9.1, anndata 0.8.0
- **설명**

  개체 변이가 큰 유전자(HVGs) 계산합니다.
- **입력 인자**

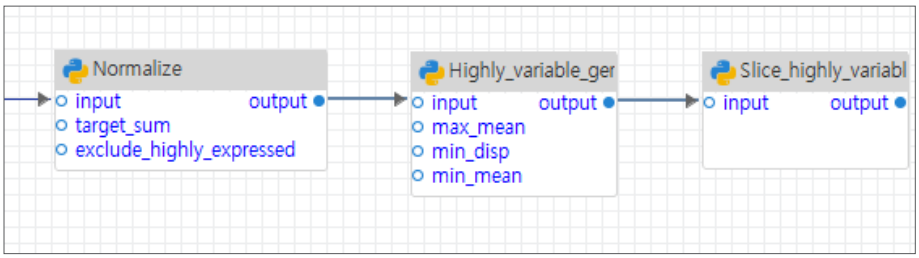
  [input 1] : input – 정규화된 Raw 데이터 anndata가 존재하는 경로

  [input 2] : max_mean – 개체변이가 큰 유전자의 최대 평균

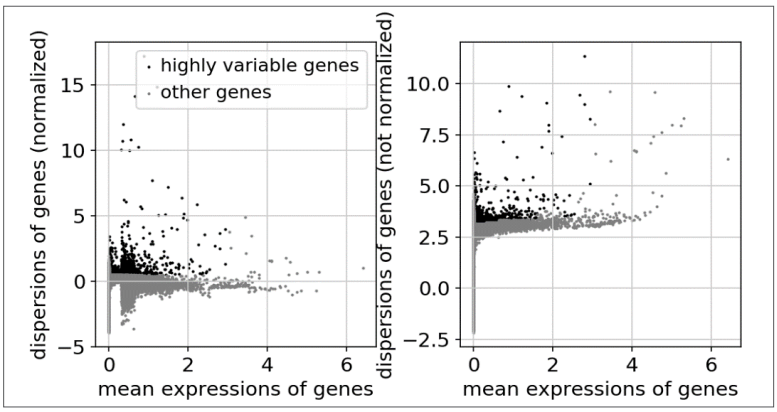
  [input 3] : min_disp – 개체변이가 큰 유전자의 최소 분산

  [input 4] : min_mean – 개체변이가 큰 유전자의 최소 평균

  [output] : output – HVGs 마킹정보가 기록된 anndata와 시각화 파일을 저장할 경로
- **파이프라인 연결**



- **결과 파일**



- **결과 파일 활용**

  이 결과는 데이터셋에서 세포간 유전자 발현 변동이 높은 유전자 집합을 계산합니다. 이러한 유전자에 초점을 맞추는 것이 단일 세포 데이터 세트에서 생물학적 신호를 강조하는데 필요한 절차 입니다. 이 결과 값은 PCA와 클러스터링 등 다운스트림 분석에 사용됩니다. 연구 내용에 따라 Highly variable genes 중에 연구에 관련이 없지만 세포간 변동이 큰 유전자는 다운스트림 분석에서 방해가 되기 때문에 제거 한 후 다시 한번 HVG를 수행하여 데이터 셋을 정제 할 수 있습니다.

Data slicing – Preprocessing

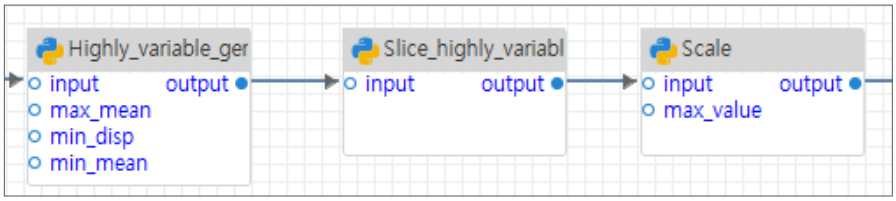
slice_highly_variable

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  컴퓨팅 자원 효율을 위해 HVGs를 제외한 나머지 유전자들을 삭제합니다.
- **입력 인자**

  [input 1] : input – HVGs 마킹정보가 저장된 anndata가 존재하는 경로

  [output] : output – 불필요한 유전자를 잘라낸 anndata를 저장할 경로
- **파이프라인 연결**



Data slicing – Preprocessing

scale

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  regress out과 scaling을 수행

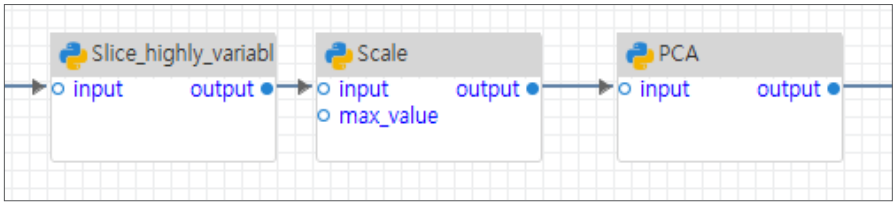
  기본 세팅은 scaling만 수행하는 것입니다. 데이터를 단위 분산 및 zero mean으로 scaling합니다. 모든 특성의 범위(또는 분포)가 같아지게 됩니다. Regress out은 특히 작은 데이터에서 과도한 가변 조정이 일어나 데이터 손실을 초래할 수 있기 때문에 필요 시에만 사용하도록 합니다.
- **입력 인자**

  [input 1] : input – HVG가 slice된 anndata가 존재하는 경로

  [input 2] : max_value – 스케일링 후 이 값에 맞춤

  [output] : output – 스케일링이 완료된 anndata를 저장할 경로

  [option] : regress_out – regress out 실행 여부
- **파이프라인 연결**



PCA - Embedding

PCA

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  PCA를 수행. FIt-SNE, T-SNE 등의 embedding을 수행하기 위해 데이터를 더 작은 차원으로 축소합니다. 단일 세포 데이터 세트는 수십만개에 이르는 세포에 대해 각 세포당 2~3만개의 유전자 발현량을 수치화한 고차원 정보입니다. 이런 고차원 정보를 있는 그대로 활용하면 계산시간이 오래 걸릴 뿐만 아니라, 각종 노이즈가 분석 결과에 포함될 수 있습니다. 따라서 세포당 약 2~3만개의 유전자들을 잘 조합해서 수 개에서 수십 개 정도의 차원으로 축소 시키는 과정이 아주 중요합니다.
- **입력 인자**

  [input 1] : input - 데이터 스케일링이 완료된 anndata의 파일 경로

  [output] : output - anndata 수행이 완료된 anndata와 pca loading, variance ratio plot 파일을 저장할 경로

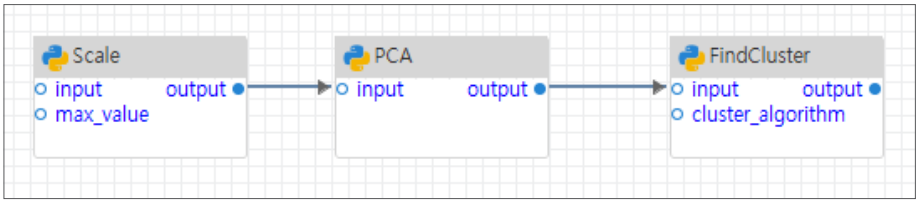
  [option 1] : 사용할 SVD solver 선택

  [option 2] : n_comps - pca의 컴포넌트 수

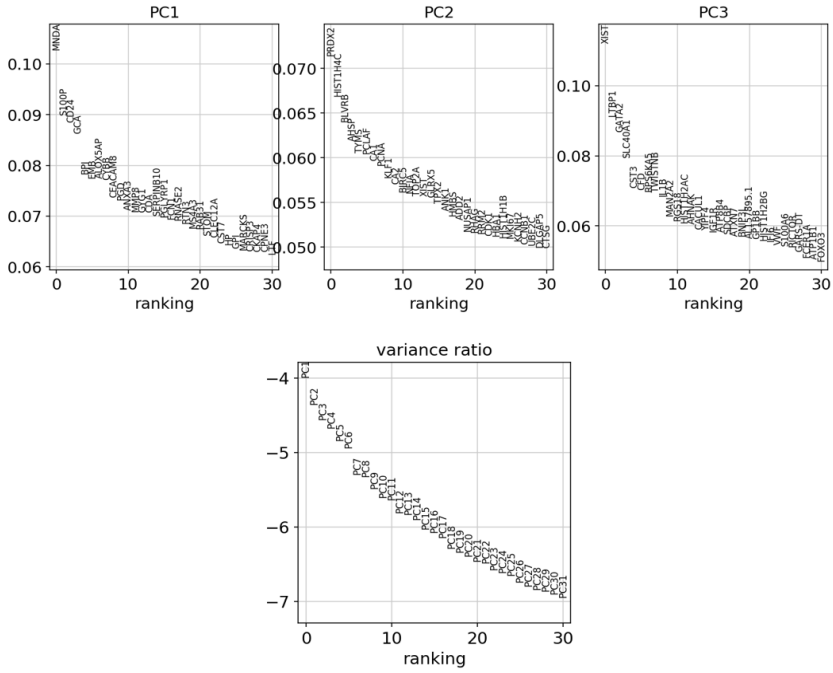
  [option 3] : components - pca_loading.png의 파라미터

  [option 4] : include_lowest - 로드가 가장 높은 변수와 가장 낮은 변수를 모두 표시할지 여부

  [option 5] : log - variance ratio를 로그 스케일로 시각화
- **파이프라인 연결**



- **결과 파일**



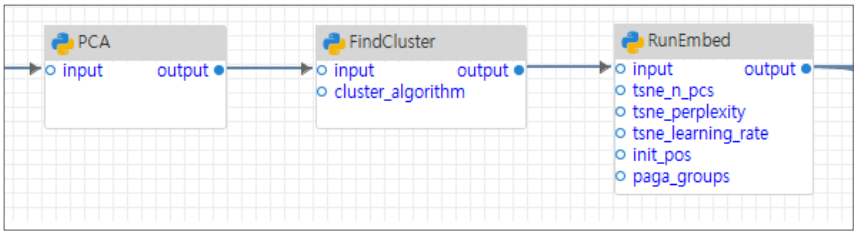
- **결과 파일 활용**

  이 결과는 단일 세포 데이터 세트에 대한 데이터의 차원을 줄이기 위한 PCA 분석 결과 입니다. 데이터 세트의 전체 분산에 대한 단일 PC의 기여도를 조사하여 세포간 이웃 관계를 계산하기 위해 고려해야 하는 PC 수에 대한 정보와 해당 유전자 정보를 제공합니다.

Neighborhood graph & Find cluster - Embedding

FildClsuter

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**
  군집화를 수행하기 위해서 neighborhood graph를 생성하고 이를 기반으로 leiden 또는 louvain 알고리즘을 수행하여 cell들의 군집 카테고리를 정의
- **입력 인자**
  [input 1] : input - pca를 수행한 anndata의 파일 경로
  [input 2] : cluster_algorithm - 클러스터링 알고리즘 선택. (leiden 또는 louvain)
  [output] : output - 클러스터링 label column이 존재하는 anndata를 저장할 경로
  [option] : n_neighbors - 인접 데이터 포인트의 수
- **파이프라인 연결**



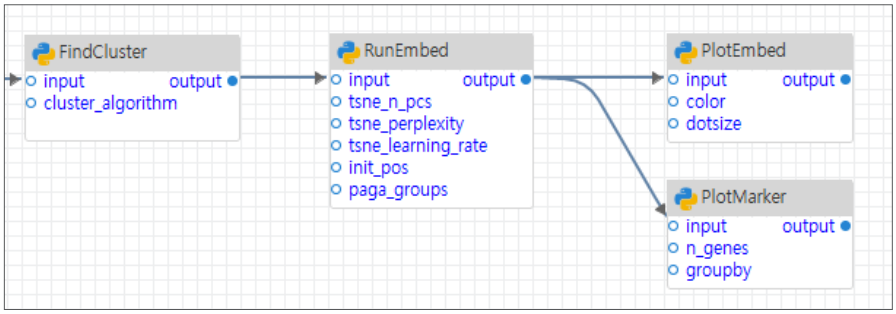
FIt-SNE, T-SNE, UMAP - Embedding

RunEmbed

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**
  FIt-SNE, T-SNE, UMAP 알고리즘을 수행합니다.
- **입력 인자**
  [input 1] : input - leiden 또는 louvain의 카테고리 column이 존재하는 anndata 파일 경로
  [input 2] : tsne_n_pcs - TNSE 수행 시 사용할 PC의 개수
  [input 3] : tsne_perplexity - perplexity는 매니폴드 학습 알고리즘에서 사용되는 nearest neighbors의 수와 관련이 있습니다. 데이터셋이 클수록 큰 perplexity값을 사용합니다. 5에서 50 사이가 가장 적절한 값입니다.
  [input 4] : tsne_learning_rate - TSNE 수행 시 필요한 learning rate. 100에서 10000 사이의 값을 선택합니다.
  [input 5] : init_pos - umap 임베딩 시 초기화될 값. paga, spectral, random 중 한가지를 선택합니다. 기본값은 paga입니다.

[input 6] : paga_groups – leiden 또는 louvain을 선택. 기본값은 leiden입니다.
[output] : output – 임베딩된 데이터가 존재하는 anndata를 저장할 경로
[option 1] : umap_min_dist – 포함된 점 사이의 유효 최소 거리
[option 2] : umap_spread – 포함된 점의 유효 척도
[option 3] : umap_alpha – 임베딩 최적화를 위한 초기 학습률
[option 4] : umap_gamma – 저차원 임베딩 최적화에서 음성 샘플에 적용되는 가중치

- **파이프라인 연결**



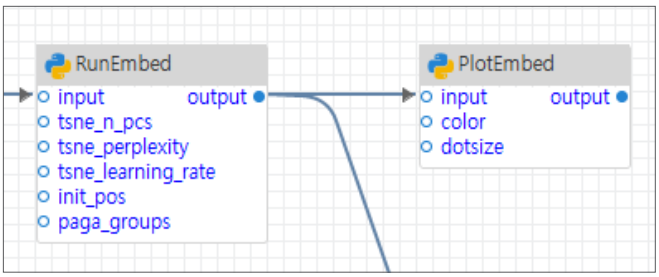
Visualize embedded data – Visualization

PlotEmbed

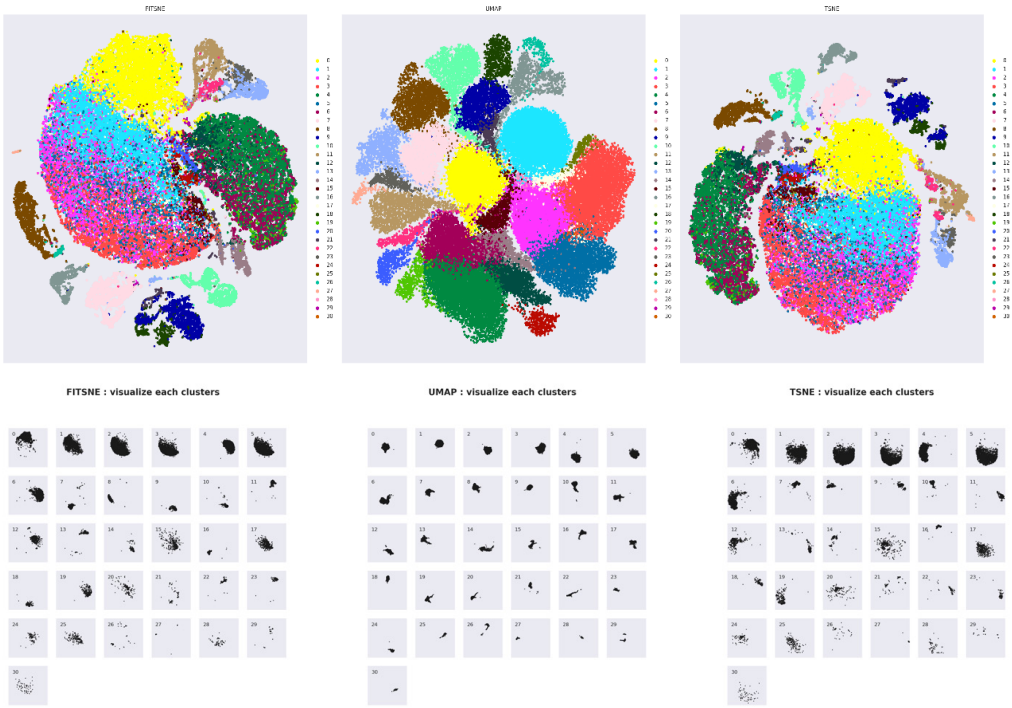
- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  FIt-SNE, T-SNE, UMAP 결과를 시각화합니다.
- **입력 인자**

  [input 1] : input –FIt-SNE, T-SNE, UMAP 결과가 포함된 anndata가 존재하는 파일 경로
  [input 2] : color – plot에 표시할 클러스터링 알고리즘. louvain 또는 leinden을 선택
  [input 3] : dotsize – plot의 dot size를 설정
  [output] : output – 시각화 파일을 저장할 경로
  [option] : figsize – 시각화 플롯의 크기 지정
- **파이프라인 연결**



- **결과 파일**



- **결과 파일 활용**

  이 결과는 단일 세포 데이터 세트를 시각화하고 탐색하기 위해 tSNE, FIt-SNE, UMAP과 같은 여러 비선형 차원 축소 기술을 통해 생산된 plot을 제공합니다. 지금까지 차원 축소를 거친 데이터는 세포 하나하나마다 저차원 공간의 좌표라고 할 수 있는 고유의 좌표를 갖게 됩니다. 따라서 비슷한 세포들은 비슷한 좌표를 갖게 되며 이 저차원 공간에서 군집을 이루어 특정한 기하학적 형태를 구성하게 됩니다. 따라서 결과로 위의 plot과 같이 특정 좌표위에 각 세포를 표현하고 비슷한 세포(생물학적 신호, 유전자 발현 패턴이 유사한) 끼리 군집을 색상별로 표현되어 제공합니다.

Visualize marker genes – Visualization

PlotMarker

- **버전 및 라이선스** : scanpy 1.9.1
- **설명**

  DEG(differentially expressed genes)를 찾아서 시각화한 다양한 plot과 Heatmap, stacked violin plot, 각각의 embedding plot을 기반으로 한 marker gene들의 활성화 plot을 제공합니다.
- **입력 인자**

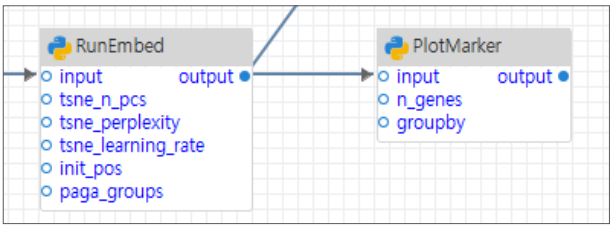
  [input 1] : input – embedding이 완료된 anndata가 존재하는 파일 경로

  [input 2] : n_genes – 각 embedding plot에 표시할 marker gene의 개수

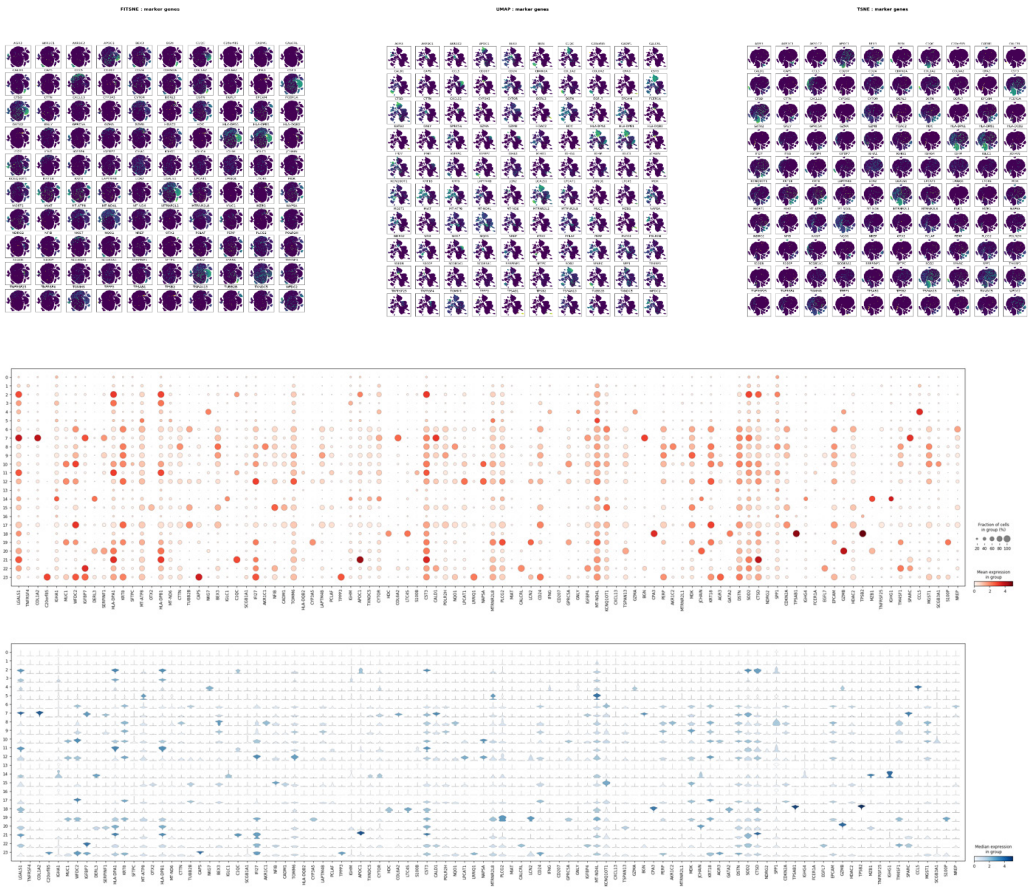
  [input 3] : groupby – marker gene을 구할 때 사용할 clustering 알고리즘의 종류를 선택

  [output] : output – 시각화 파일을 저장할 경로

  [option] : gene_name – 이 유전자 이름을 사용하여 바이올린 플롯을 보여주며 None인 경우 첫 번째 마커 유전자를 표시
- **파이프라인 연결**



- **결과 파일**



- **결과 파일 활용**

  이 결과는 단일 세포 데이터 세트에서 전처리, 차원축소, 군집화가 완료된 데이터에서 관심 있는 유전자를 입력하면 해당 유전자 또는 생물학적 신호를 구성하는 유전자 세트가 어느 군집에서 얼마나 발현하고 있는지를 결과로 제공합니다.

BIO-EXPRESS 2.0
USER MANUAL
바이오 익스프레스 2.0 사용자 매뉴얼

kobic 국가생명연구자원정보센터 Korea Bioinformation Center

34141 대전광역시 유성구 과학로 125
E-mail. cloud_team@kobic.kr www.kobic.re.kr