ebiogen

User Manual

WinSeurat v3.0

Tel. 02-3141-0791 Fax. 02-3141-0792 E-mail: service@e-biogen.com

<목 차>

1. R installation

㈜이바이오젠은 Single cell RNA-seq 데이터에 대하여 연구자의 필요에 따라 쉽게 분석을 진행할 수 있도록 WinSeurat 을 제공한다. Single cell RNA-seq 을 분석하는 데 기본적으로 사용되는 프로그램인 Cell Ranger 또는 Loupe Browser 에서는 연구자가 분석 옵션을 지정할 수 없기 때문에 원하는 결과를 얻지 못할 수 있다. WinSeurat 은 연구자가 분석 옵션을 조정하여 결과를 확인하고, Cell Ranger 또는 Loupe Browser 에서는 진행하기 어려운 부가적인 분석을 진행할 수 있는 프로그램이다. WinSeurat 은 연구자의 필요성에 따라 지속적으로 업데이트 될 예정이다.

WinSeurat 은 Seurat 이라는 R 패키지를 기반으로 제작되었다. Cell Ranger 를 이용하여 Generate count matrix 결과를 얻은 후 Seurat 을 통하여 분석을 진행한다. Seurat 은 Single cell RNA-seq 분석에서 가장 보편적으로 사용되는 강력한 R 패키지이다.

WinSeurat 을 원활하게 사용하기 위해서는 R 이라는 분석 프로그램을 설치해야 한다. R 프로그램을 설치하기 위해서는 먼저 다음 URL 을 통하여 R 공식 홈페이지에 접속한다 (https://www.r-project.org/). 홈페이지에 접속하면 그림 1-1 과 같은 화면을 볼 수 있는데, 여기서 CRAN 를 클릭한다.

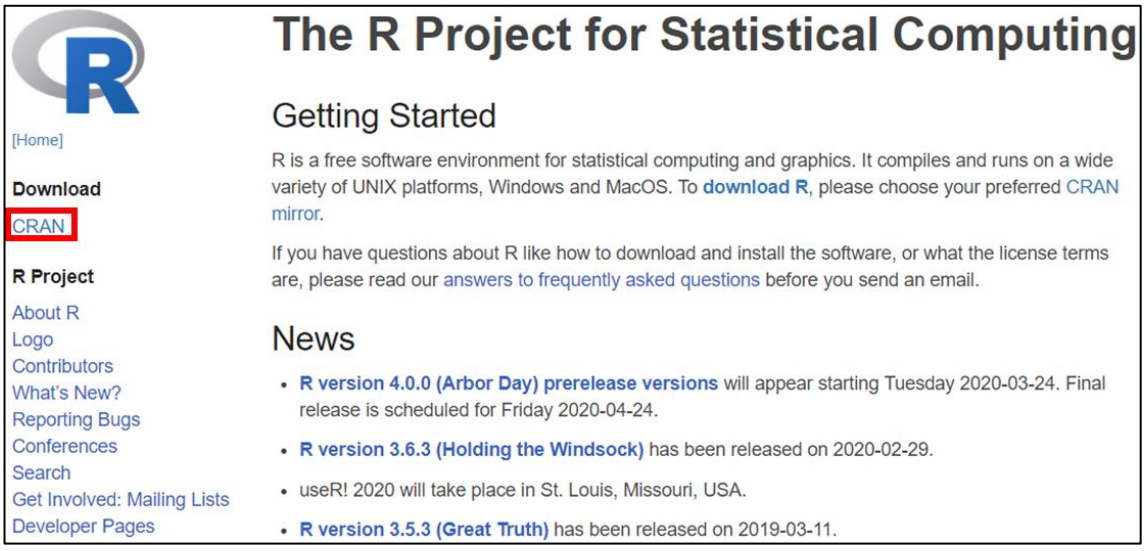

그림 1-1. R 공식 홈페이지 화면

CRAN 를 클릭하면 그림 1-2 와 같이 다운로드 서버를 선택하는 화면이 출력된다. 여기서 0-Cloud 를 선택하면 자동으로 서버가 선택되어 파일을 다운로드 받을 수 있다.

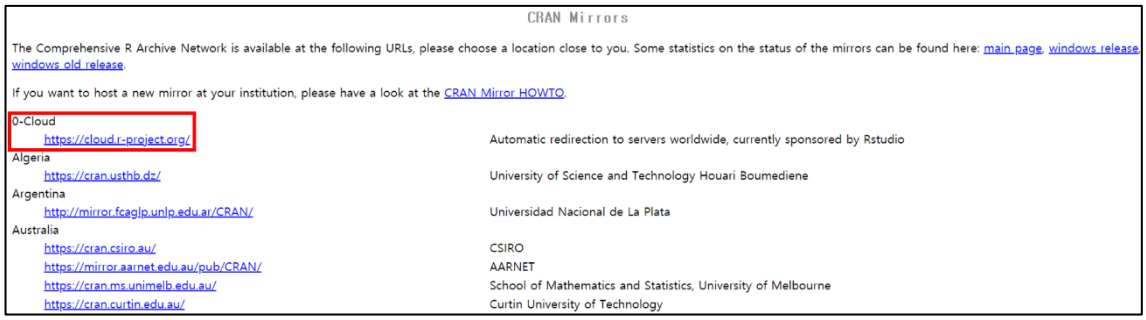

그림 1-2. 다운로드 서버 선택

WinSeurat 은 현재 Windows 운영체제에서만 사용이 가능하다. 그림 1-3 과 같이 Windows 운영체제에서 사용 가능한 R 프로그램을 선택한다.

그림 1-3. Windows 운영 체제 선택

기본적으로 설정된 R 프로그램을 설치하기 위해서는 그림 1-4 와 같이 처음 사용자를 위한 R 설치 항목을 클릭한다.

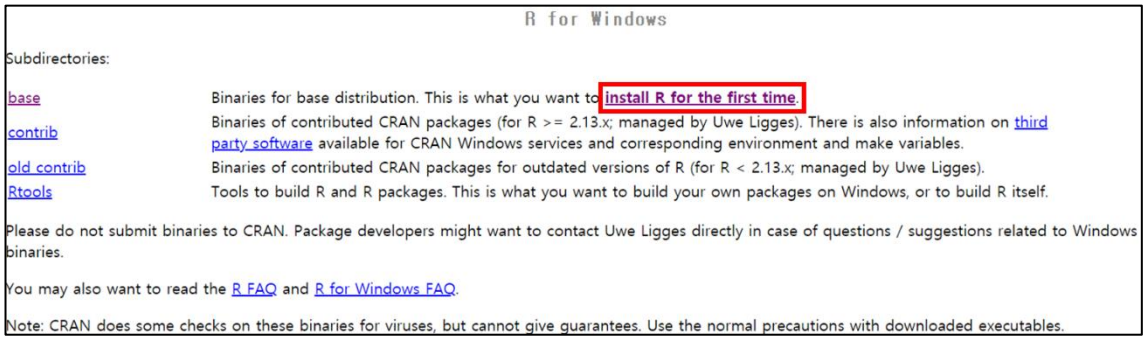

그림 1-4. 처음 사용자를 위한 R 설치

R 프로그램은 지속적으로 업데이트되기 때문에 설치할 때마다 버전이 상이할 수 있다. WinSeurat 을 원활하게 사용하기 위해서는 최신 버전의 R 프로그램보다는 모든 패키지를 정상적으로 사용할 수 있는 4.2.1 버전 이상의 R 프로그램 설치를 추천한다. 그림 1-5 에서 Previous releases 를 클릭하여 이전 버전의 R 프로그램을 검색한다.

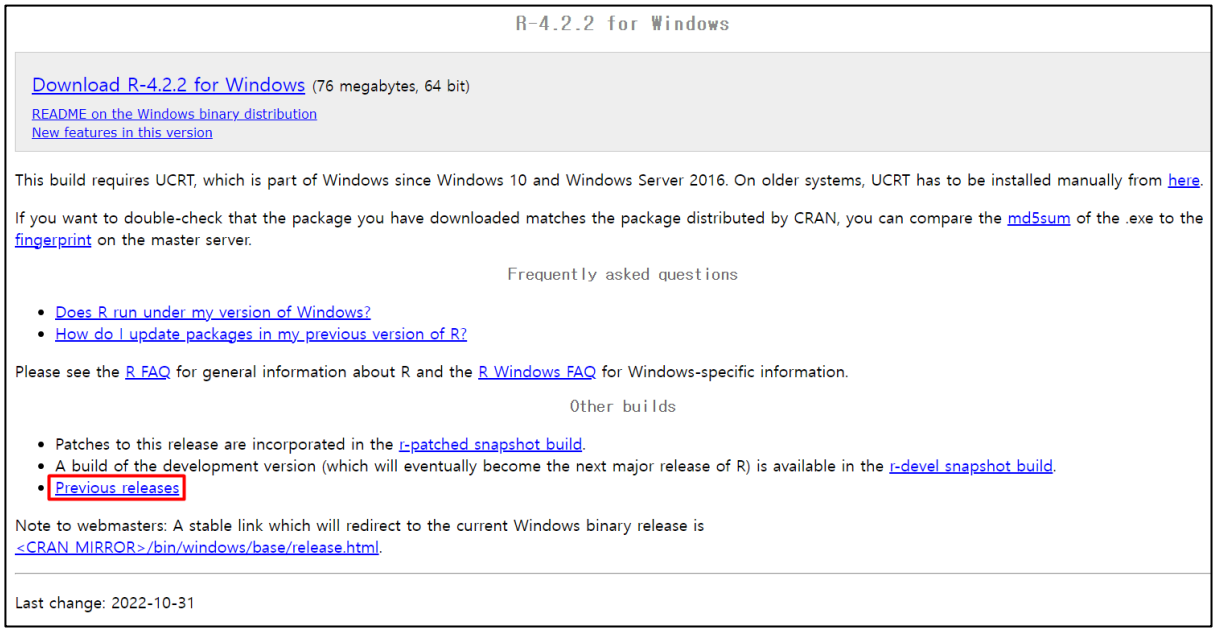

그림 1-5. 이전 버전 R 프로그램 검색

그림 1-6 과 같이 배포된 다양한 버전의 R 프로그램 중 4.2.1 버전을 클릭한다.

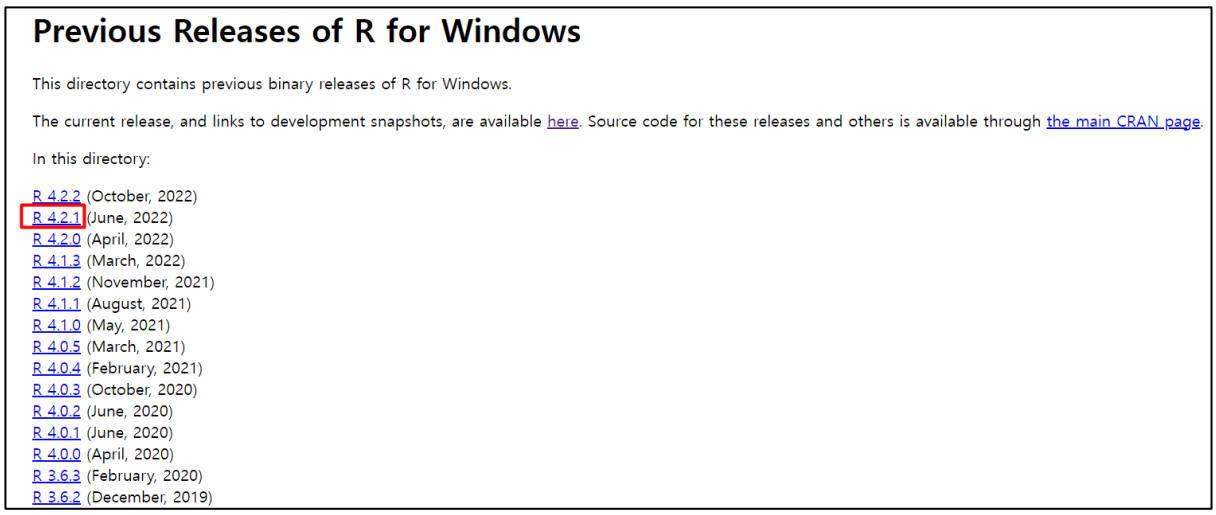

그림 1-6. 4.2.1 버전 R 프로그램 선택

그림 1-7 과 같이 R 프로그램의 버전을 4.2.1 로 맞춘 후 설치 파일을 다운로드 받는다.

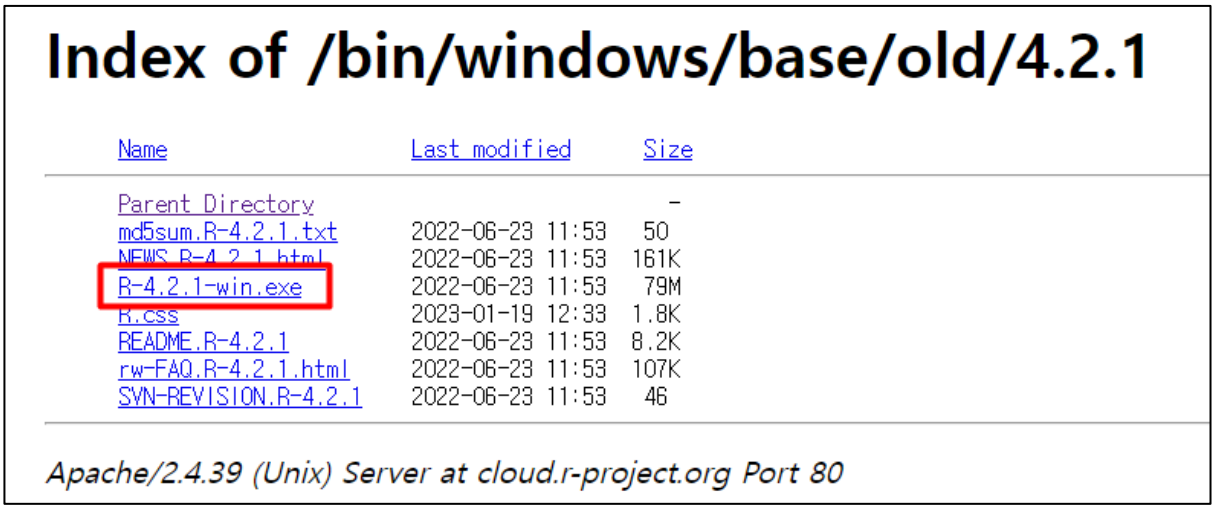

그림 1-7. 설치 프로그램 다운로드

다운로드가 완료된 설치 프로그램을 실행하면 그림 1-8 과 같은 화면이 출력된다. 확인 버튼을 클릭한 후, 그림 1-9 와 같이 다음 버튼을 클릭하여 설치를 진행한다. 그림 1-10 과 같은 화면이 출력되면 4.2.1 버전의 R 프로그램 설치가 마무리된다.

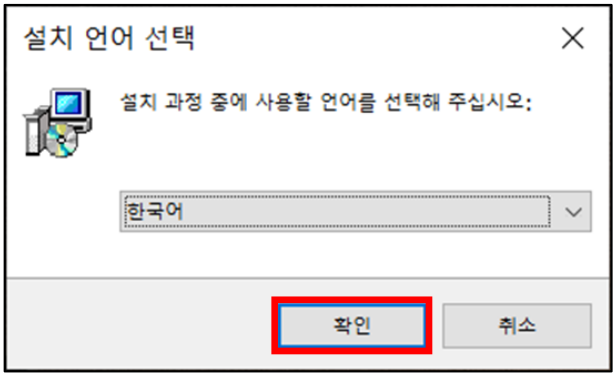

그림 1-8. R 프로그램 설치 초기 화면

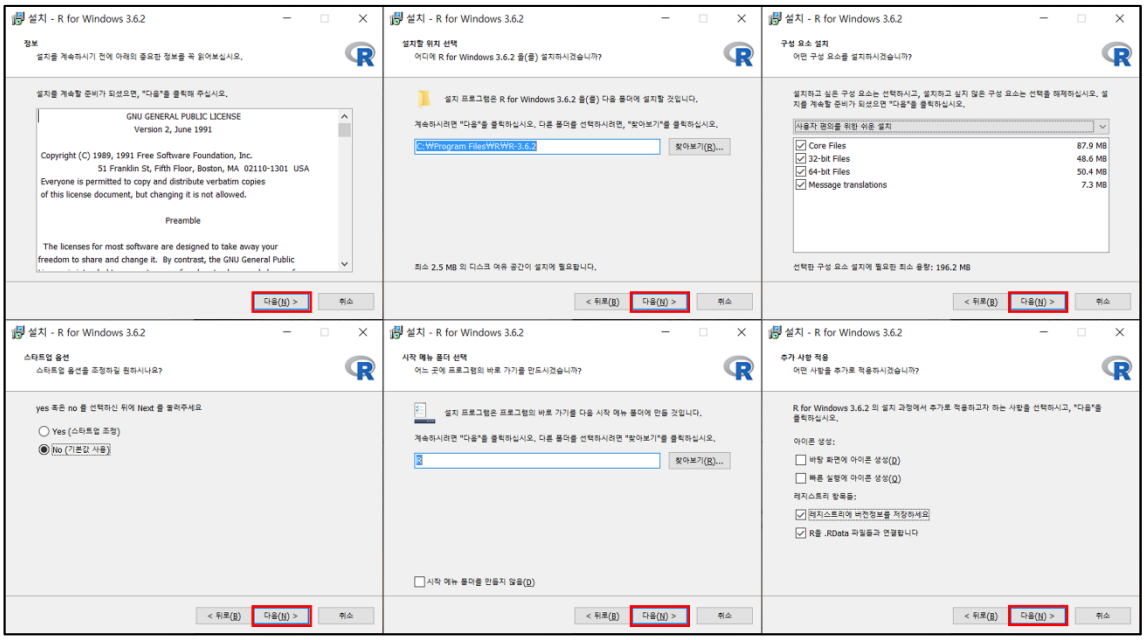

그림 1-9. R 프로그램 설치 화면

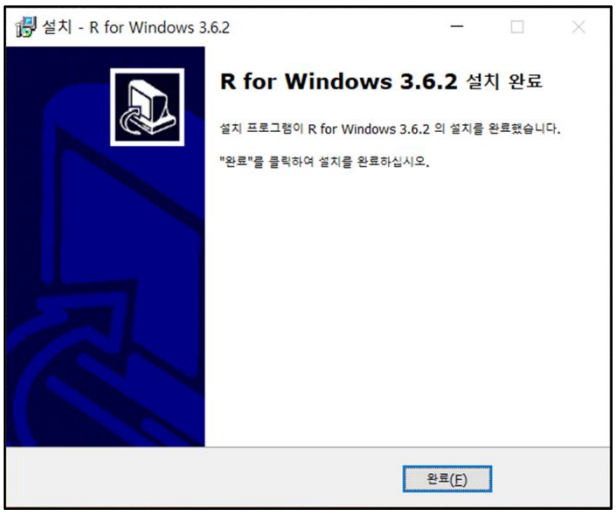

그림 1-10. R 프로그램 설치 완료 화면

2. Home

Home 에서는 WinSeurat 을 사용하기 위하여 사용자 인증을 진행하는 단계인 Login 과 분석을 수행하기 위하여 데이터를 불러오는 단계인 Load data 로 구성된다.

2.1. Login

Login 에서는 FTP 아이디/비밀번호를 이용하여 WinSeurat 을 사용하기 위한 인증을 진행한다. 인증이 완료되어야 WinSeurat 의 기능을 사용할 수 있다. FTP 아이디와 비밀번호를 입력한 후 Login 버튼을 클릭하여 인증을 진행한다.

그림 2-1. Login 기능

2.2. Load data

WinSeurat 을 사용하기 위해서 먼저 ExSCEA Report 파일을 불러와야 한다. 이 파일은 scRNA-Seq 분석 결과 중 하나이며, Import ExSCEA Report 버튼을 클릭하여 불러올 수 있다. ExSCEA 가 정상적인 파일로 확인된 경우에는 data 를 input 할 수 있는 버튼이 활성화된다.

분석 파일을 불러올 수 있는 방법은 크게 두가지가 있다. 첫번째 방법은 Input data 버튼을 통해 WinSeurat input 파일이 들어있는 샘플 폴더 경로를 선택하는 것이다. scRNA-Seq 결과보고 자료를 컴퓨터에 다운로드 받은 뒤 6. WinSeurat > WinSeurat Input 폴더로 접속하면 된다. WinSeurat input 파일은 각 샘플명으로 존재하는 폴더 안에 barcodes, genes (또는 features), matrix 3 가지 파일로 구성되며, 각각의 파일들은 압축 유무에 영향을 받지 않는다. 그림 2-2 는 데이터를 정상적으로 불러오기 위한 폴더 구조를 보여준다. 그림 2-2 와 같이 각 샘플마다 3 가지

파일이 들어있는 폴더 경로를 선택하면 WinSeurat 에서 해당 샘플에 대한 분석을 진행할 수 있다.

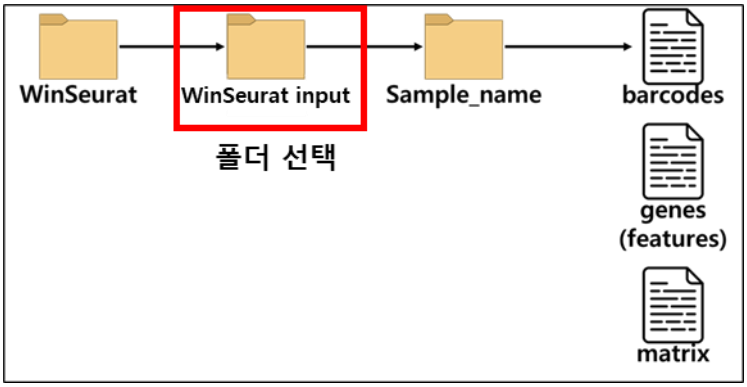

그림 2-2. Load data 수행을 위한 폴더 구조

두번째 방법은 input RDS 버튼을 선택하여 RDS 파일을 불러오는 방법으로 이미 분석이 진행된 결과를 이어서 분석할 수 있다. 이 경우 step1 부터 step3 까지의 데이터 전처리 과정은 이미 진행된 것으로 간주하고 step4 부터 분석을 시작한다. 이를 통해 이전에 분석한 결과를 사용하여 추가분석을 진행할 수 있다.

첫번째 혹은 두번째 방식으로 데이터를 불러오면 로딩되는 동안 프로그램 하단에 진행 아이콘이 표시된다. 데이터 로딩이 완료되면 Done 표시가 뜨면서 그림 2-3 에 보이는 것처럼 Single or Integration 버튼이 활성화 된다. 단일 샘플 분석일 경우 Single 버튼을, 여러 샘플을 통합하여 진행하는 분석일 경우 Integration 버튼을 클릭하여 분석을 시작할 수 있다.

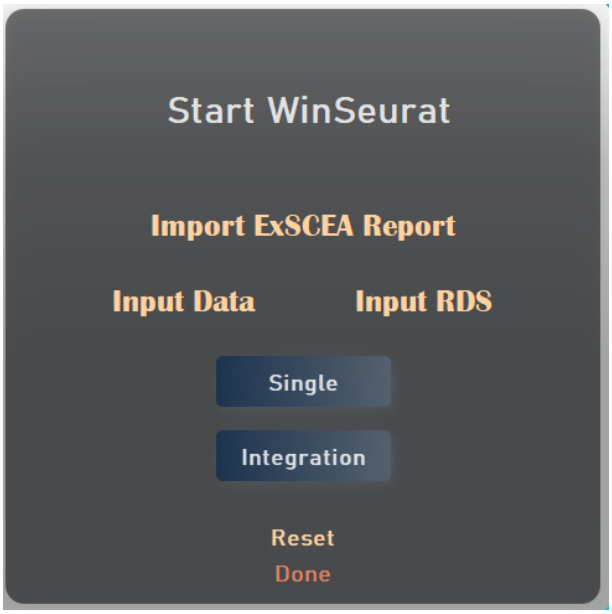

그림 2-3. Load data 기능

3. Analysis

Analysis 는 단일 샘플로 분석을 진행하는 Single sample 과 여러 샘플을 통합하여 분석을 진행하는 Integration 으로 구성된다. Single sample 과 Integration 의 분석 단계는 Data generation, Data processing, Exploratory analysis, Marker identification, Cell type assignment, DE analysis, Additional analysis 로 동일하게 구성된다.

분석을 진행하는 전반적인 흐름은 Single sample 과 Integration 이 유사하기 때문에 본 매뉴얼에서는 Single sample 을 기준으로 설명하며, Integration 에 해당되는 내용은 파란색 글씨로 설명한다.

3.1. Data generation

Integration 에서 Data generation 은 먼저 Data merge 를 진행한다. Data merge 에서는 각 샘플에 대하여 비교를 위한 그룹 이름을 지정한다. Sample list 에 정상적으로 불러온 샘플 데이터들이 표시되고 각 샘플들의 그룹명을 지정한다. 그림 3-1 은 Data merge 화면을 보여준다.

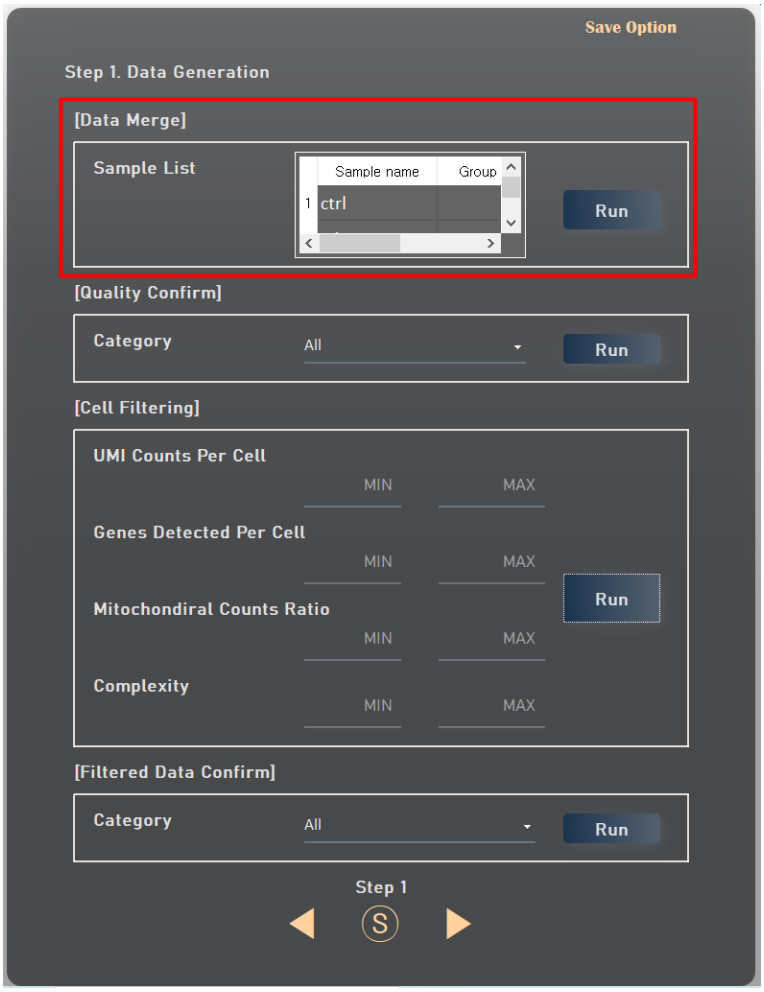

그림 3-1. Data merge 기능

Data generation 은 Quality confirm 과 Cell filtering, Filtered data confirm 으로 구성된다. Quality confirm 단계에서는 입력 데이터의 초기 품질을 확인할 수 있으며, 여섯 가지 지표를 통하여 해당 샘플의 품질을 평가할 수 있다. 샘플의 품질을 평가할 수 있는 지표는 다음과 같다. Cell counts 는 샘플에서 검출된 세포의 수를 의미한다. UMI counts per cell 은 세포 당 검출된 UMI 의 수를 의미한다. UMI 는 고유 분자 식별자로 불리며, 전사체를 검출하고 정량화 하는 데 사용되는 태그이다. Genes detected per cell 은 세포 당 검출된 유전자의 수를 의미한다. UMIs vs Genes detected 는 UMI 당 유전자의 수를 확인할 수 있는 지표를 의미한다. Mitochondrial counts ratio 는 세포에 포함된 미토콘드리아의 비율을 의미한다. Complexity 는 UMI 당 검출된 유전자의 비율에 log10 을 취한 값을 나타낸다. Complexity 가 낮다는 것은 UMI 대비 유전자의 수가 적다는 것을 의미하며, 이는 세포에 포함된 유전자의 종류가 다양하지 않다는 것을 의미한다. 그림 3-1 은 Quality confirm 화면과 결과물인 여섯 가지 지표 그래프를 보여준다.

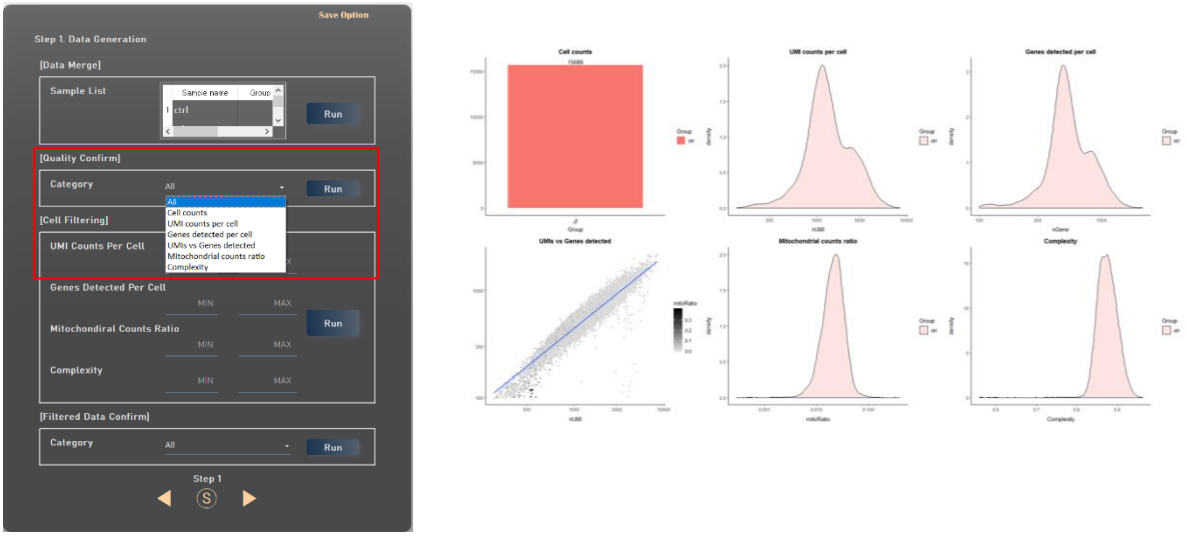

그림 3-2. Quality confirm 기능

Cell filtering 단계에서는 분석에 부적합한 세포를 제거한다. 각 지표에 대한 임계값은 연구자가 지정할 수 있으며, Cell filtering 에 기본적으로 적용되는 임계값은 표 1 과 같다.

표 1. Cell filtering 기본값

Category	최소값	최대값
UMI range	500	세포의 최대 Gene count per cell
Gene range	300	세포의 최대 UMI count per cell
Mitochondrial ratio	세포의 최소 Mitochondrial ratio	세포의 최대 Mitochondrial ratio
Complexity	세포의 최소 Complexity	세포의 최대 Complexity

그림 3-3 은 Cell filtering 화면과 최소/최대값이 이미지 상에 적용된 지표 그래프를 보여준다. 파란색 선은 각 지표의 최소값 (값을 입력한 경우 해당 값)을, 빨간색 선은 각 지표의 최대값 (값을 입력한 경우 해당 값)을 의미한다. 각 지표를 모두 만족하는 세포들이 추출된다.

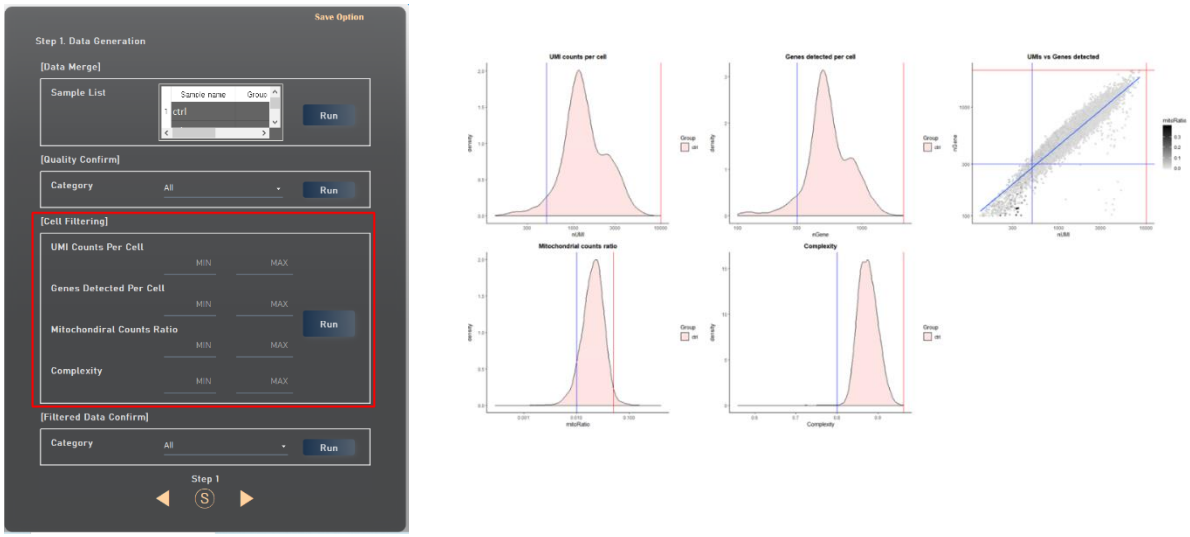

그림 3-3. Cell filtering 기능

Filtered data confirm 단계에서는 앞서 필터링 된 데이터의 여섯 가지 지표를 최종적으로 확인한다. 이 단계는 반드시 거쳐야 하는 단계가 아니며, Cell filtering 단계에서 원하는 결과를 확인한 경우에는 해당 단계를 진행하지 않아도 된다. 그림 3-4 는 Filtered data confirm 화면과 필터링 된 데이터에 대한 지표 그래프를 보여준다.

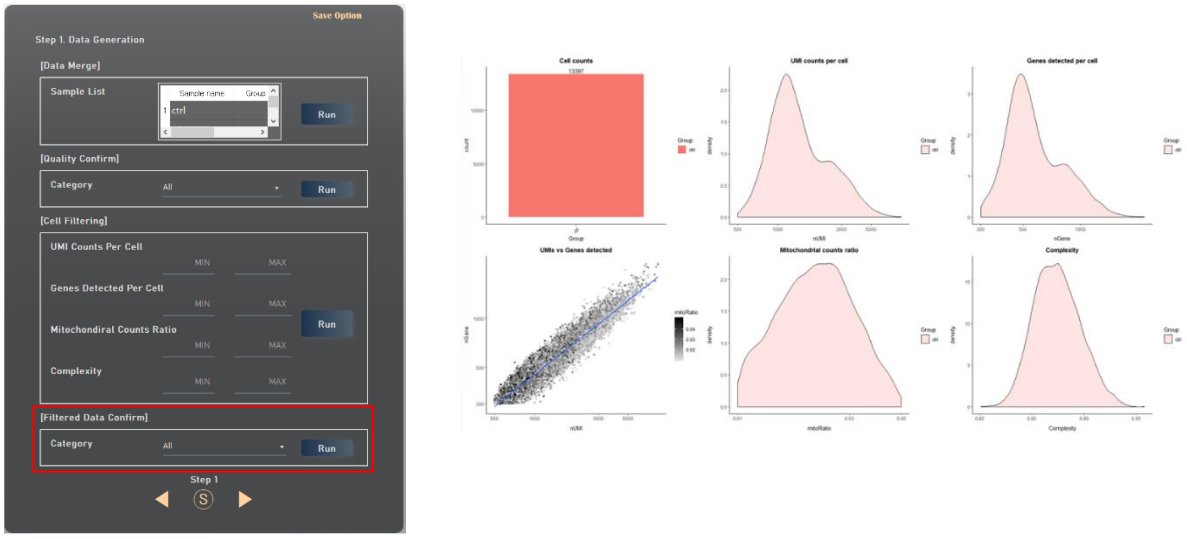

그림 3-4. Filtered data confirm 기능

참고로 WinSeurat 으로 분석된 이미지는 그림 3-5.와 같이 파일 > 다른 이름으로 저장 > 원하는 형식 지정하여 저장할 수 있다.

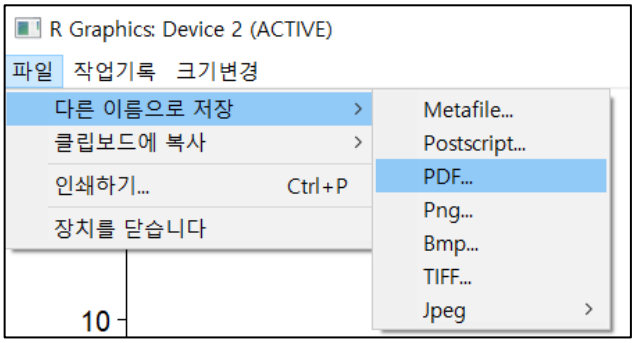

그림 3-5. 이미지 저장

3.2. Data processing

Data processing 은 Cell cycle confirm 과 Normalization, Feature Selection 으로 구성된다. 연구 방향에 따라 세포 주기를 고려해야 하는 경우가 발생한다. WinSeurat 에서는 Species 가 Human 을 포함한 3 가지 종에서 세포 주기에 따른 세포 발현을 확인하고, 이후 분석 단계에서 세포 주기 고려 여부를 결정할 수 있다. Cell cycle confirm 단계는 Species 항목에서 샘플의 종을 선택하고, Number of features 항목에 이미지로 표현하기 위하여 차원 축소에 사용하려는 유전자의 수를 입력한다. 분석하려는 종이 Species 항목에 없을 경우 other 를 선택한다. Number of features 의 기본값은 2000 이다. 그림 3-6 는 Cell cycle confirm 화면과 Species 가 Human 인 경우에 대한 세포 주기에 따른 세포 발현 그래프를 보여준다.

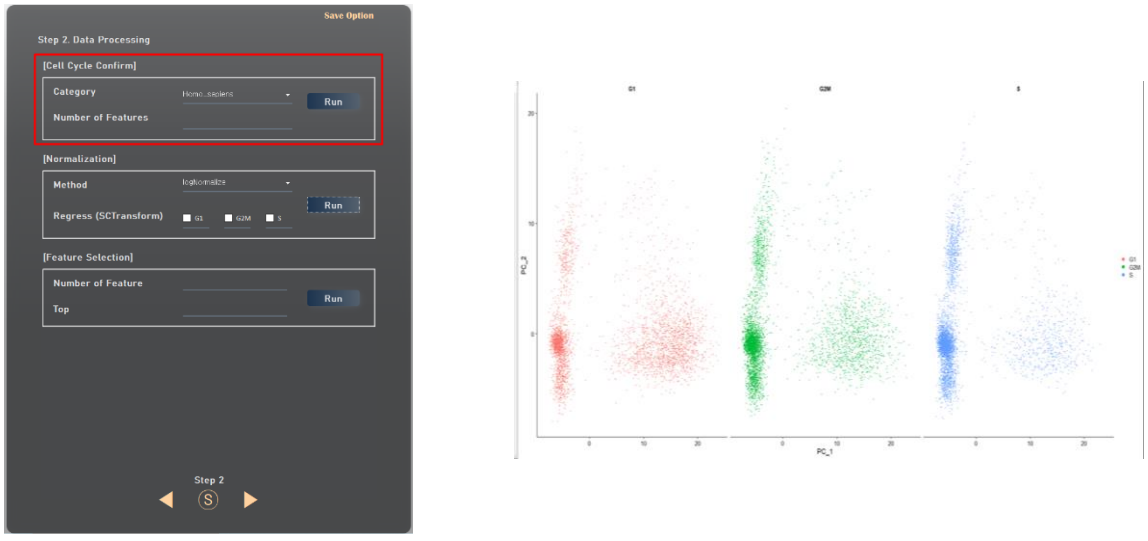

그림 3-6. Cell cycle confirm 기능

Normalization 단계에서는 세포 간의 발현에 대한 정확한 비교를 위하여 Read count 를 기반으로 정규화를 진행한다. WinSeurat 은 정규화 방법으로 SCTransform 과 LogNormalize 가 제공된다.

SCTransform 은 Sequencing depth 를 고려함과 동시에 미토콘드리아와 세포 주기 단계를 고려할 수 있는 정규화 방법이며, logNormalize 는 Sequencing depth 만을 고려하는 정규화 방법이다. Regress 는 정규화 방법이 SCTransform 인 경우에만 적용된다. 세포 주기로 인하여 데이터가 변질된 경우에는 해당 세포 주기를 분석에서 제외할 수 있다. Regress 항목에서 각 세포 주기 (G1, G2M, S) 중 분석에서 제외할 단계를 선택할 수 있다. 그림 3-7 은 Normalization 화면을 보여준다.

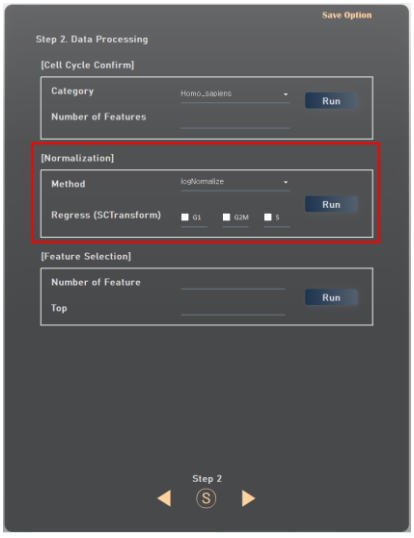

그림 3-7. Normalization 기능

Feature selection 단계는 단일세포 데이터 세트에서 유의한 유전자를 선택하기 위한 중요한 단계이다. 이 단계는 셀 간 분산 차가 큰 유전자를 선택하여, 세포 유형 간의 차이점을 파악할 수 있도록 하는 것이다. Seurat 프로세스는 이러한 평균-분산 관계를 직접 모델링 하여 단일 셀 데이터에 내제된 관계를 잘 반영한다. Seurat 는 기본적으로 데이터 세트 당 2000 개의 feature(gene)을 선택하도록 되어있으며, feature 수를 조정할 수도 있다. 데이터 세트 크기에 따라 feature 수를 조정하여 분석 속도를 향상시키거나, 높은 수준의 분석을 위해 feature 수를 늘릴 수 있다. 이 단계에서 단일 셀 데이터에서 평균-분산 관계를 계산하고 상위 N 개의 Feature 를 반환한다. 이 N 값은 원하는 수로 지정할 수 있다.

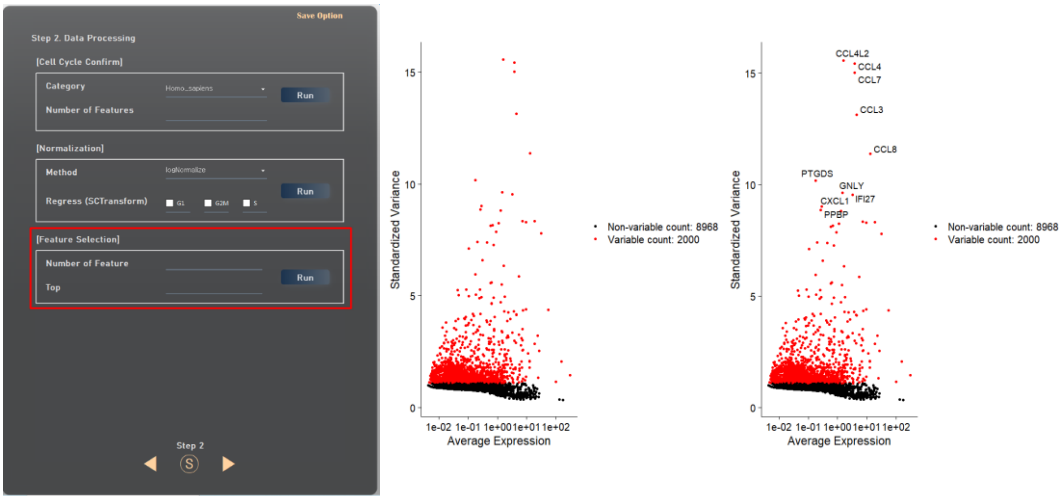

그림 3-8. Feature Selection 기능

3.3 Exploratory analysis

Exploratory analysis 는 PCA 와 Confirm the optimal number of PC, Clustering 으로 구성된다. PCA 단계에서는 앞서 정규화 된 데이터를 기반으로 주성분 분석을 진행한다. 주성분 분석은 고차원 데이터의 특성을 유지하면서 저차원 데이터로 차원을 축소하는 기법을 의미한다. Number of PC 항목은 축소하려는 차원의 개수를 의미하며, 기본값은 30 이다. 그림 3-9 은 PCA 화면을 보여준다.

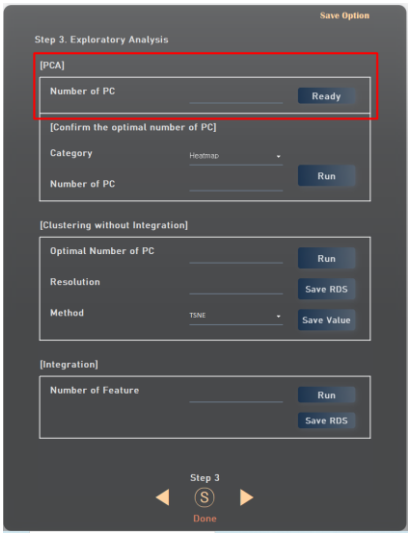

그림 3-9. PCA 기능

Confirm the optimal number of PC 에서는 세 가지 지표를 통하여 최적의 주성분 개수를 식별한다. SCTransform 으로 정규화를 진행한 경우 데이터의 편향을 제거했기 때문에 주성분 개수에 크게 영향을 받지 않지만, logNormalize 로 정규화를 진행한 경우에는 주성분의 개수가 분석에 큰 영향을 미친다. 최적의 주성분 개수를 식별하기 위한 지표로 Heatmap, Jackstraw plot, Elbow plot 이 제공된다. Heatmap 은 각각의 주성분을 구성하는 유전자들을 기반으로 데이터를 명확하게 구분할 수 있는지 평가할 수 있는 지표이다. WinSeurat 에서는 모든 주성분의 Heatmap 을 그리는 것이 어렵기 때문에, 가급적 상위 20 개 이내의 주성분에 대한 Heatmap 을 그리는 것을 권장한다. Jackstraw plot 은 각각의 주성분과 이들에 포함된 유전자 사이의 연관성을 계산하여 P-value 로 제공하는 그래프이다. JackStraw plot 에서는 P-value 를 기반으로 최적의 주성분 개수를 식별한다. 참고로 scTransform 으로 보정을 했을 경우 Jackstraw plot 방식은 seurat 상에서 권장하지 않기에 비활성화된다. Elbow plot 은 주성분의 표준 편차를 기반으로 구축된 그래프이다. Elbow plot 에서는 표준 편차의 변화가 적은 부분을 최적의 주성분 개수로 식별한다. Heatmap 과 JackStraw plot, Elbow plot 을 참고하여 연구자가 최적의 주성분 개수를 식별한다. 그림 3-10 은 Confirm the optimal number of PC 화면과 결과물인 Heatmap, Jackstraw plot, Elbow plot 이미지를 보여준다.

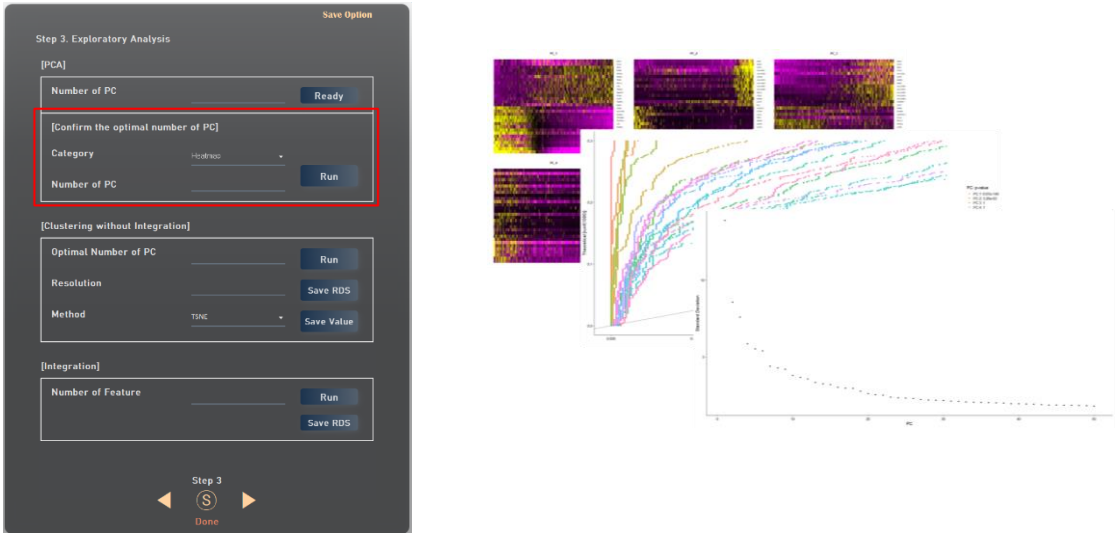

그림 3-10. Confirm the optimal number of PC 기능

Clustering 에서는 앞서 식별된 최적의 주성분 개수를 이용하여 세포 군집을 구성한다. Optimal number of PC 에는 앞서 세 가지 지표로 확인된 최적의 주성분 개수를 입력한다. Resolution 은 Clustering 단계에서 밀집도를 조절하여 군집의 개수를 조정할 수 있으며, 기본값은 0.8 이다. 군집 개수를 줄이려면 resolution 값을 내리고, 반대로 군집 개수를 올리려면 resolution 값을 올리면 적용할 수 있다. Method 는 군집 알고리즘을 선택할 수 있는 항목으로, TSNE 와 UMAP 중 하나를 선택할 수 있다. Trajectory analysis 와 같은 추가 분석은 현재 UMAP 을 기반으로 구축된 데이터로만 지원되기 때문에, 이러한 분석을 진행하고자 한다면 UMAP 을 이용하여 군집 분석을 진행해야 한다. 그림 3-11 는 Clustering 화면과 결과물인 군집 이미지를 보여준다. 제작한 UMAP/tSNE 에 대한 분석결과를 저장하고자 할 경우에는 Save RDS 버튼을 클릭하여 RDS 형식으로 저장한 뒤, 추후 WinSeurat 에서 다시 불러올 수 있다. UMAP/tSNE 에 대한 좌표 정보를 저장하고자 할 경우 Save value 버튼을 클릭하면 된다.

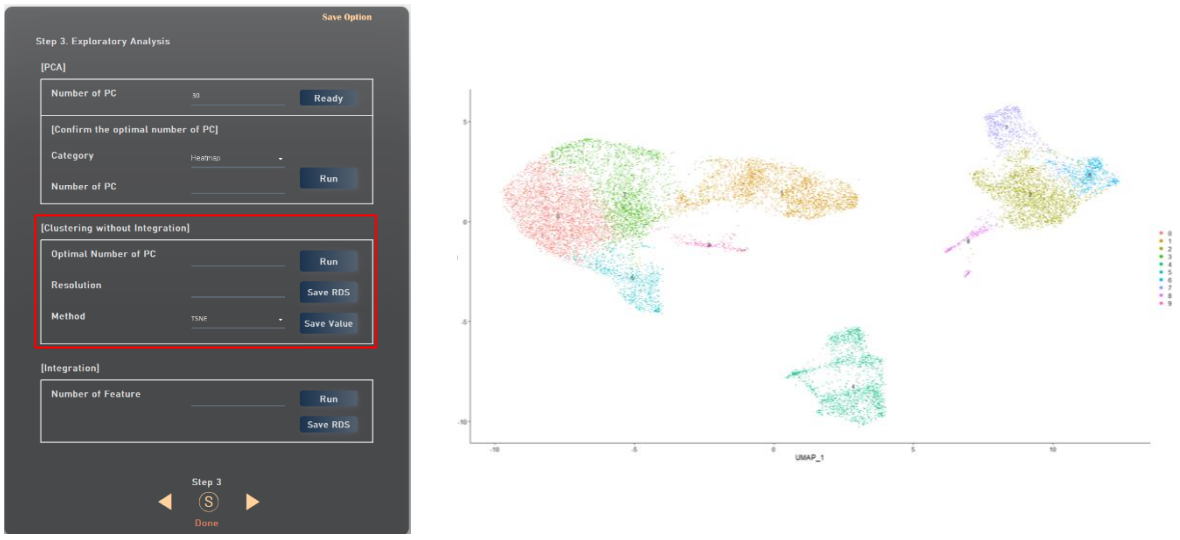

그림 3-11. Clustering 기능

Integration 에서는 Clustering 이 두 단계로 구분된다. Clustering without integration 은 샘플 간의 Batch 를 고려하지 않은 상태에서 군집을 형성하는 단계이며, Integration 은 샘플 간 유사한 상태의 세포들의 위치를 기반으로 서로 다른 데이터를 하나의 공간에 배치하는 단계이다. Clustering without integration 은 Single sample 의 Clustering 단계와 동일한 옵션들이 사용된다. Integration 에서는 서로 다른 샘플들을 하나의 공간에 배치하기 위하여 기준점이 되는 유전자의 수를 입력한다. 기준점으로 사용되는 옵션인 Number of features 의 기본값은 2000 이다. 그림 3-12 은 Clustering without integration 의 결과물인 군집 이미지를, 3-13 은 Integration 의 결과물인 군집 이미지를 보여준다. Integration 까지 진행된 분석 데이터를 저장하고자 할 경우 Save RDS 를 통해 RDS 형식의 데이터로 저장할 수 있다.

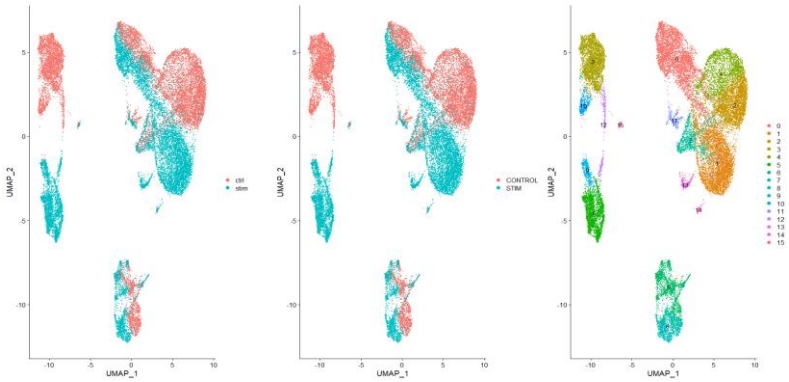

그림 3-12. Clustering without integration 기능

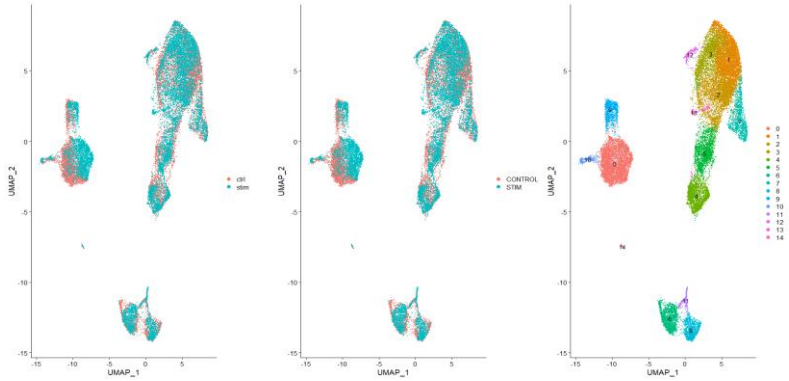

그림 3-13. Integration 기능

3.4 Marker identification

Marker identification 은 Finding markers 와 Marker expression 으로 구성된다. Finding markers 단계는 P-value 를 기반으로 다른 군집들과 비교하여 발현 차이가 나타나는 유전자들이 추출된다. WinSeurat 은 P-value 를 계산하는 다양한 방법들을 제공한다. 표 2 는 WinSeurat 에서 제공하는 P-value 계산 방법들과 이들의 간략한 설명을 보여준다. Loupe Browser 에서는 wilcox 를 이용하여 P-value 를 계산하기 때문에, WinSeurat 에서도 동일하게 기본값으로 wilcox 를 사용한다. Integration 에서는 wilcox 를 이용하여 P-value 가 계산된다.

표 2. P-value 계산 방법

Finding method	Description
wilcox	Wilcoxon rank sum test (Default).
bimod	Likelihood-ratio test for single cell feature expression.
roc	Standard AUC classifier.
t	Student's t-test.
negbinom	Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets.
poisson	Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets.
LR	Uses a logistic regression framework to determine differentially expressed genes. Constructs a logistic regression model predicting group membership based on each feature individually and compares this to a null model with a likelihood ratio test.
MAST	GLM-framework that treates cellular detection rate as a covariate.
DESeq2	DE based on a model using the negative binomial distribution.

Cluster 항목에서는 마커 유전자를 식별하려는 군집을 선택할 수 있으며, 모든 군집 또는 특정 군집을 지정할 수 있다. Positive/Negative marker 항목에서는 Fold change 가 0 보다 큰 항목만 추출할 것인지 또는 절대값이 큰 항목을 추출할 것인지 선택할 수 있다. Number of markers 는 식별하고자 하는 마커 유전자의 수를 의미하며, 기본값은 10 이다. Finding markers 의 결과물은 해당 군집에 대하여 P-value 가 가장 유의한 상위 유전자들에 대한 정보가 텍스트 파일로 제공되며, Cluster 항목에서 모든 군집을 선택한 경우 각 군집에 대한 Clustering heatmap 이미지가 생성된다. 그림 3-14 는 Finding markers 화면과 결과물인 Clustering heatmap 이미지를 보여준다.

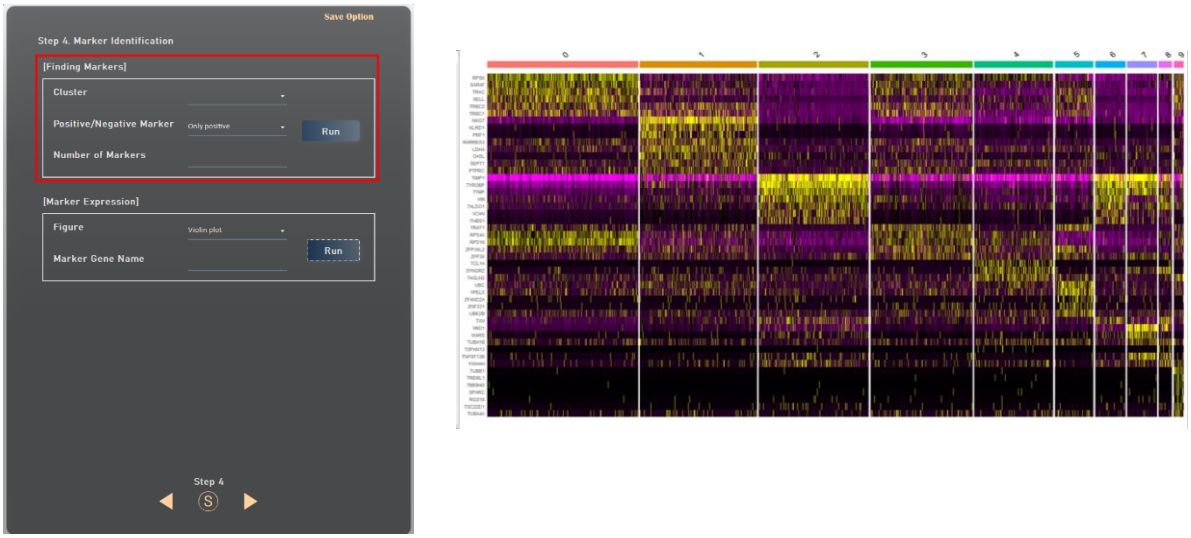

그림 3-14. Finding markers 기능

Marker expression 단계는 각 군집에 대하여 마커 유전자의 발현 수준을 확인한다. 앞서 추출된 마커 유전자가 다른 군집들과 비교하여 해당 군집에서 유의한 발현 수준을 보이는지 시각적으로 확인할 수 있다. Figure 항목에서 Violin plot 또는 Clustering 을 선택하여 원하는 이미지를 생성할 수 있다. Marker gene name 항목에 유전자 이름을 입력하면 모든 군집에 대하여 해당 유전자의 발현 정보를 볼 수 있다. 여러 유전자들의 발현 수준을 동시에 보기 위해서는 유전자 이름을 쉼표 (,)로 구분되도록 입력한다. 그림 3-15 은 Marker expression 화면과 결과물인 Violin plot 및 Clustering 상의 발현 수준 이미지를 보여준다.

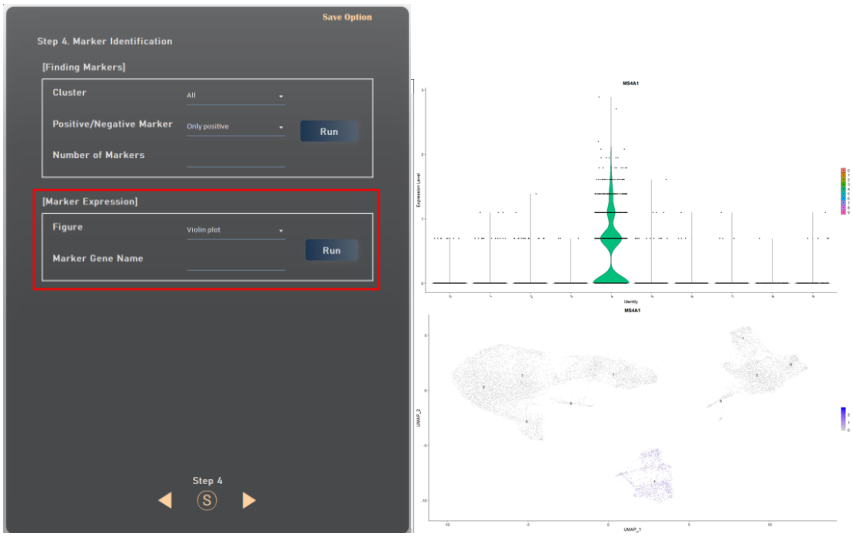

그림 3-15. Marker expression 기능

3.5 Cell type assignment

Cell type assignment 는 cell type 을 지정하는 탭으로써 알고 있는 정보를 기반으로 직접 지정하는 Specify manual cell type 한 방식과 DB 를 이용하는 Specify automatic cell type 두가지 방식으로 구성된다. 먼저, Specify manual cell type 방식은 연구자가 알고 있는 세포 타입과 이를 구성하는 마커 유전자들을 이용하여 어떤 군집이 해당 세포 타입인지 확인한 후, 군집 이미지 상에 해당 세포 타입 이름을 적용한다. Cluster name 항목에 이름을 변경하려는 군집을 선택하고 Assign cell type 항목에 변경하고자 하는 이름을 입력한다. Assign 버튼을 이용하여 변경하려는 군집 이름을 적용하고, Run 버튼을 이용하여 적용이 완료된 군집 이미지를 생성한다. 예를 들어, MS4A1 은 B cell 의 마커 유전자이며, GNLY 와 NKG7 은 NK cell 의 마커 유전자이다. Marker expression 을 확인한 결과 1 번 군집이 NK cell, 4 번 군집이 B cell 인 것으로 예측되었으며, 해당 세포 타입을 군집 이미지에 적용하였다. 그림 3-16 는 Assign cell type 화면과 결과물인 세포 타입의 이름이 적용된 Clustering 이미지를 보여준다.

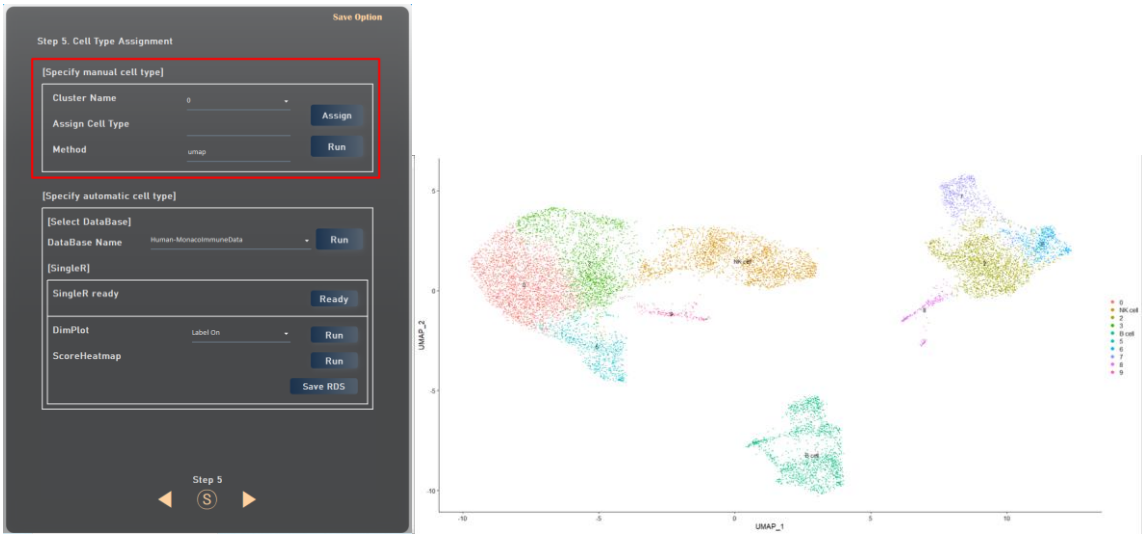

그림 3-16. Specify manual cell type 기능

만약 매뉴얼로 직접 지정하지 않고 Database 를 이용하여 cell type 을 자동 지정하고자 할 경우, 그림 3-17 처럼 Specify automatic cell type 에서 지정할 수 있다. SingleR 패키지를 이용하면 자주 분석되는 인간과 마우스 종에 대해 미리 구성된 데이터베이스를 활용하여 cell type 을 자동으로 매칭할 수 있다. 이를 위해서 Select database 탭에서 사용하려는 database 를 선택하고 Ready 버튼을 클릭하여 패키지를 준비해야 한다. 이후 DimPlot 에서 label on/off 을 선택하여 UMAP/tSNE 이미지에 cell type 을 표시하거나 표시하지 않을 수 있다. 여기까지 진행된 데이터를 저장하고자 할 경우 Save RDS 버튼을 이용하여 RDS 형식으로 데이터를 저장할 수 있다. 또한 Score Heatmap 을 이용하여 cell type 별 발현 양상을 heatmap 으로 제작할 수 있다. 이 때, 값은 패키지에서 제공되지 않기 때문에 참고용 이미지로 사용할 수 있다. 자동으로 매칭되는 기능이 잘 동작하지 않는 경우에는 다시 위쪽의 Specify manual cell type 방식을 이용하여 수동으로 cell type 을 지정할 수도 있다.

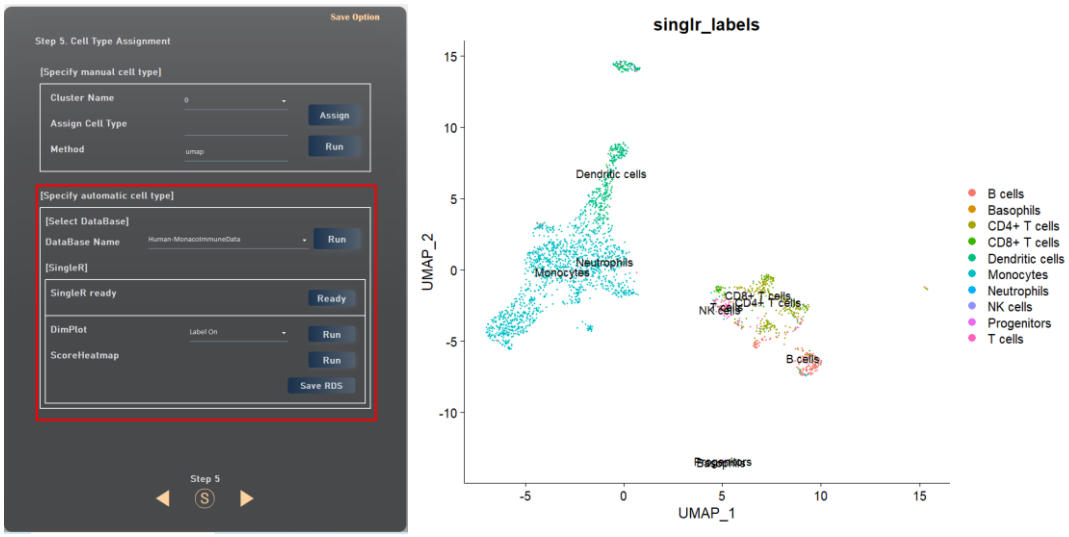

그림 3-17. Specify automatic cell type 기능

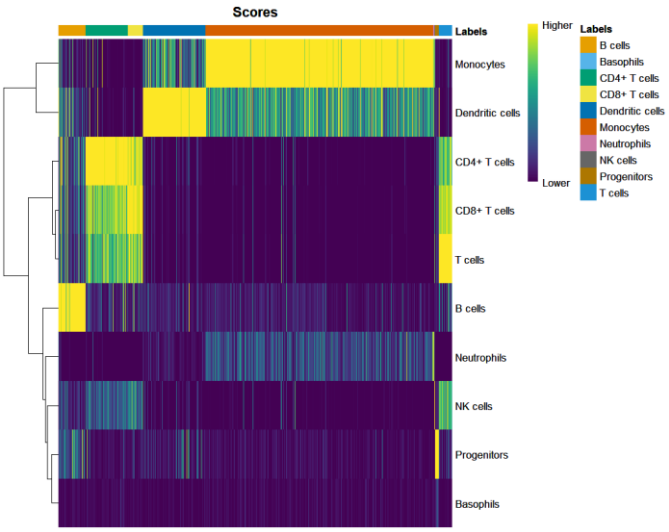

그림 3-18. Score Heatmap 기능

3.6 DE analysis

DE analysis 는 Volcano plot 과 Expression level 로 구성된다. Volcano plot 단계는 서로 다른 두 비교 집단에 대하여 P-value 와 Fold change 를 기반으로 비교한 이미지를 보여준다. Group1 항목과 Group2 항목에 비교하고자 하는 군집의 번호를 쉼표 (,)로 구분되도록 입력한다. P-value 와 Fold change 항목은 연구자가 유의하다고 판단하는 임계값을 입력한다. 그리고 유의한 유전자를 그래프상에 표현하고 싶으면 Number of Gene mark 에 표현하고 싶은 상위 유전자 개수를 입력한다. 그림 3-19 는 Volcano plot 화면과 결과물인 두 비교 집단의 Volcano plot 비교 그래프를 보여준다.

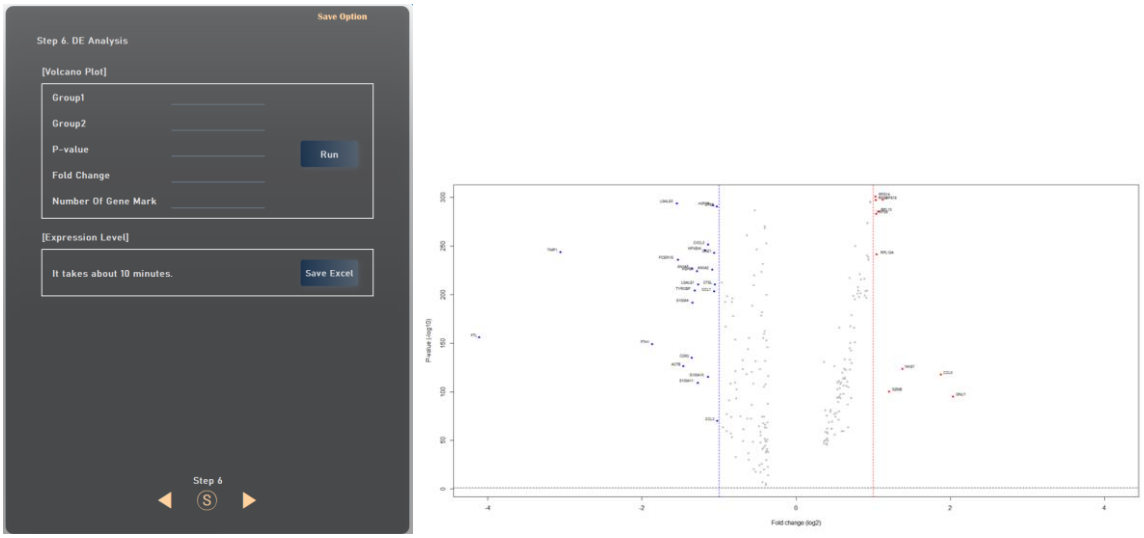

그림 3-19. Volcano plot 기능

빨간색 점선과 파란색 점선은 입력한 Fold change 임계값을, 검은색 점선은 입력한 P-value 임계값을 표시한다. 빨간색 영역에는 Group2 보다 Group1 에서 유의한 유전자들이 표시되며, 파란색 영역에는 Group1 보다 Group2 에서 유의한 유전자들이 표시된다.

Expression level 에는 volcano plot 에 해당하는 value 값을 엑셀로 저장하는 기능이다. 컴퓨터 사양에 따라 차이는 있지만 대략 10 분 전후로 소요되기에 필요할 경우 여유를 두고 진행하는 것을 권장한다.

3.7 Additional Analysis

Additional Analysis 는 Multiple plot 으로 각 Cluster 별 유전자 발현을 시각화 하는 기능이다. 그림 3-20 처럼 Ridge plot 과 dot plot 중 원하는 plot 을 선택하고, 직접 gene name 을 작성하거나 텍스트파일 형식으로 작성한 gene list 를 불러와서 분석할 수 있다. 직접 gene name 을 작성할 경우에는 Input gene symbol 탭에서 (,)형태로 공백 없이 gene 을 표기한다. 예를 들어 2 가지 gene name 을 기입할 경우 다음과 같이 CD79A,TCL1A 형식으로 기입한다. Gene list 가 많아서 여러 개를 표기하고자 할 경우에는 텍스트 형식의 파일로 gene list 를 만든 뒤, Import gene list(.txt) 탭에 import 버튼으로 파일을 불러올 수 있다. 파일 제작 방식은 관심있는 gene list 를 엑셀파일 1 행에 Enter 형식으로 쭉 나열한 뒤 다른 이름으로 저장 > 텍스트 탭으로 분리(.txt) 형식으로 저장한다. 단, 한 화면에 plot 개수가 많아져 복잡해질 수 있기에 gene 개수는 1 개에서 10 개 이내로 제작하는 것을 권장한다.

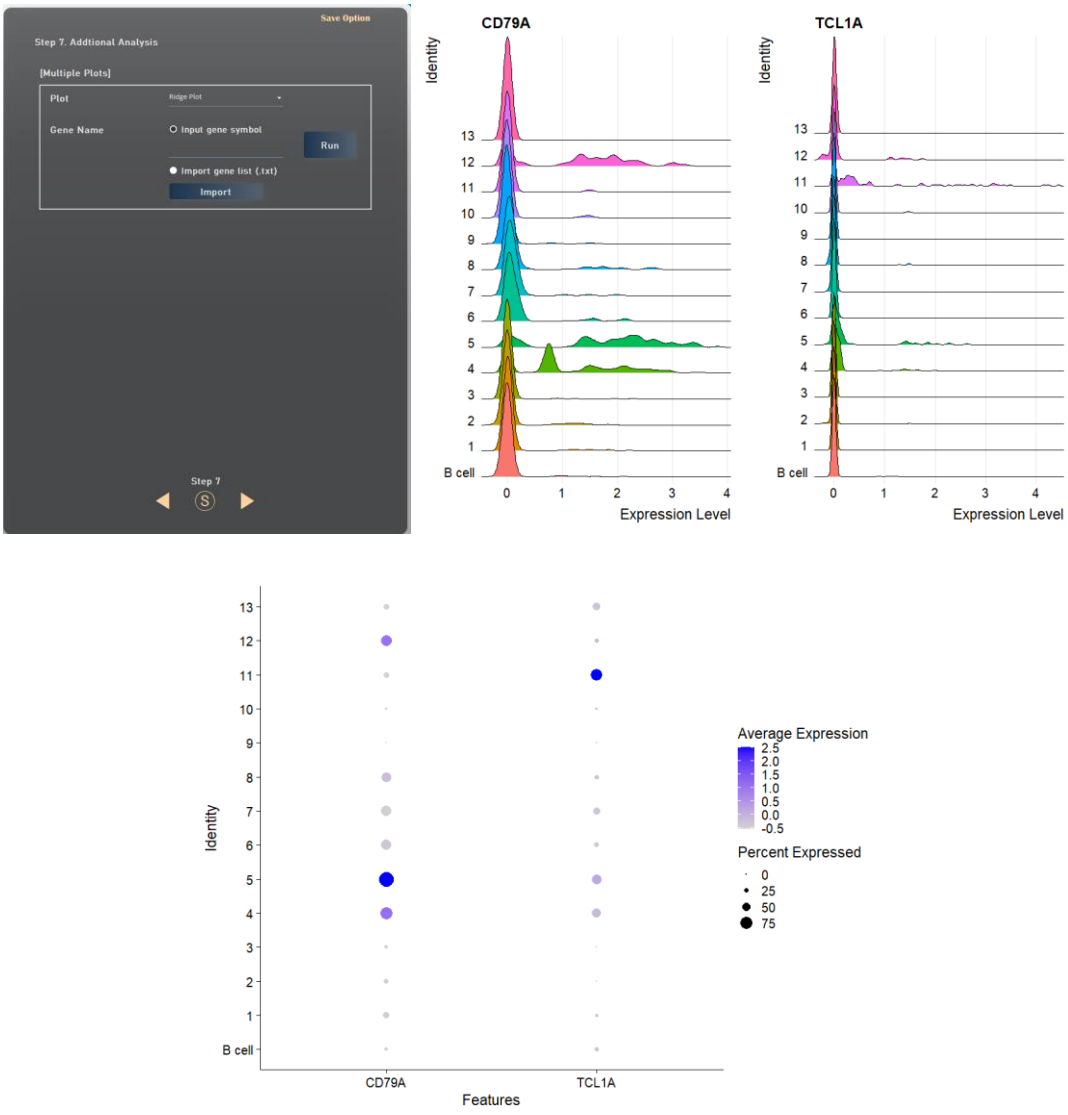

그림 3-20. Multiple plot 기능

4. Additional Function

4.1 Save option

분석에 사용된 옵션들은 그림 4-1 과 같이 Save option 버튼을 클릭하여 텍스트 파일로 저장할 수 있다. WinSeurat 프로그램을 종료한 후에 동일한 옵션으로 분석을 다시 진행하려는 경우에는 저장된 옵션 정보를 참고하여 진행할 수 있다.

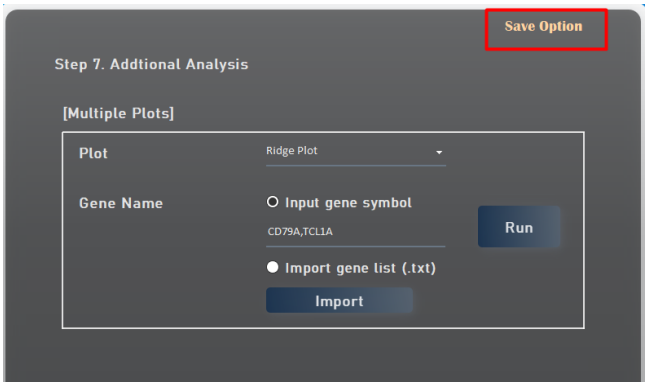

그림 4-1. Save option 기능

4.2 Reset data

분석이 완료된 후 새로운 데이터로 분석을 진행하고자 한다면, 그림 4-2 와 같이 화면 하단의 S 버튼을 클릭하여 파일 import 하는 홈화면으로 이동할 수 있다. 그림 4-3 과 같이 Reset 버튼으로 메모리 캐쉬를 비워주고 원하는 데이터를 다시 Input 하면 새로운 데이터로 WinSeurat 분석을 진행할 수 있다.

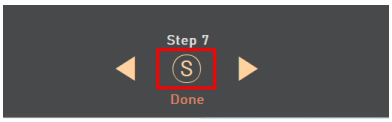

그림 4-2. 홈 화면 이동

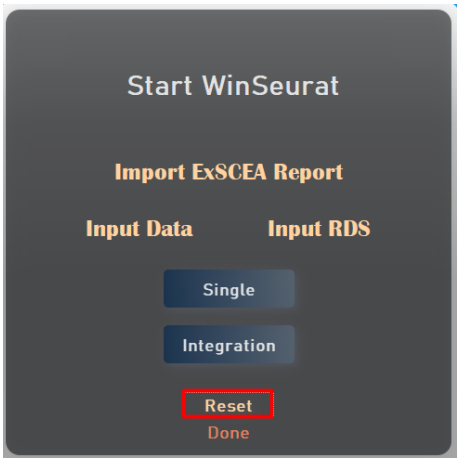

그림 4-3. Reset 기능