

BIO-EXPRESS 2.0

USER MANUAL

바이오 익스프레스 2.0 사용자 매뉴얼

BIO-EXPRESS 워크벤치





BIO-EXPRESS 2.0 USER MANUAL

바이오 익스프레스 2.0 사용자 매뉴얼

1. Bio-Express 워크벤치 설치

04 1.1 Bio-Express 워크벤치 다운로드

05 1.2 Bio-Express 워크벤치 실행

2. Bio-Express 워크벤치 상세 기능 소개

08 2.1 화면 및 기능 소개

09 2.2 사용자 정의 기반 프로그램 등록

20 2.3 분석 파이프라인 개발

29 2.4 분석 파이프라인 실행

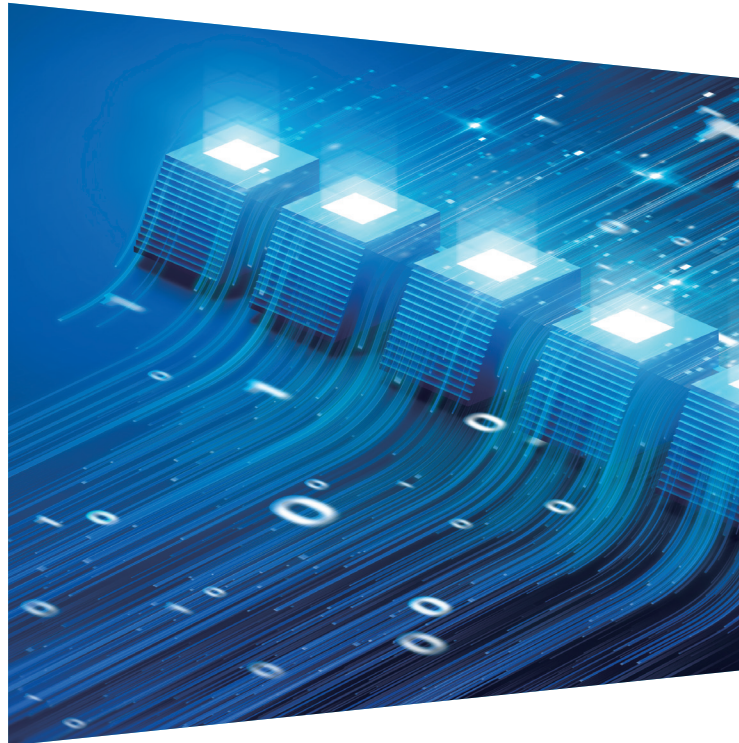
34 2.5 공개 서비스 등록



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.



EX 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

Bio-Express 워크벤치 다운로드

- Bio-Express 웹 서비스 메인 화면 우측의 Quick Downloads 버튼을 클릭한다.

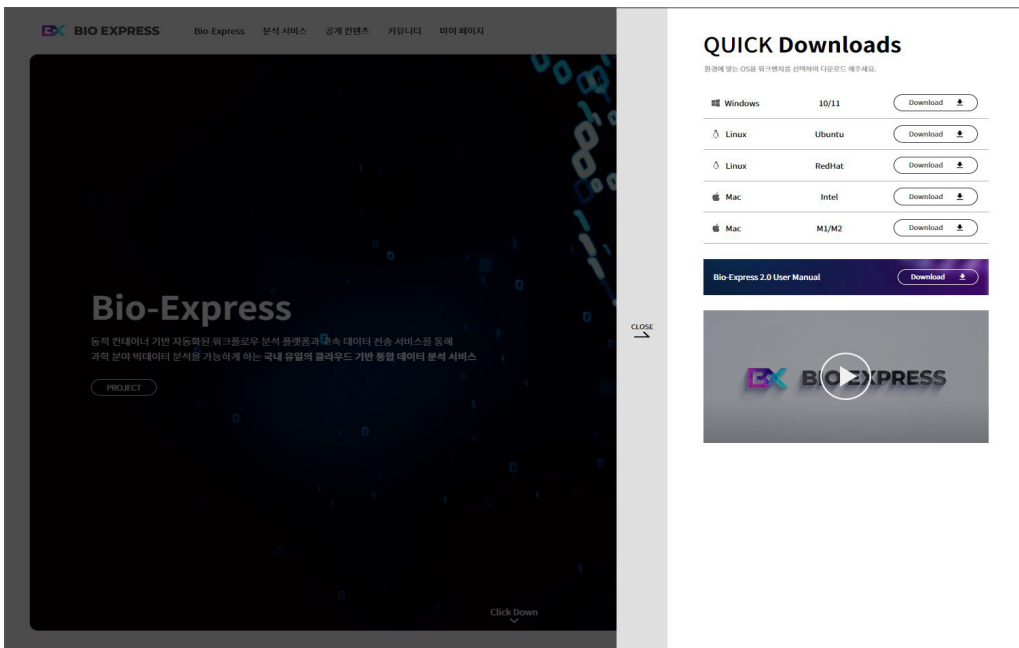
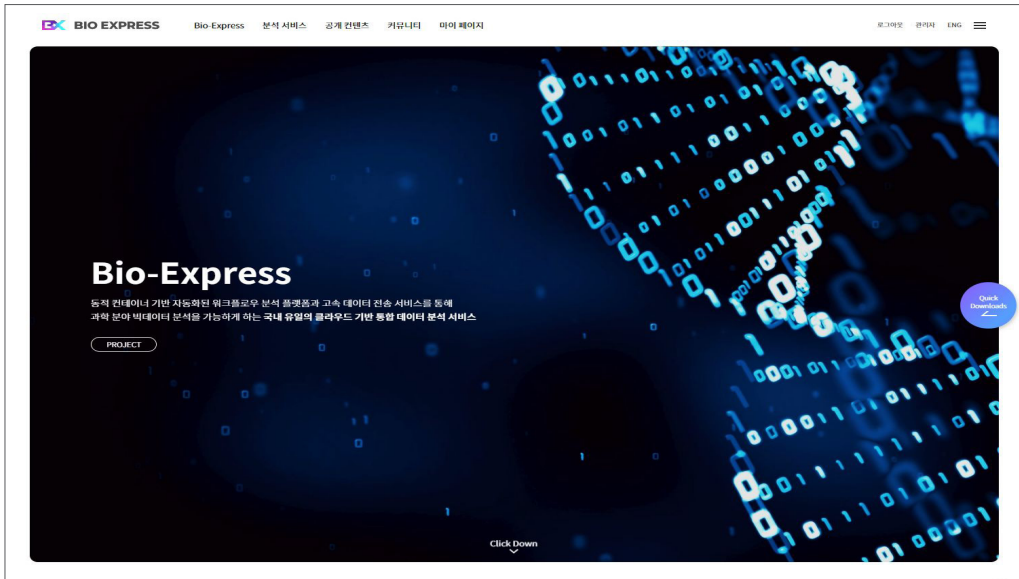


그림 1.1-1. 바이오익스프레스 워크벤치 다운로드

- 사용자 OS 및 CPU 환경에 맞는 배포버전을 다운로드하여 압축을 해제한다.

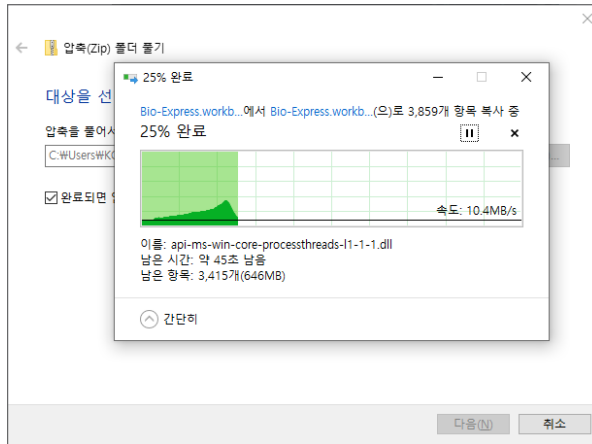


그림 1.1-2. 바이오익스프레스 압축해제

1.2

Bio-Express 워크벤치 실행

- 압축이 해제된 폴더의 Bioexpress.exe 파일을 실행한다.

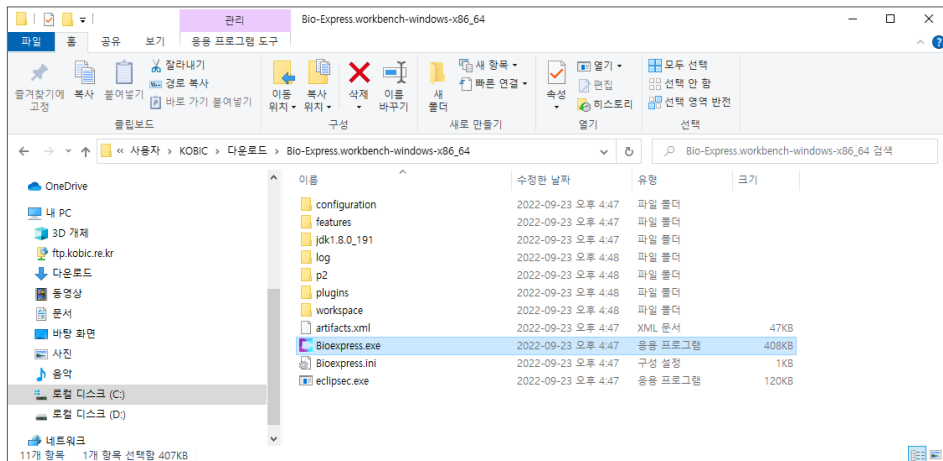


그림 1.2-1 프로그램 실행

1.2.1. Bio-Express 워크벤치 로그인

- KOBIC 통합로그인(SSO)에 가입 되어 있는 아이디로 로그인 한다.
(Bio-Express 웹 매뉴얼 목차 1.2.1 참조)
- 아이디가 없으면 왼쪽 하단 Register 버튼을 통하여 회원가입 후 로그인한다.

Single sign-on for **BIO EXPRESS**

CLOSHA

ID

Password

Starts the Bio-Express Workbench.

LOGIN CANCEL

Find ID/Password | Register Go Bio-Express Service

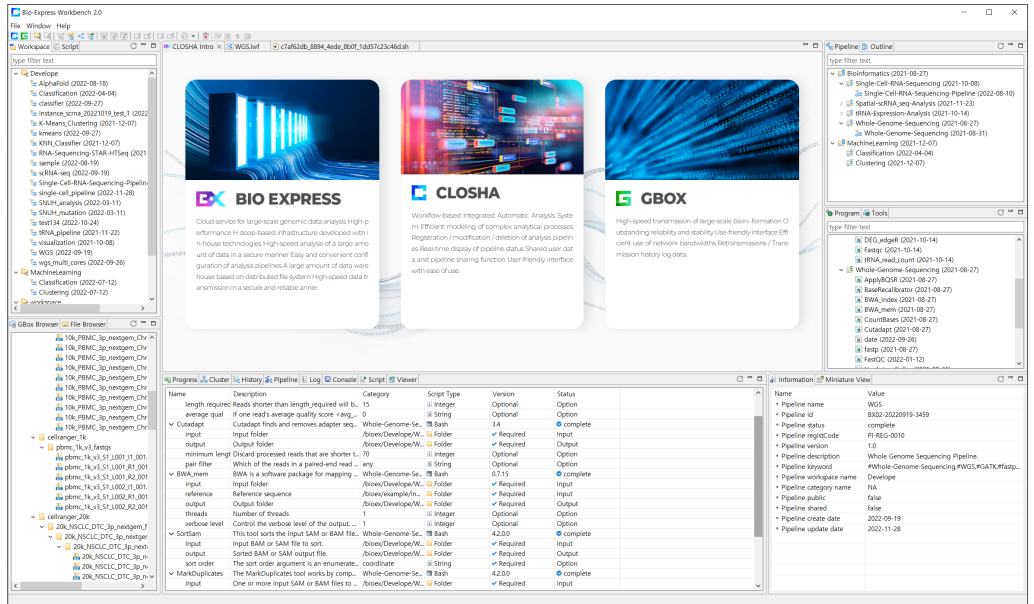
1.2.1-1. Bio-Express 로그인 화면

1.2.2. Bio-Express 워크벤치 시작화면

- CLOSHA intro 페이지가 활성화되고 각 서비스의 카드영역 부분을 누르면 Bio-Express 웹 페이지로 연결되어 자세한 소개와 사용법을 익힐 수 있다.
- 메인 메뉴 [File], [Window], [Help]는 사용자 편의기능과 워크벤치 업데이트가 가능하다.

- ✓ [File] > [New] : 새로운 워크스페이스, 카테고리, 파이프라인, 프로그램 생성
- ✓ [File] > [Register] : 파이프라인 등록, 프로그램 등록
- ✓ [Window] > [Open Perspective] : CLOSHA / GBox 퍼스펙티브 이동
- ✓ [Window] > [Reset Perspective] : 기본(Default)으로 설정된 퍼스펙티브로 되돌리기
- ✓ [Window] > [Preference] : Bio-Express 워크벤치 설정 값 세팅
- ✓ [Help] > [Help Contents] : Bio-Express 사용자 매뉴얼
- ✓ [Help] > [Community] : Bio-Express 웹 커뮤니티 페이지 이동
- ✓ [Help] > [Feedback] : Bio-Express 웹 피드백 페이지 이동
- ✓ [Help] > [About Bio-Express] : Bio-Express 웹 소개 페이지 이동
- ✓ [Help] > [Update CLOSHA Workbench] : Bio-Express 워크벤치 업데이트

- 툴바 가장 좌측 메뉴를 통하여 빅데이터 고속 전송 시스템 GBox과 클라우드 기반 통합 분석 시스템 CLOSHA의 퍼스펙티브 이동이 가능하다.



1.2.2-1. Bio-Express 워크벤치 메인 화면

2.1

화면 및 기능 소개

Bio-Express 워크벤치는 대용량 데이터 분석 서비스를 제공하는 클라우드 기반의 프로그램이다. 사용자 친화적인 인터페이스로 구성되어 복잡한 생명정보 분석과정을 손쉽게 파이프라인으로 디자인 할 수 있다.

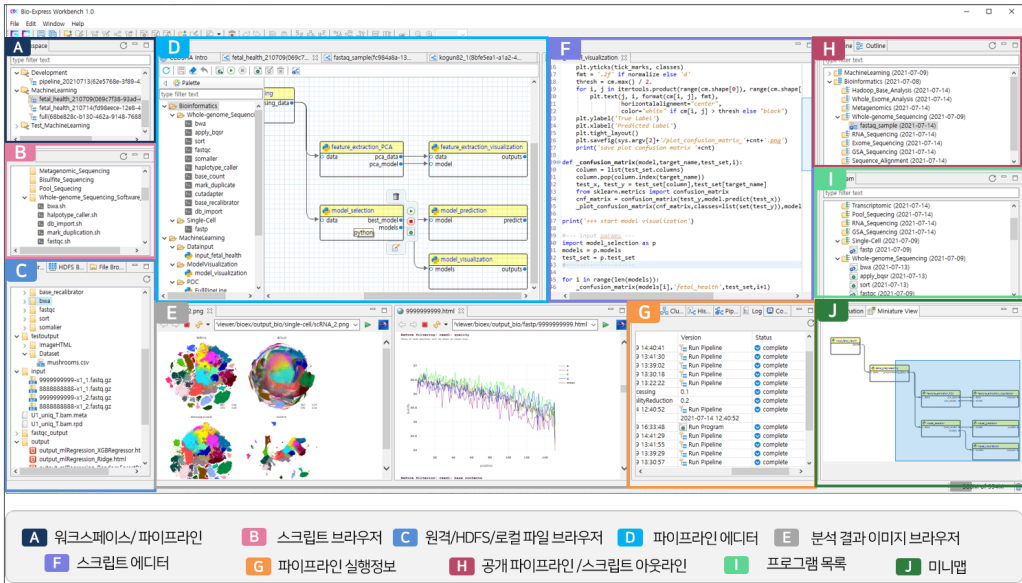


그림 2.1-1 바이오 익스프레스 CLOSHA 화면

2.2

사용자 정의 기반 프로그램 등록

2.2.1. 사용자 분석 코드 작성

- 사용자 정의 기반 프로그램을 등록하기 위해 먼저 사용자 분석 코드를 등록해야한다.
- 사용자 분석 코드를 등록할 스크립트 폴더를 생성하기 위하여 [Script] View 컨텍스트 메뉴의 [New Folder] 버튼을 클릭한다.

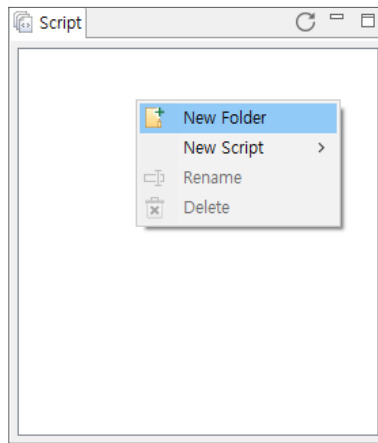


그림 2.2.1-1 사용자 스크립트 폴더 생성

- 생성 폴더의 이름을 작성한 후 [OK] 버튼을 클릭하여 폴더를 생성한다.

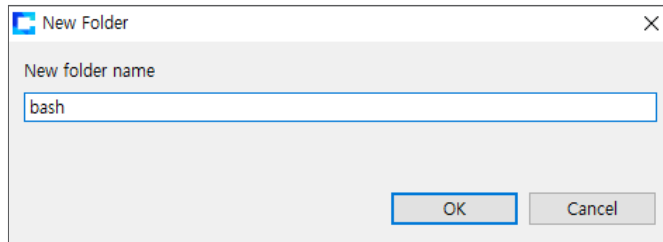


그림 2.2.1-2 사용자 스크립트 폴더 생성 다이얼로그

- [그림2.2.1-3]과 같이 생성된 스크립트 폴더에서 컨텍스트 메뉴의 [New Script] 버튼을 클릭하여 사용자 분석 코드를 생성한다. 사용자 분석 코드의 파일 형식으로는 Python, Bash, R을 지원한다.

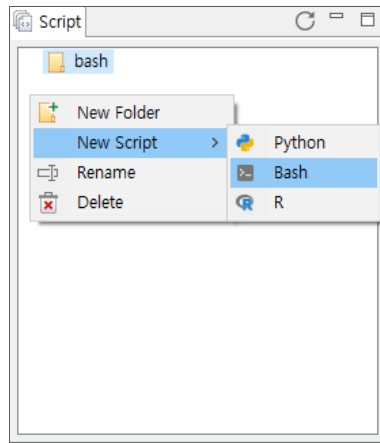


그림 2.2.1-3 사용자 스크립트 생성

- 생성할 스크립트의 이름을 입력한 후 [OK] 버튼을 클릭하여 빈 스크립트를 생성한다.

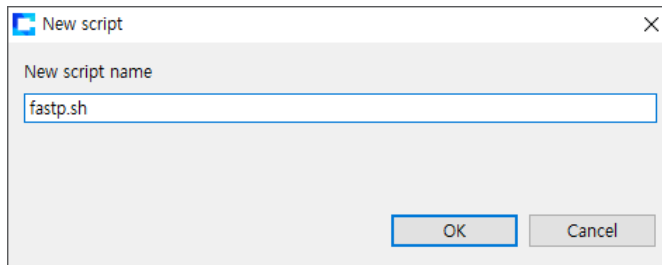


그림2.2.1-4 사용자 스크립트 생성 다이얼로그

- 생성한 사용자 분석 코드는 [그림2.2.1-5]와 같이 [Script] View에서 더블 클릭하여 에디터를 통해 작성 및 편집이 가능하며, 에디터에서 코드 자동 완성 기능을 제공하여 보다 편리한 코드 작성을 지원한다.

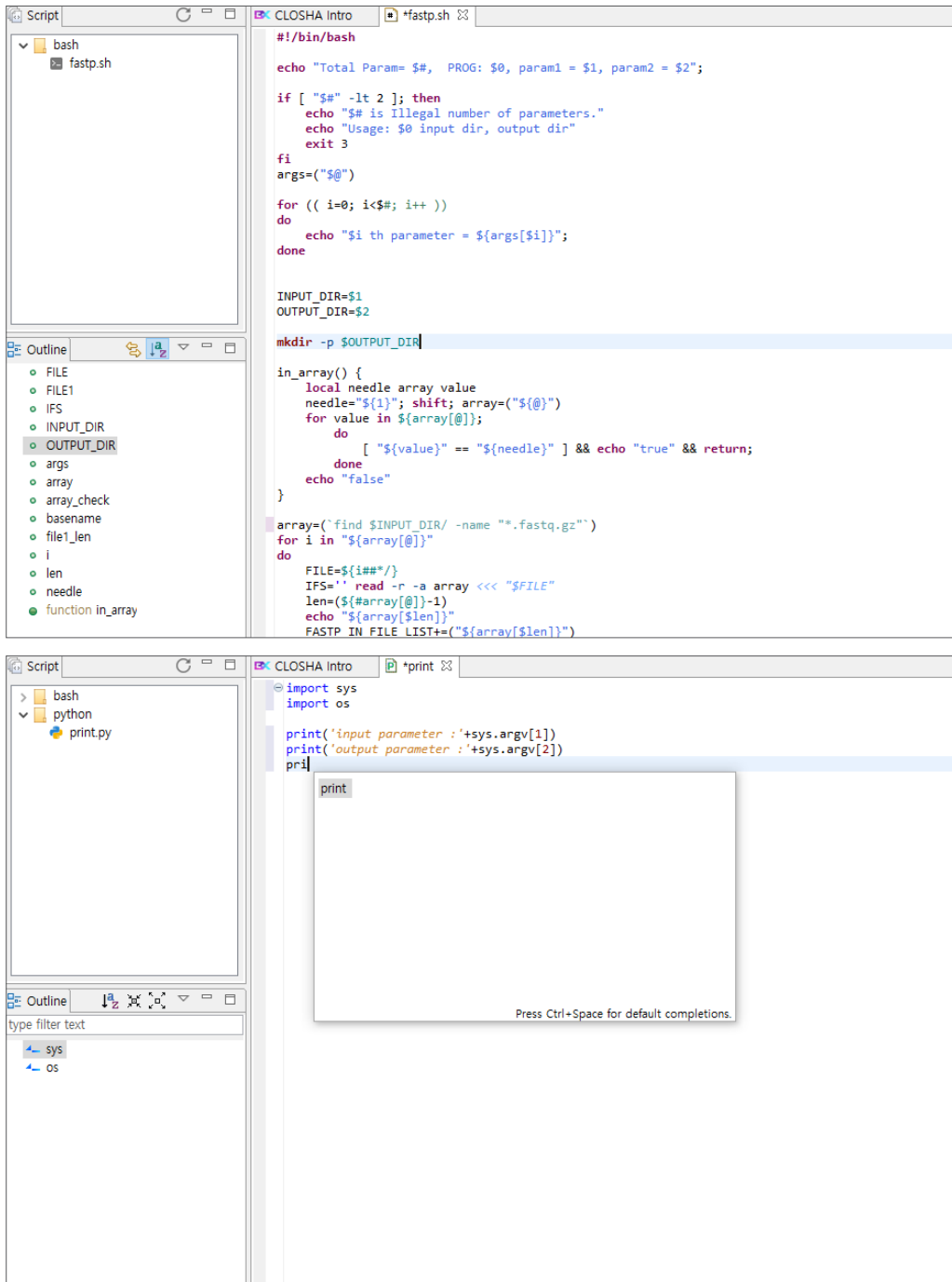


그림 2.2.1- 5 사용자 스크립트 작성

2.2.2. 사용자 프로그램 등록

- 작성한 사용자 분석 코드를 실행하기 위해서 사용자 프로그램을 등록한다.
- [Program] View에서 컨텍스트 메뉴의 [New Category] 버튼을 클릭하여 사용자 정의 프로그램을 등록할 루트 카테고리 및 서브 카테고리를 생성한다.

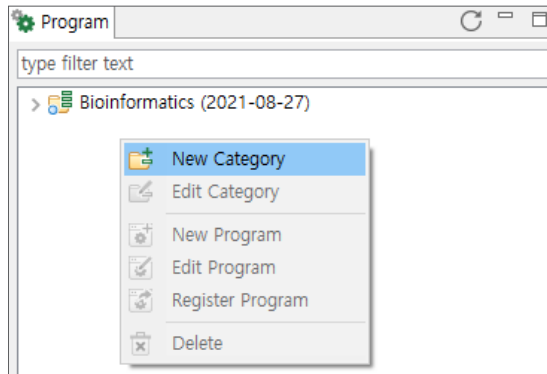


그림 2.2.2-1 사용자 프로그램 루트 카테고리 생성

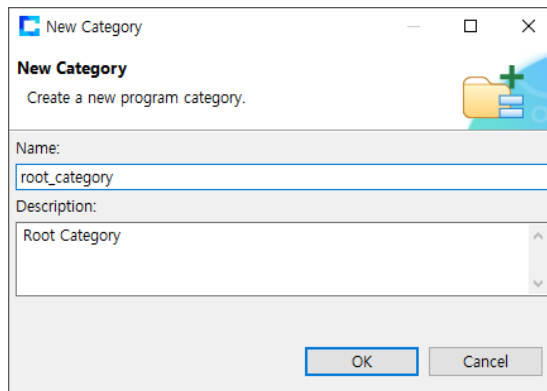


그림 2.2.2-2 사용자 프로그램 루트 카테고리 생성 다이얼로그

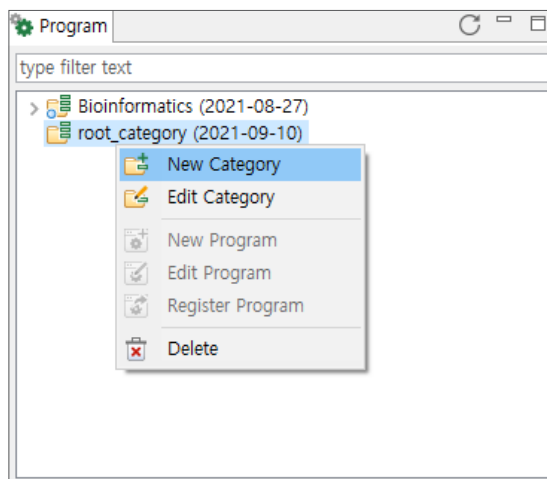


그림 2.2.2-3 사용자 프로그램 서브 카테고리 생성

- 컨텍스트 메뉴로 루트 카테고리 하위에 서브 카테고리를 생성한다.

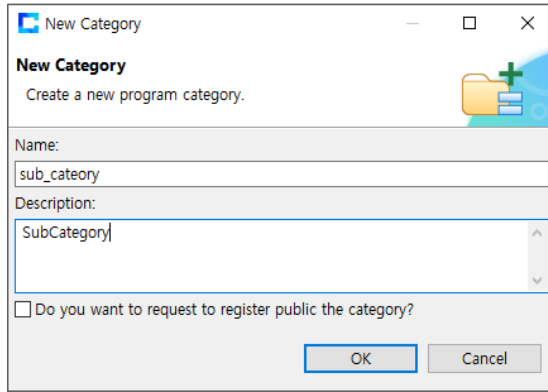


그림 2.2.2 - 4 사용자 프로그램 서브 카테고리 생성 다이얼로그

- 생성한 서브 카테고리의 컨텍스트 메뉴 [New Program] 버튼을 클릭하여 사용자 프로그램을 등록한다.

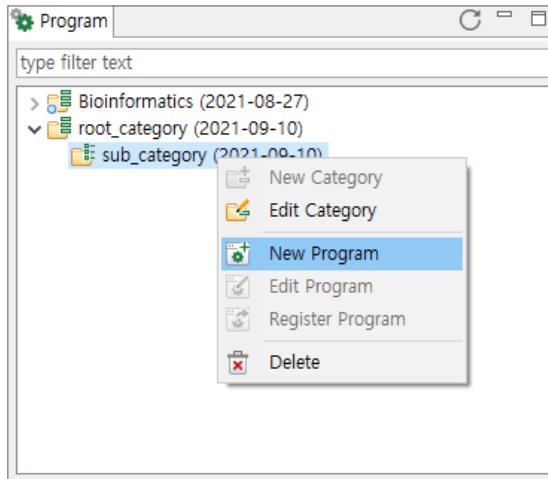


그림 2.2.2 - 5 사용자 프로그램 생성

- 등록할 사용자 프로그램의 기본정보, 추가정보, 인자값을 설정하여 프로그램을 등록한다.

New Program
Create New Analysis Program
Please enter a description

Program Name: fastp

Category Name: sub_category Select...

ID: BX04-20220913-0500

Basic Information

Version: 0.20.1

Keyword: fastp

URL: http://github.com/OpenGene/fastp

Description
Atool designed to provide fast all-in-one preprocessing for FastQ files

< Back **Next >** Finish Cancel

그림 2.2.2-6 사용자 프로그램 생성 다이얼로그 - 프로그램 기본정보

New Program
Create New Analysis Program
Setting Program Execution Information

Script Information

Script Type: Bash

Script Path: /BIO/K-BDS/BIO-EXPRESS/script/bioex/bash/fastp.sh Select...

Execute Core

Run analysis tools using a single core
 Run analysis tools with up to 6 multi-cores

Platform Environment

Use the high performance computing cluster
 Use a GPU(Graphics Processing Unit) server (supported for machine learning and deep learning)

Relation Program

Name	Description	Category
> <input type="checkbox"/> train_test_split	Split data into train, test set	Preprocessing
> <input type="checkbox"/> tRNA_read_count	Counting of mapped tRNA reads	tRNA-Expression
> <input type="checkbox"/> somalier	rapid relatedness estimation for cancer and germline...	Whole-Genome-
> <input type="checkbox"/> model_visualization	Model Visualization	Visualization
> <input type="checkbox"/> maftools_summary	.	Annovar
> <input type="checkbox"/> maftools_pathway_d	.	Annovar
> <input type="checkbox"/> maftools_oncplot	.	Annovar
> <input type="checkbox"/> maftools AAchange	.	Annovar

< Back **Next >** Finish Cancel

그림 2.2.2-7 사용자 프로그램 생성 다이얼로그 - 프로그램 스크립트 및 실행 설정

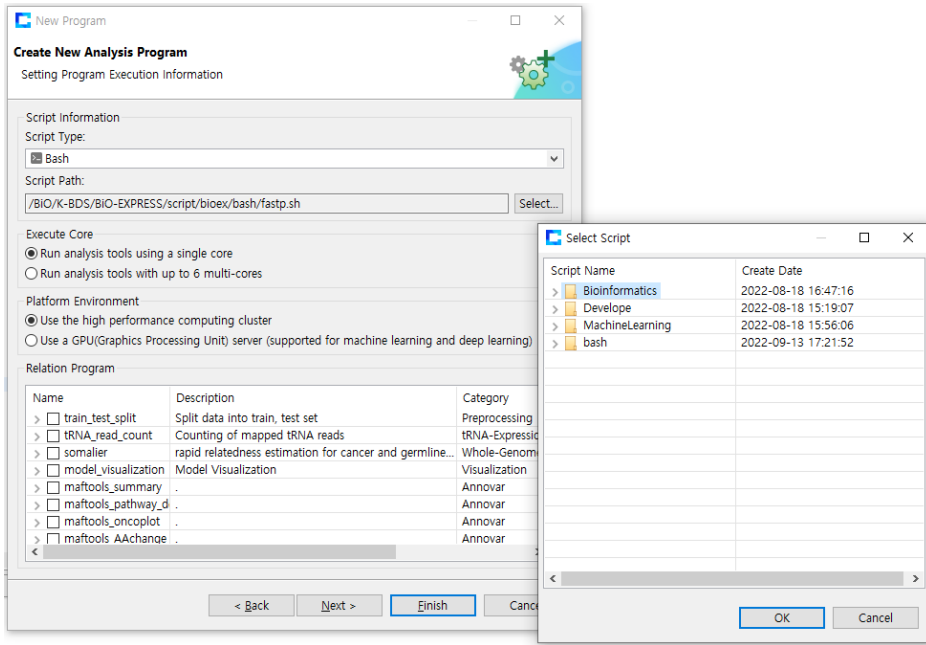


그림 2.2.2 - 8 사용자 프로그램 생성 다이얼로그 - 프로그램 스크립트 파일 탐색기

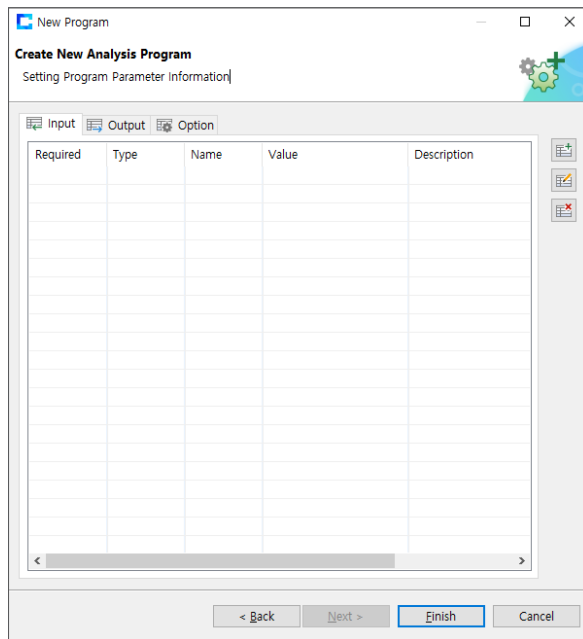


그림 2.2.2 - 9 사용자 프로그램 생성 다이얼로그 - 프로그램 Input 파라미터 설정

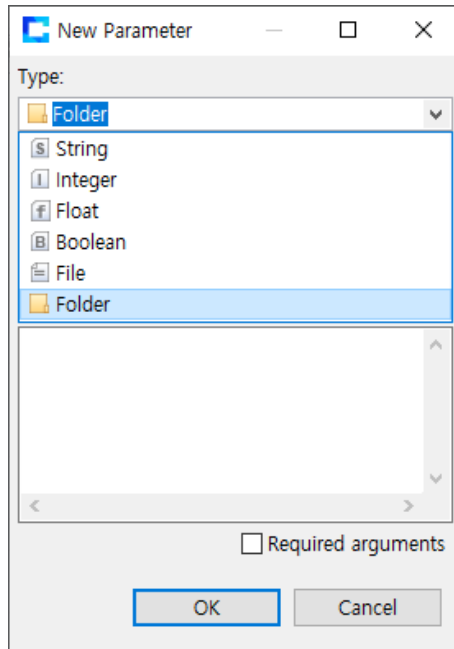


그림 2.2.2- 10 사용자 프로그램 생성 다이얼로그 - 프로그램 Input 파라미터 타입 설정

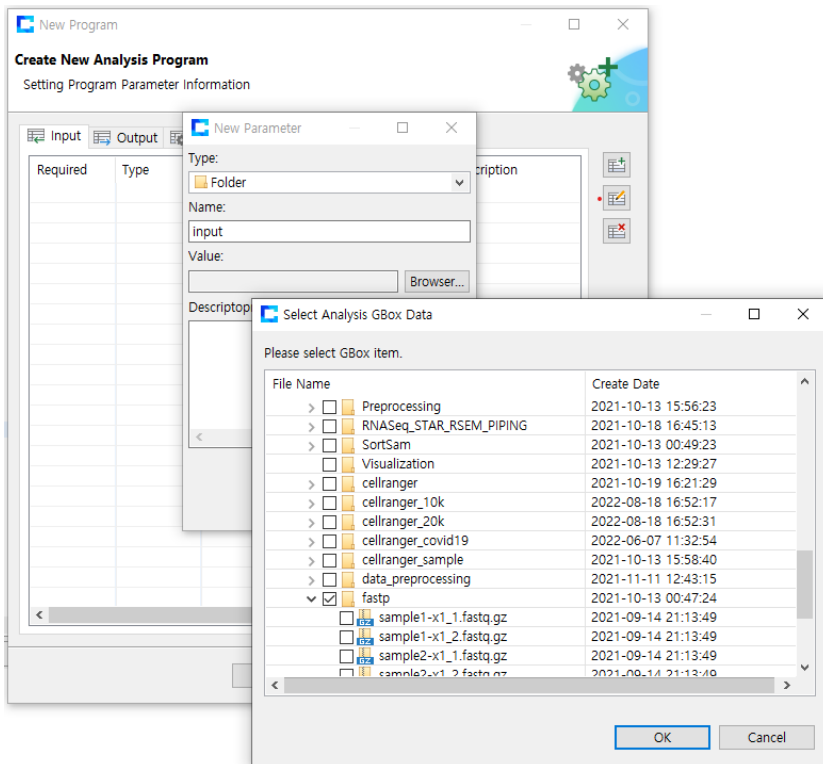


그림 2.2.2- 11 사용자 프로그램 생성 다이얼로그 - 프로그램 Input 파라미터 입력 폴더 설정

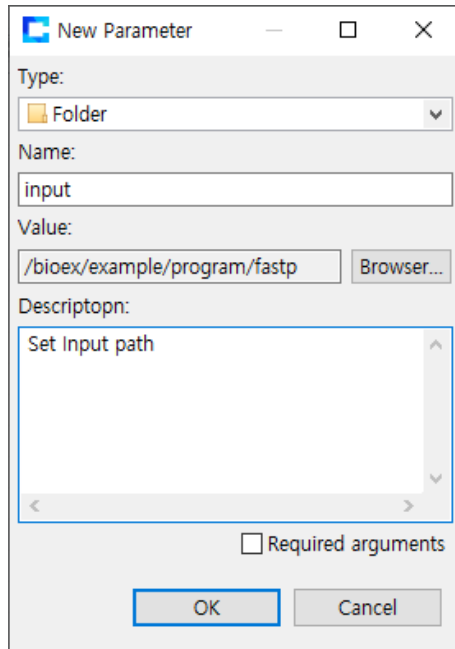


그림 2.2.2 - 12 사용자 프로그램 생성 다이얼로그 - 프로그램 Input 파라미터 추가 다이얼로그

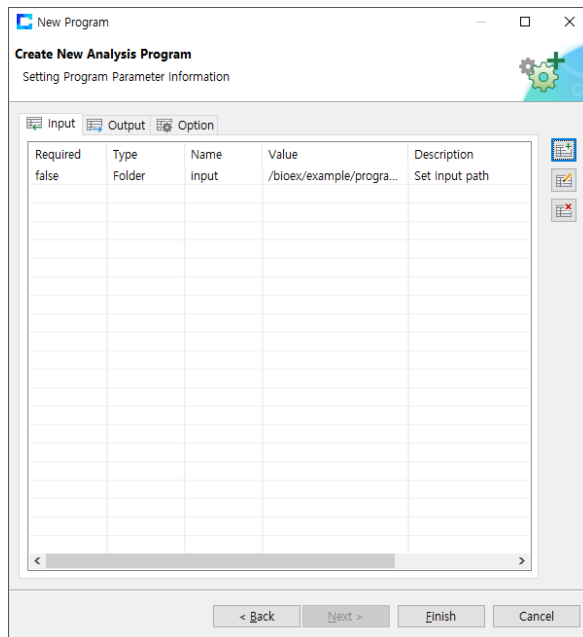


그림 2.2.2 - 13 사용자 프로그램 생성 다이얼로그 - 프로그램 Input 파라미터 추가 완료 화면

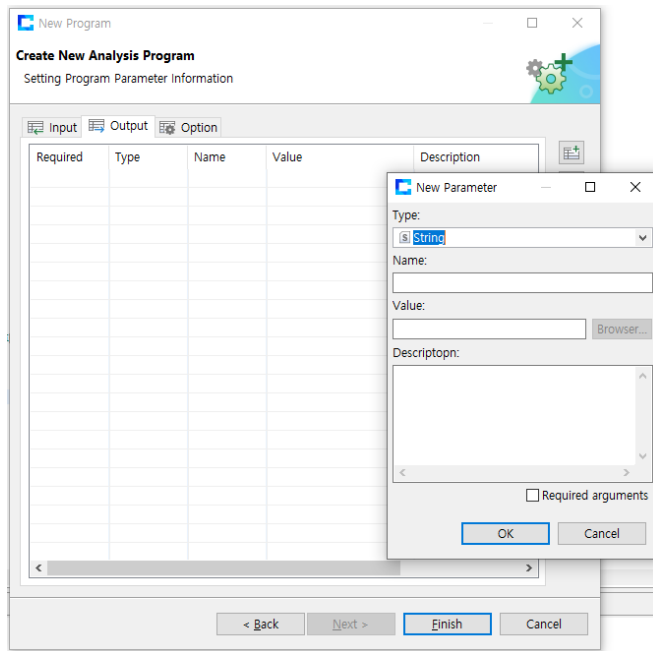


그림 2.2.2 - 14 사용자 프로그램 생성 다이얼로그 - 프로그램 Output 파라미터 설정

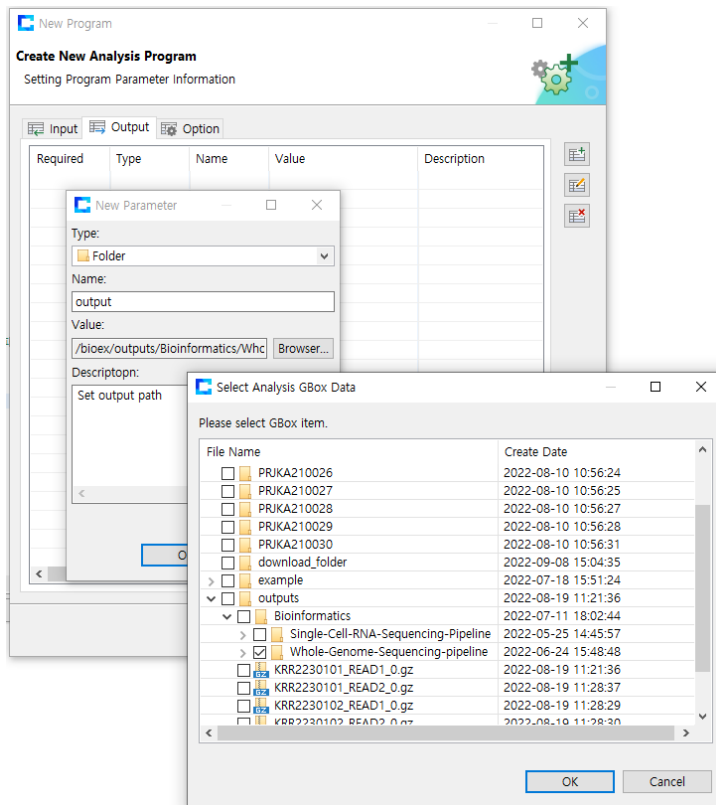


그림 2.2.2 - 15 사용자 프로그램 생성 다이얼로그 - 프로그램 Output 파라미터 출력 폴더 설정

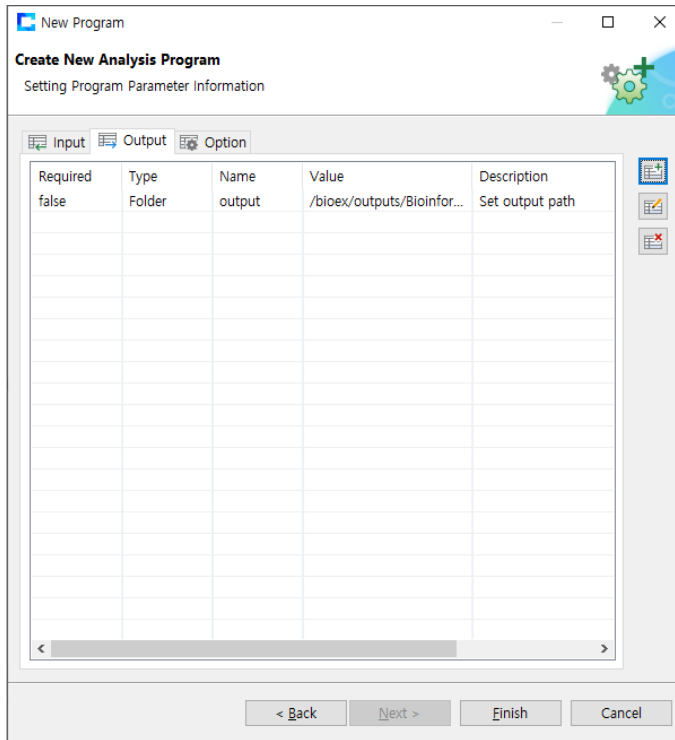


그림 2.2.2 - 16 사용자 프로그램 생성 다이얼로그 - 프로그램 Output 파라미터 추가 완료 화면

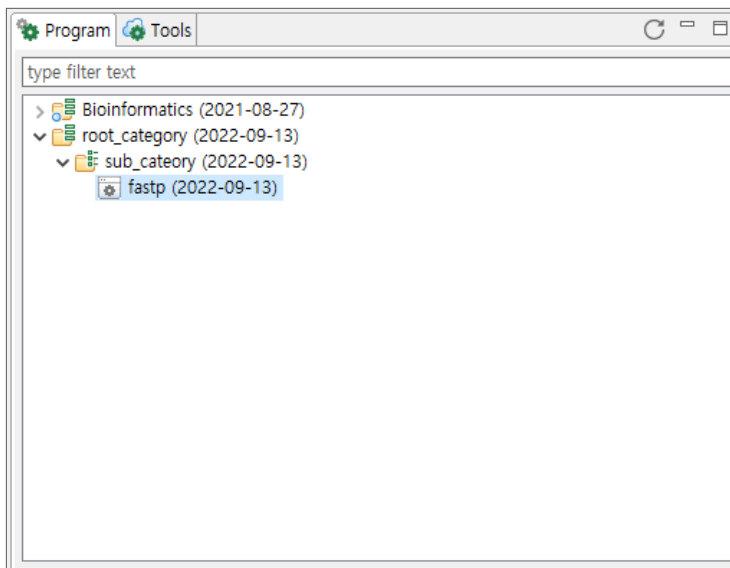


그림 2.2.2 - 17 사용자 프로그램 생성 완료 화면

2.3

분석 파이프라인 개발

2.3.1. 워크스페이스 생성

- [그림 2.3.1-1]과 같이 [Workspace] View의 컨텍스트 메뉴 [New Workspace] 버튼을 클릭하여 워크스페이스를 생성한다.

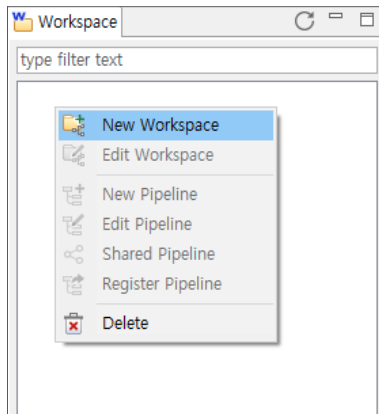


그림 2.3.1-1 사용자 워크스페이스 생성

- 생성할 워크스페이스의 이름과 설명을 작성하고 연관 키워드를 등록한 후 [OK] 버튼을 클릭한다.

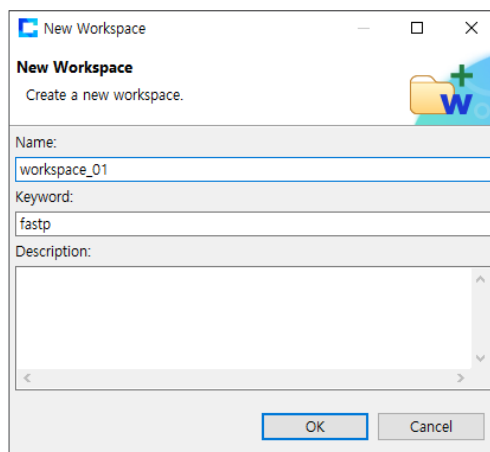


그림 2.3.1-2 사용자 워크스페이스 생성 다이얼로그

- 생성한 워크스페이스의 컨텍스트 메뉴 [New Pipeline] 버튼을 클릭하여 분석 파이프라인을 생성한다.

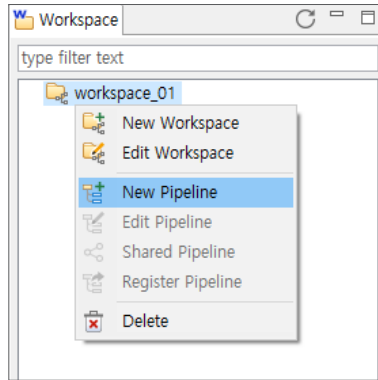


그림 2.3.1- 3 파이프라인 생성

- 파이프라인 유형에 따라 사용자 정의 파이프라인을 생성하거나, 기존에 등록되어 있던 파이프라인을 선택하여 인스턴스 파이프라인을 생성할 수도 있다.

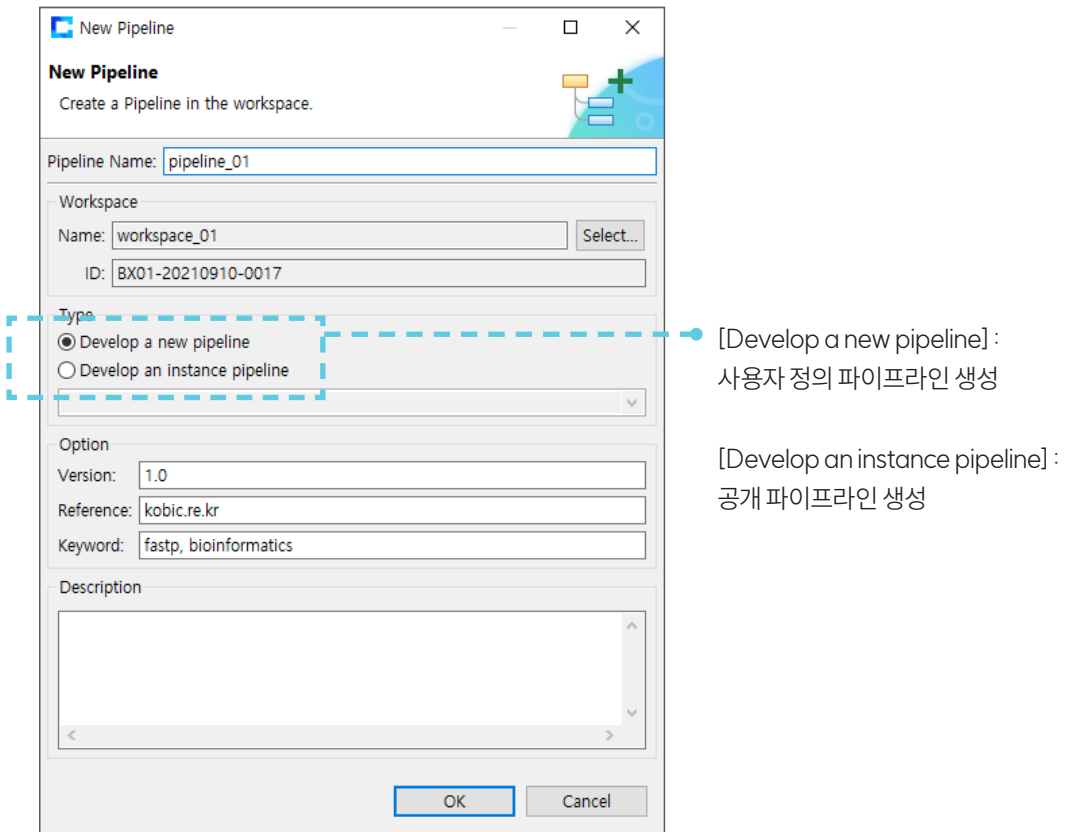


그림 2.3.1- 4 사용자 정의 파이프라인 생성

- [그림 2.3.1-5]과 같이 [Workspace] View를 통해 생성된 사용자 정의 파이프라인을 확인할 수 있다.

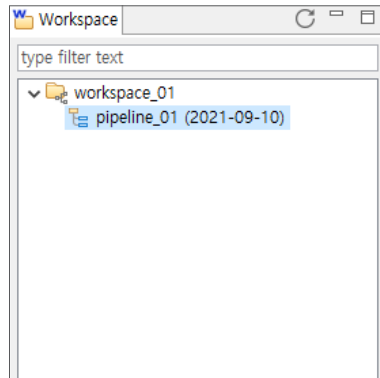


그림 2.3.1 - 5 사용자 정의 파이프라인 생성 완료 화면

2.3.2. 분석 파이프라인 디자인

● 사용자 정의 파이프라인 디자인

- 분석 파이프라인 디자인에 필요한 분석 프로그램을 에디터의 Palette에서 찾는다.

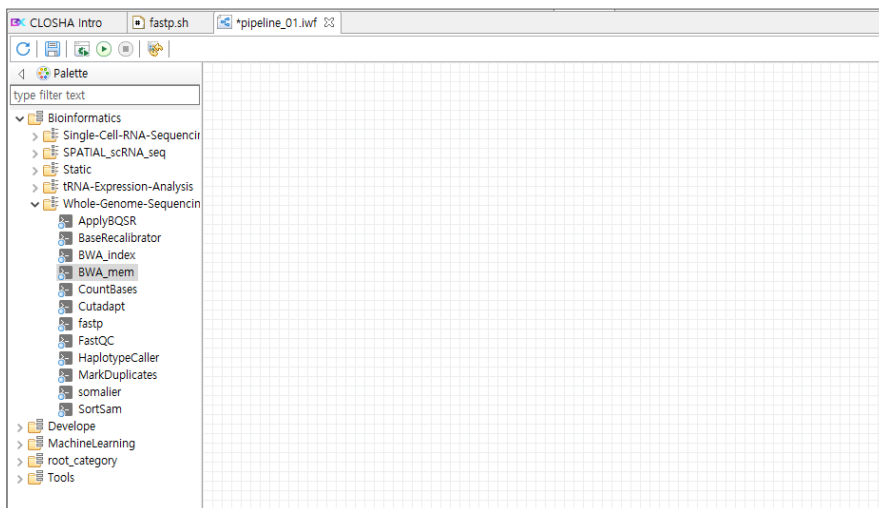


그림 2.3.2 - 1-1 생성된 파이프라인 에디터 활성화 화면

- 프로그램을 선택한 후 캔버스 위에 마우스로 드래그 앤 드롭 한다.

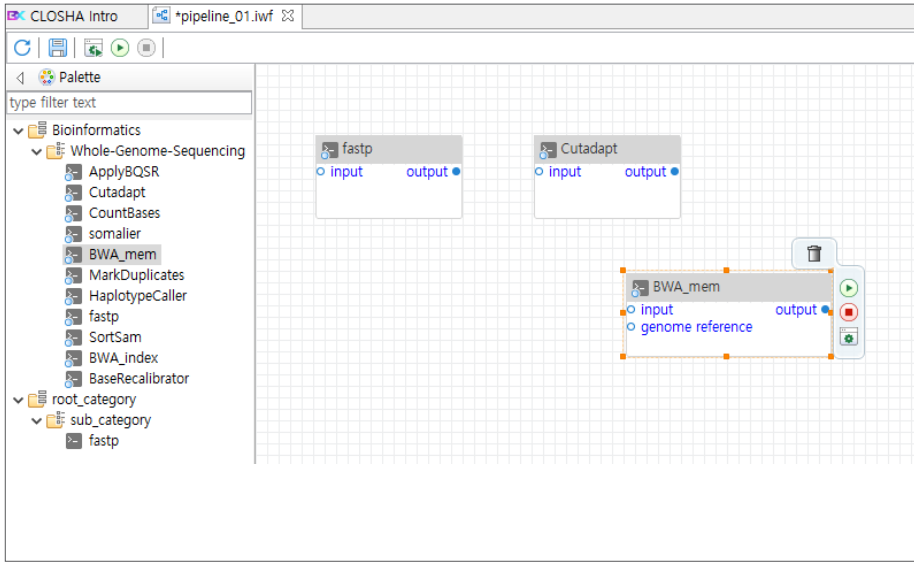


그림 2.3.2-1-2 Drag & Drop 을 이용한 파이프라인 디자인

- 필요한 분석 프로그램으로 디자인한 파이프라인을 저장하고, 각각의 분석 프로그램을 더블 클릭하여 실행에 필요한 옵션을 설정한다.

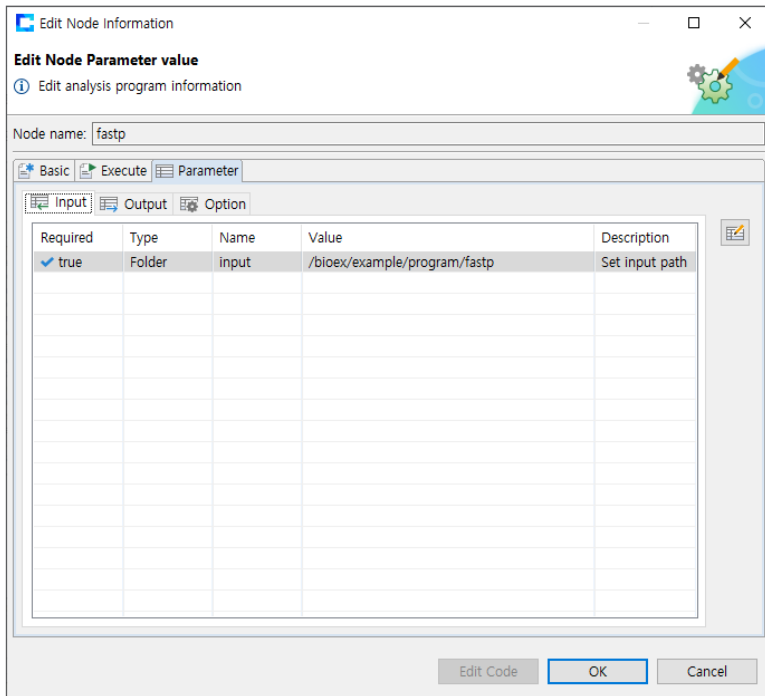


그림 2.3.2-1-3 활성화된 프로그램 파라미터 설정 다이얼로그

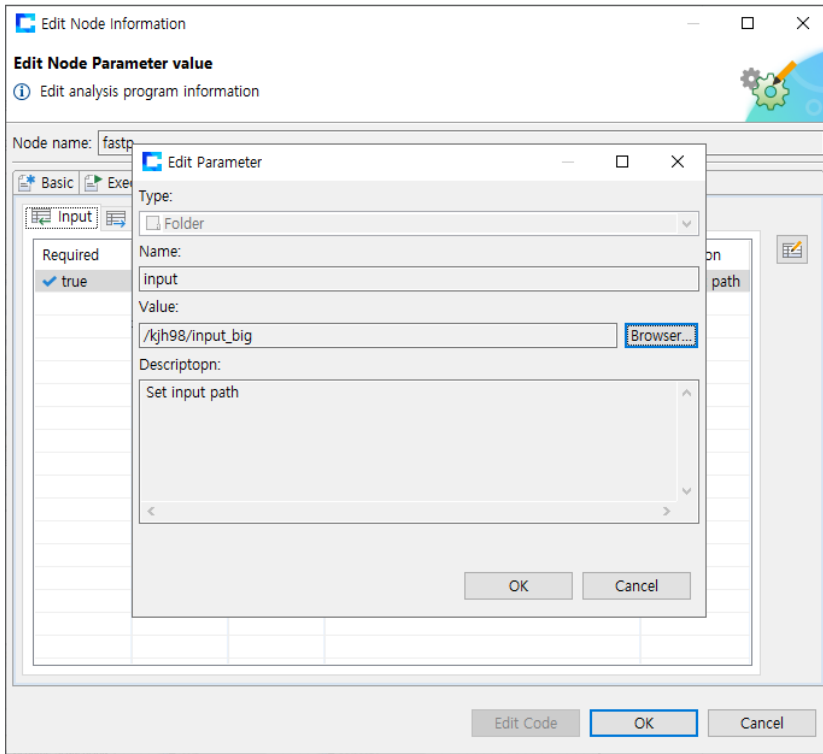


그림 2.3.2- 1 - 4 Input 파라미터 설정 다이얼로그

- 분석 파이프라인 실행 옵션 설정 후, 캔버스에 위치한 분석 프로그램을 연결하여 하나의 파이프라인을 생성한다.
- 분석 프로그램간의 연결과 옵션 설정의 순서가 바뀌어도 기능상 문제가 없지만, 분석 프로그램을 연결하면 자동으로 앞 단계 프로그램의 Output을 Input으로 설정하게 되므로 프로그램 설정 후 연결하는 방식을 추천한다.

Name	Description	Category	Script Type	Version	Status
▼ fastp	A tool designed to provide fast all-in-one preproces...	Whole-Genome-Sequencing	Bash	0.20.1	wait
input	Set input path	/kjh98/input_big	Folder	Required	Input
output	Set output path	/kjh98/workspace_01/pipeline_01/fastp/output	Folder	Required	Output
▼ Cutadapt	Cutadapt finds and removes adapter sequences, pri...	Whole-Genome-Sequencing	Bash	3.4	wait
input	Set input path	/kjh98/workspace_01/pipeline_01/fastp/output	Folder	Required	Input
output	Set output path	/kjh98/workspace_01/pipeline_01/Cutadapt/output	Folder	Required	Output
▼ BWA_mem	BWA is a software package for mapping low-diverge...	Whole-Genome-Sequencing	Bash	0.7.15	wait
input	Set input path	/bioex/example/program/BWA_mem	Folder	Required	Input
genome reference	Set genome reference fasta folder path	/bioex/example/index/BWA/hg38	Folder	Required	Input
output	Set output path	/kjh98/workspace_01/pipeline_01/BWA_mem/output	Folder	Required	Output

그림 2.3.2- 1 - 5 연결선 연결

- 분석 프로그램의 Output 파라미터 우측 끝에 위치한 원 모양의 연결 영역을 다음 단계 분석 프로그램의 Input 파라미터 좌측 연결 영역으로 드래그 앤 드롭하여 사용자 정의 파이프 라인을 디자인한다.

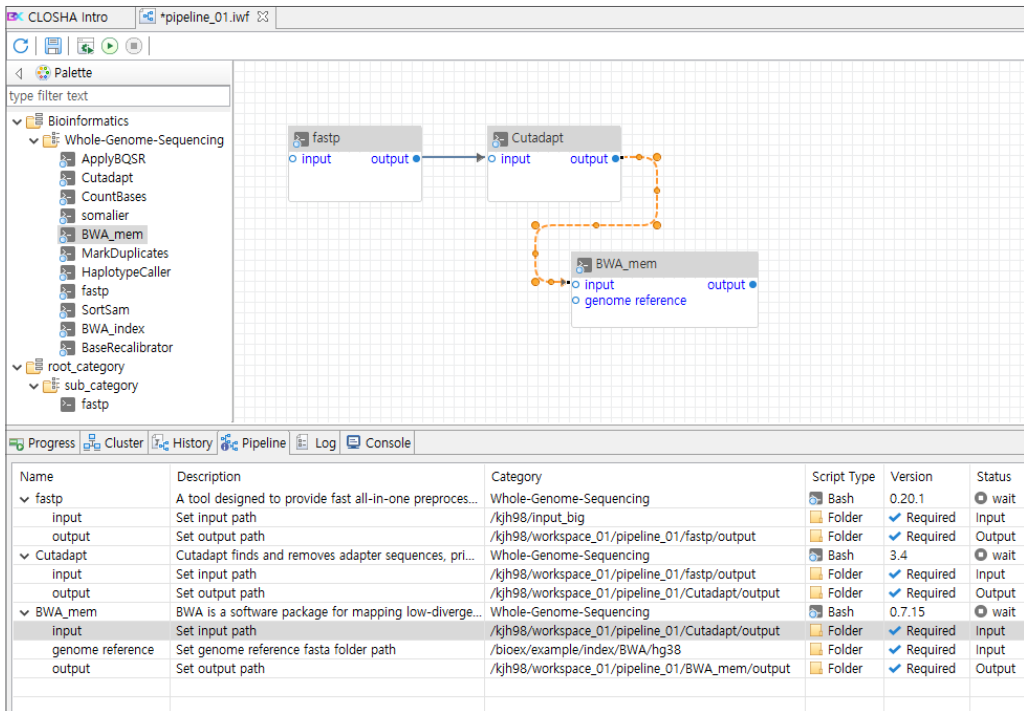


그림 2.3.2- 1- 6 연결선 디자인

- 연결선의 형태는 [그림 2.3.2-1-6]과 같이 마우스 드래그를 통하여 사용자가 원하는 대로 자유롭게 수정할 수 있다.

● 인스턴스 파이프라인 디자인

- 그림 2.3.2-2-2과 같이 파이프라인 생성 단계에서 인스턴스 파이프라인 유형을 선택하여 기존에 등록되어 있던 파이프라인을 선택할 수 있다.

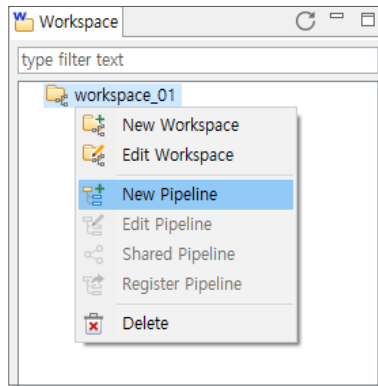


그림 2.3.2 - 2 - 1 파이프라인 생성

- 인스턴스 파이프라인 역시 사용자 정의 파이프라인과 같이 각각의 프로그램 별 실행에 필요한 설정을 변경할 수 있다.

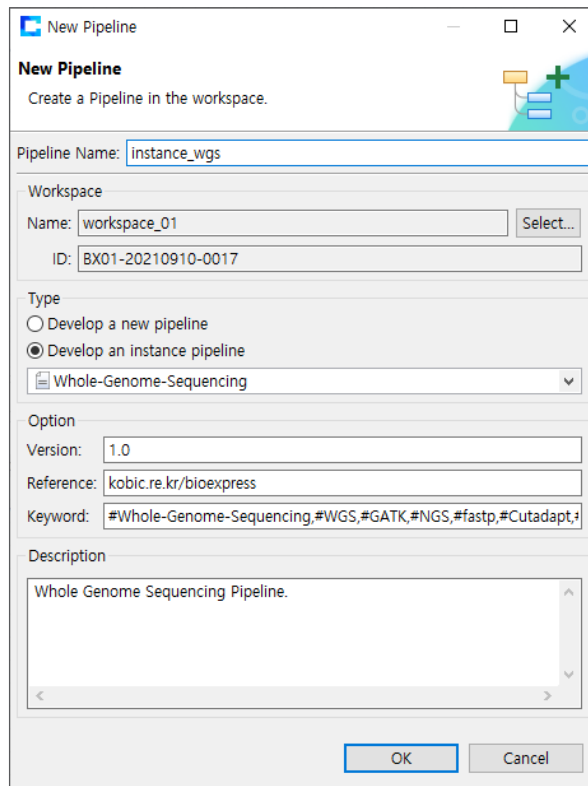


그림 2.3.2 - 2 - 2 인스턴스 파이프라인 생성

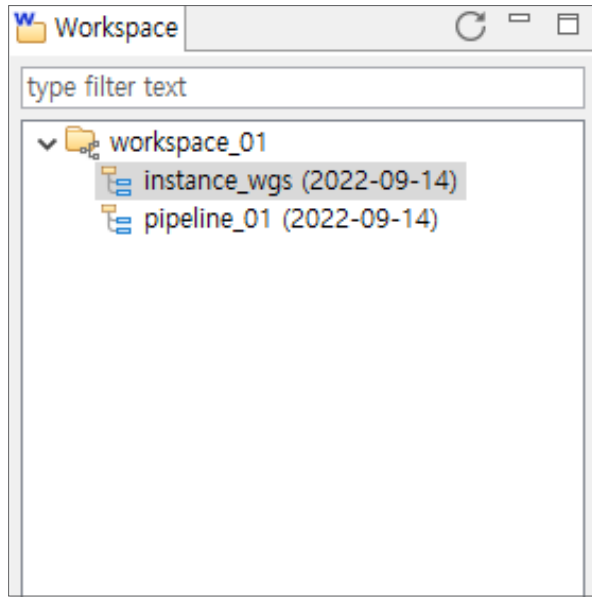


그림 2.3.2 - 2 - 3 인스턴스 파이프라인 생성 완료 화면

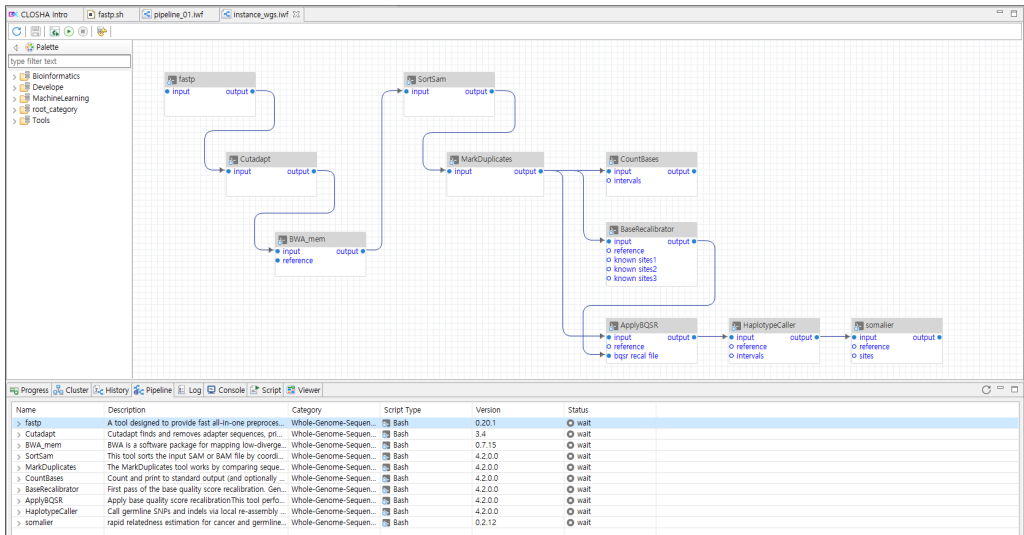


그림 2.3.2 - 2 - 4 생성된 인스턴스 파이프라인 에디터 활성화 화면

2.4

분석 파이프라인 실행

2.4.1. 분석 파이프라인 실행

- 에디터 상단의 [Run] 버튼을 클릭하여 디자인한 분석 파이프라인을 실행할 수 있다.
- 분석이 실행되면 [Pipeline] View, [Log] View와 에디터를 통해 해당 분석 파이프라인의 진행 상태를 확인할 수 있다. 각 View와 에디터의 [Refresh] 버튼을 통해 분석 파이프라인 실시간 진행 상태를 확인할 수 있다. 분석 파이프라인 실행 시 순차적으로 분석 프로그램이 진행된다.

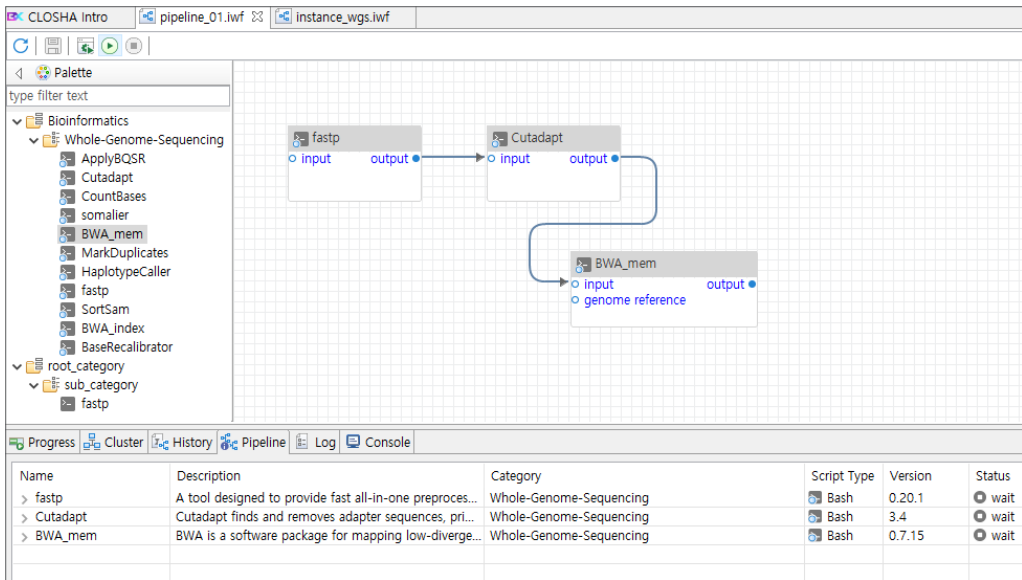


그림 2.4.1-1 파이프라인 에디터 및 파이프라인 정보 화면

- 실행 중인 프로그램은 run 상태이며 에디터에서 녹색으로 표시되고, 실행을 마친 complete 상태의 프로그램은 파란색으로 표시된다.

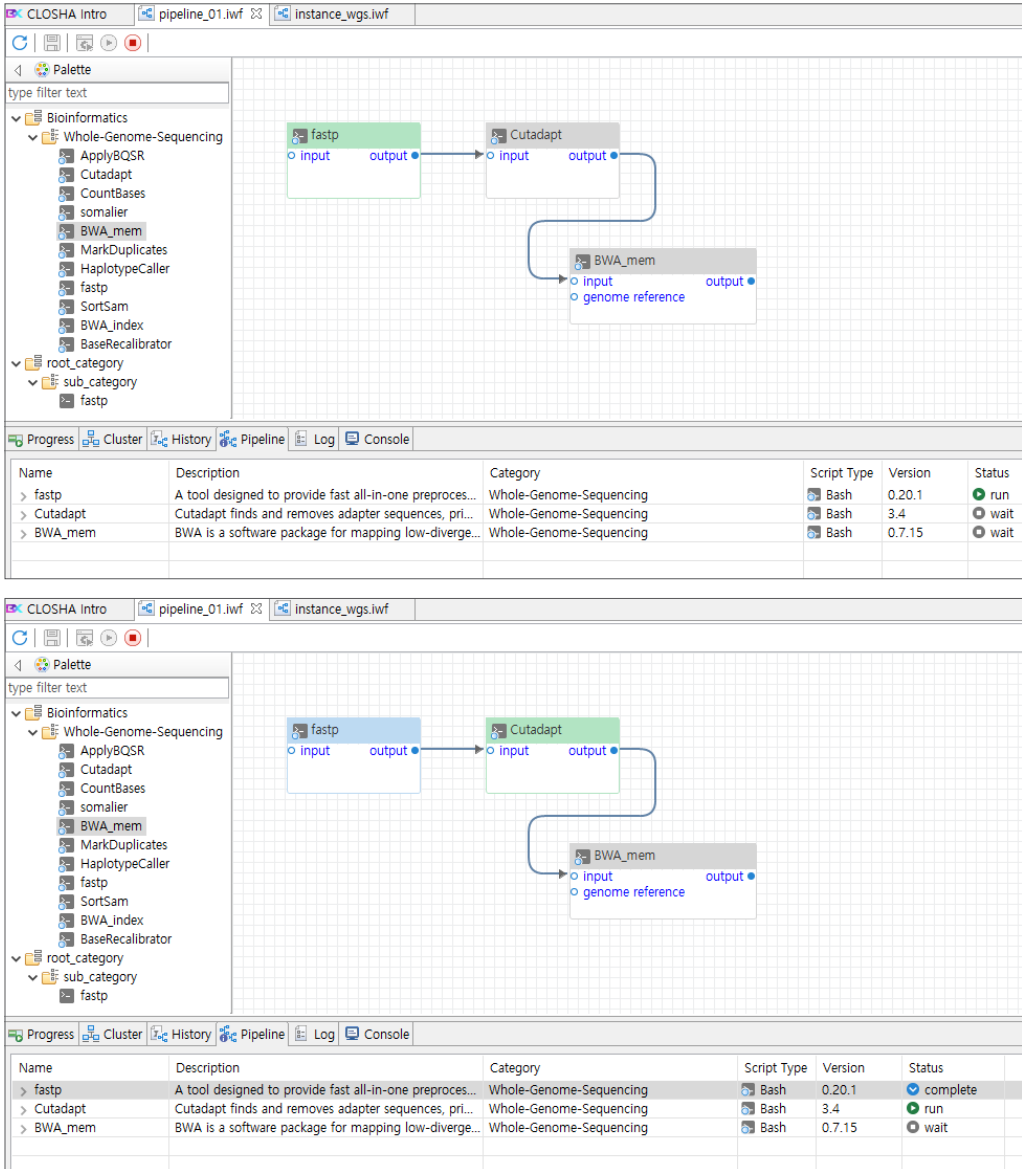


그림 2.4.1- 2 파이프라인 실행 화면

2.4.2. 단일 분석 프로그램 실행

- 각 분석 프로그램의 단일 실행도 가능하다. [캔버스에서](#) 단일 실행할 분석 프로그램에 마우스 오버 시 나타나는 바로 가기 메뉴에서 [Start Node]버튼을 클릭하고 메시지창에서 OK를 클릭하면 정상 실행 후 Pipeline view에서 실행결과를 확인 할 수 있다.

Name	Description	Category	Script Type	Version	Status
> fastp	A tool designed to provide fast all-in-one preproces...	Whole-Genome-Sequencing	Bash	0.20.1	wait
> Cutadapt	Cutadapt finds and removes adapter sequences, pri...	Whole-Genome-Sequencing	Bash	3.4	wait
> BWA_mem	BWA is a software package for mapping low-diverge...	Whole-Genome-Sequencing	Bash	0.7.15	wait

2.4.2 - 1 파이프라인 에디터 및 파이프라인 정보 화면

2.4.3. 데이터 분석 결과 확인

- [Log] View를 통해 각 분석 프로그램의 로그 정보를 확인한다.

Name	Description	Category	Version	Status
▼ fastp	A tool designed to provide fast all-in-one preprocess...	Whole-Genome-Sequen...	0.20.1	complete
▼	524685f6-374d-465e-8d89-1160d601d23e	2022-09-14 10:32:23	Run Pipeline	complete
		21033	2022-09-14 10:32:23	
▼	53647a3d-beba-4c7d-adbf-f2e552b0da82	2022-09-14 10:15:34	Run Program	complete
		20991	2022-09-14 10:15:34	
▼ Cutadapt	Cutadapt finds and removes adapter sequences, pri...	Whole-Genome-Sequen...	3.4	complete
▼	b16897c0-1d3d-47d4-8453-670ed38f1f9e	2022-09-14 10:33:23	Run Pipeline	complete
		21034	2022-09-14 10:33:23	
▼	bab1d289-93f0-48fd-850f-10b0bd45dfda	2022-09-14 10:15:46	Run Program	complete
		20992	2022-09-14 10:15:46	
▼ BWA_mem	BWA is a software package for mapping low-diverge...	Whole-Genome-Sequen...	0.7.15	complete
▼	f5c0282b-245a-4067-96f9-f52dcbdf556d	2022-09-14 10:34:23	Run Pipeline	complete
		21035	2022-09-14 10:34:23	
▼	44bb5129-8e7d-4409-8a0c-9f4d35248202	2022-09-14 10:24:42	Run Program	complete
		21032	2022-09-14 10:24:42	

그림 2.4.3 - 1 파이프라인 상세 로그 목록 화면

- 해당 로그를 더블 클릭하면 [Console] View에서 분석 과정 및 오류 등의 로그를 확인할 수 있다.

```

Console
Q20 bases: 368626(97.6493%)
Q30 bases: 353708(93.6975%)

Read2 before filtering:
total reads: 2500
total bases: 377500
Q20 bases: 364691(96.6069%)
Q30 bases: 346723(91.8472%)

Read1 after filtering:
total reads: 2464
total bases: 369221
Q20 bases: 361627(97.9432%)
Q30 bases: 347381(94.0848%)

Read2 after filtering:
total reads: 2464
total bases: 369289
Q20 bases: 359500(97.3492%)
Q30 bases: 342526(92.7528%)

Filtering result:
reads passed filter: 4928
reads failed due to low quality: 62
reads failed due to too many N: 0
reads failed due to too short: 10
reads with adapter trimmed: 80
bases trimmed due to adapters: 5528
reads corrected by overlap analysis: 39
bases corrected by overlap analysis: 140

Duplication rate: 1.65724%

Insert size peak (evaluated by paired-end reads): 233

```

그림 2.4.3 - 2 파이프라인 상세 로그 확인

- [GBox Browse] View 를 통해 각 분석 프로그램 별로 지정한 Output 경로로 이동하여 결과를 확인할 수 있다.

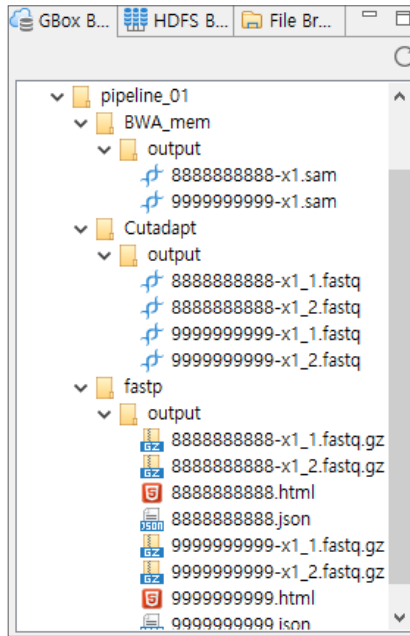


그림 2.4.3 - 3 파이프라인 실행 결과 확인

- [그림2.4.3-4]는 Bioinformatics 카테고리의 공개 프로그램 fastp의 분석 결과이다.

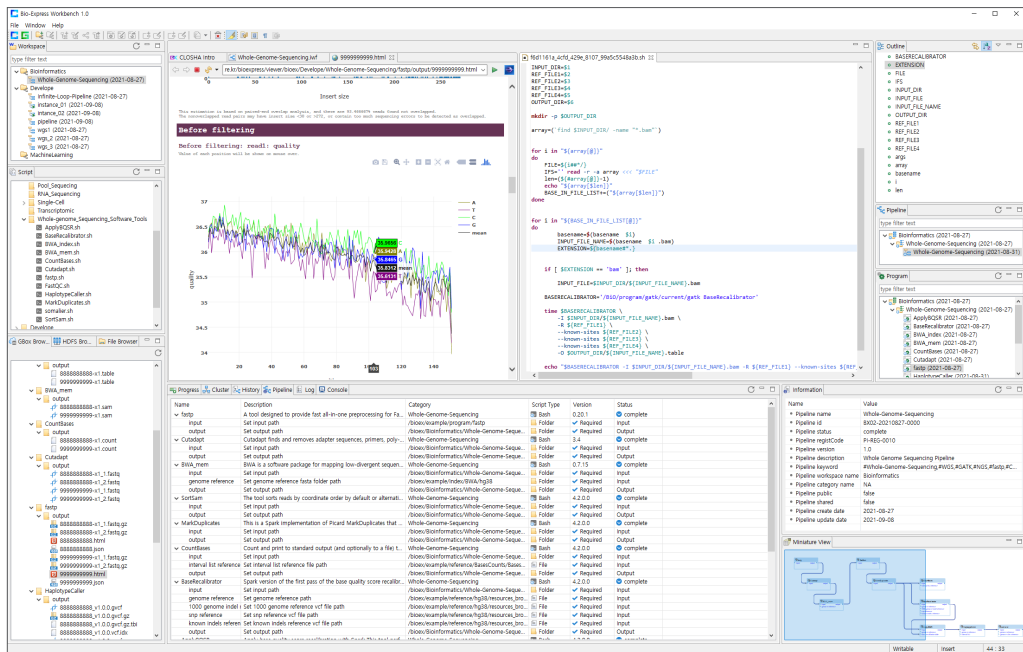


그림 2.4.3 - 4 Bio-Express fastp 프로그램 분석 결과



공개 서비스 등록

2.5.1. 파이프라인 공개 등록

- 특정 사용자와의 분석 파이프라인 공유를 위하여 분석 파이프라인을 공개 등록이 가능하다.
- [그림2.5.1-1]과 같이 [Workspace] View에서 원하는 분석 파이프라인을 선택하고, **컨텍스트** 메뉴 [Register Pipeline] 버튼을 클릭한다.

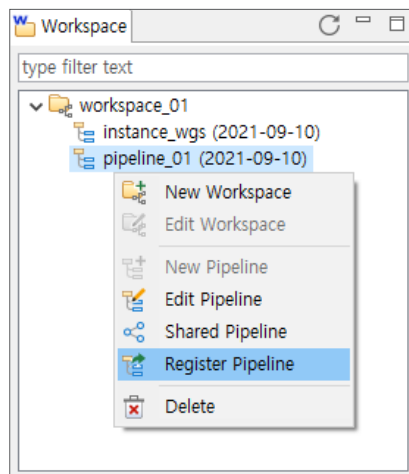


그림 2.5.1-1 파이프라인 공개 등록 요청

- 공개 등록이 요청된 분석 파이프라인은 관리자 승인 절차를 거쳐 공개 분석 파이프라인으로 사용이 가능하다.

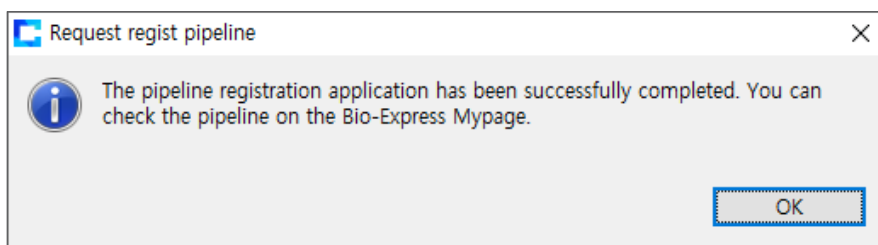


그림 2.5.1-2 파이프라인 공개 등록 요청 완료



BIO-EXPRESS 2.0 USER MANUAL

바이오 익스프레스 2.0 사용자 매뉴얼