발 간 등 록 번 호
11-1480592-001568-01

최종보고서/ 2019.11

독도 자생식물 보전 및 관리를 위한 유전자 분석 연구 (5차년도)

2019

책임운영기관
환경부
국립생물자원관

제 출 문

국립생물자원관장 귀하

본 보고서를 “독도 자생식물 보전 및 관리를 위한 유전자 분석연구 5차년도)”의 최종보고서로 제출합니다.

2019년 11월

연구수행기간: 2019. 4. 8.~ 2019. 11. 30.

연구책임자: 나 경 주 (서울대학교 NICEM)

연 구 원: 이 정 호 (녹 색 식 물 연구소)

연 구 원: 임 종 성 (서울대학교 NICEM)

연 구 원: 이 준 기 (서울대학교 NICEM)

연 구 원: 엄 상 희 (서울대학교 NICEM)

연 구 원: 장 그 린 (서울대학교 NICEM)

연구보조원: 김 효 진 (서울대학교 NICEM)

연구보조원: 강 소 연 (서울대학교 NICEM)

요 약 문

1. 과제명: 독도 자생식물 보전 및 관리를 위한 유전자 분석연구 (5차년도)

2. 용역사업 배경 및 목적

▷ 배경

1) 나고야 의정서가 내포한 유전자원의 접근 및 이익공유에 따라 국가별 생물자원
주권주장을 위한 유전적 근거 확보가 필요함

2) 독도는 울릉도와 더불어 한반도 대륙과 격리되어 섬지역 고유의 식물자원을 다수 보유함

3) 특히 독도 주권에 대한 국민적 관심이 높아지는 시점에 독도 자생식물에 대한 유전자 증빙자료 확보는 매우 중요함

▷ 목적

독도 자생 관속식물인 섬초롱꽃 (*Campanula takesimana* Nakai) 1종의 핵유전체 염기서열 생산, 조립 및 유전자 지역 annotation 정보 확보

3. 과업의 범위

1) 독도 관속식물인 섬초롱꽃 (*Campanula takesimana* Nakai)의 표준 게놈지도 분석 수립

2) 대상종의 게놈 크기 측정

3) 분석 전략에 따라 표준 게놈지도 조립을 위한 염기서열 생산 및 조립

4) 섬생물 다양성 및 진화 분야 생물학적 난제 규명을 위한 향후 유전체 연구 방향 제시

4. 연구내용 및 방법

▷ 연구 내용

1) 대상종의 genome 크기 추정

대상종의 genome 크기 측정은 아래 2가지 방법을 이용하여 진행함

○ Flow-cytometry를 이용하여 세포내 핵을 염색하여 이미 genome 크기가 알려진 표준 식물종 (본 연구에서는 genome 크기 1.1Gb인 콩 (*Glycine max*))과 비교하여 genome 크기를 예측함

○ K-mer 분석을 통한 genome 크기 추정은 HiSeq short read data로 KmerGenie (http://kmergenie.bx.psu.edu/)를 사용하여 k-value 값을 선별한 후 genome 크기를 추정함

2) 독도 관속식물인 섬초롱꽃 1종의 표준 genome 데이터 생성

본 제안은 Illumina를 이용한 short read와 PacBio를 이용한 long read의 장점을 혼합하여 독도 섬초롱꽃의 표준유전체의 염기서열을 생성함

○ K-mer 분석 및 Scaffold용:

10XChromium 기반 HiSeq2500을 이용한 250bp PE sequencing data 생성

○ *De novo* assembly용:

PacBio 기반 장거리 (10kb library) 염기서열 data 생성

○ Transcript 기반 annotation용:

HiSeq 150bp PE 기반 RNAseq data 생성

3) 독도 관속식물인 섬초롱꽃 1종의 표준 genome 데이터 조립 및 분석

본 제안은 Illumina를 이용한 short read와 PacBio를 이용한 long read의 장점을 혼합하여 생성한 독도 섬초롱꽃의 표준유전체 데이터를 조립하여 유전자 및 repeat을 분석함

○ *de novo* assembly:

PacBio read로 염기서열 조립 후 contig 생성

○ Polishing:

PacBio로 생성된 contig의 error correction을 위해 PacBio raw read로 mapping하여 1차 correction을, HiSeq raw read로 mapping하여 2차 correction을 진행함

○ Scaffolding:

수백 kb~수 Mb 단위의 Linked-read contig 정보로 수천개 PacBio contig들의 order, orientation, continuity 등을 보정함

○ 최종 Assembly 점검:

Assembly 점검을 위해 BUSCO (https://busco.ezlab.org/)를 실행하여 Plant Nr Database와 90% 이상 match를 보이는지 점검함

○ Repeat 분석:

유전자 annotation을 위해 repeat masker로 repeat 지역을 cover하고 전체 genome 상에 존재하는 repeat 종류, 위치, 크기를 파악함

○ Structural annotation:

유전자 부위는 실제 발현되는 data인 RNAseq data와 유전자 발굴 알고리즘을 이용한 AUGUSTUS program을 모두 사용하여 분석함

○ Functional annotation:

Annotation 결과를 Plant DB에 match시켜 gene function을 예측하고 Gene Ontology 분석 및 KEGG 분석을 진행함

▷ 연구 방법

1) 1단계: 핵형분석

대상종의 유전체 크기는 유세포분석기(Flow-cytometry) 분석과 K-mer 분석의 두가지 방법을 사용하여 genome 크기 추정함

○ Flow-cytometry (CyFlow® Ploidy Analyser, Görlits, Germany)를 사용하여 본 과제 대상종인 식물 1종의 genome 크기를 측정함

○ K-mer 분석을 통한 genome 크기 추정은 Illumina short read data로 KmerGenie를 사용하여 k-value 값을 선별한 후 genome 크기를 추정함

2) 2단계: DNA 추출 및 QC

○ DNA purity:

Nanodrop으로 O.D. 260/280 ratio 1.8~2.0 확인

○ DNA quantity:

Picogreen으로 >100ng/ul 농도 점검, HiSeq(RNAseq 포함)용으로 Total 5ug, PacBio용으로 10ug이 요구됨

○ DNA integrity:

Gel electrophoresis로 DNA degradation 정도 점검, Bluepippin으로 10kb 이상 DNA fragment 선발 (PacBio용)

3) 3단계: PacBio sequencing 및 assembly, polishing, linked-read contig와 hybrid scaffolding 시행.

○ 10kb 이상의 DNA fragment를 생산 후 제조사의 manual에 따라 SMRT Bell shape library를 제작

○ Sequel platform에서 >60G data 생성

○ Read quality >Q80이상, read length >500bp를 filtering

○ 가장 긴 seed read에 subread를 mapping하여 corrected read 생산

○ Corrected read들로 *de novo* assembly 진행

○ 이후 assemble된 contig들의 염기서열 정확성을 위해 polishing 단계 진행. 1차 polishing은 PacBio Subreads를 다시 contig에 mapping하여 진행하고 2차 polishing은 HiSeq의 read를 mapping하여 진행함

4) 4단계: 10XGenomics linked-read sequencing 진행

○ Long range scaffold 제공을 위하여 10XGenomics 사의 linked-read sequencing을 추가로 진행함 (실험결과에 따라 Mb level의 scaffold 생성 가능)

○ ARCS program을 이용하여 PacBio contig와 hybrid scaffolding 진행

5) 5단계: 표준지도 품질 점검 방법 제시

○ PacBio사에서 제공하는 분석 툴을 (SMRT Link) 활용하여 PacBio raw data QC 수행 (Read quality 0.8, 500bp, 30Xcoverage >10kb)

○ 최종 염기서열 assembly 품질 평가 (Contig N50: >50kb, Contig 수: <10,000개, Scaffold N50: >200kb, Scaffold 수: <5,000개)

○ BUSCO를 활용하여 최종 염기서열 completeness 평가 (Complete BUSCO >80%)

6) 6단계: 유전자 해독을 위한 RNAseq 및 annotation

○ 대상종의 leaf, root, stem, flower, fruit 등 5개 tissue에서 추출한 RNA로 cDNA library

1개 제작

○ HiSeq을 이용하여 5G data 생성

○ Augustus 또는 Maker로 gene annotation시 Mapping 활용

7) 7단계: Annotation, Repeat masker, GO 분석, KEGG 분석 진행

○ 유전자 부위 및 반복서열 등 유전체 구조 판독

- 반복서열 (repeat) 지역은 RepeatMasker로 판독

- 반복서열 중 SSR 판독의 경우 misa.pl를 활용

- 유전자 부위는 RNAseq data 및 AUGUSTUS 등을 활용하여 판독

○ 유전자 기능 예측

- NCBI DB (ex. Nr, Plant Reference protein db)를 활용한 기능 주석

- Gene Ontology 및 KEGG 분석을 통하여 특정 pathway에 소속된 유전자 분류

- BLAST2GO, Interpro-scan을 활용하여 기존의 알려진 유전자 기능 및 protein motif 정보를 활용

5. 결과 및 고찰

▷ 결과

○ 독도 섬초롱꽃 게놈사이즈 측정 결과 Flow-cytometry로 약 0.97Gb로, K-mer 분석으로 약 0.84Gb로 추정되었고, 최종 assembly 결과 약 1.25Gb가 확보되었음

○ PacBio assembly 및 HiSeq polishing 후 약 8,800개의 contig를 확보하여 총 길이 약 1.25Gb, contig N50 230Kb, Maximum contig 1.6Mb 확보함

○ Linked-read seq으로 hybrid scaffolding 후 총 3,995 scaffold 확보, Scaffold N50 731kb, Maximum scaffold 4.3Mb 확보함

○ BUSCO로 근연종의 core gene으로 분석결과 84% 일치

○ Repeat 분석 결과 전체 repeat 중 LTR이 50%를 차지함

○ SSR은 총 196,182개 지역에서 발견되었고 이는 3,903개 scaffold에 해당

○ Annotation은 잎, 줄기, 뿌리, 꽃, 열매 RNA를 pooling한 RNAseq 기반 AUGUSTUS를 사용하여 총 168,217개를 얻었고 NCBI Plant reference DB 기반 functional annotation 수행 결과 135,438개로 1차 annotation 되었음

○ 현재 CD-hit으로 redundant한 sequence를 제거하고 coding region을 정확히 파악하기 위해 ORF identifying tool로 분석 중에 있음. 이 결과로 KEGG 및 GO 분석 진행

▷ 고찰 및 향후 연구방향 제시

1) 유전체 해독 의미 및 활용방안

○ 우리나라 고유식물인 독도 섬초롱꽃 유전체 정보의 보유는 유전자 자원의 보유라는 측면에서 국가의 주요 연구 자원으로서의 가치가 있음

○ 본 연구에서 발굴된 SSR (Simple Sequence Repeat)은 핵유전체의 마커로 향후 종간 및 종내 집단분석을 위한 마커개발의 토대가 됨

○ 본 연구에서 진행한 유전자 annotation 정보는 향후 섬초롱꽃과 식물의 고유 유전자 발굴 및 유용 물질 탐색의 주요 source가 됨

○ 본 연구에서 생산된 유전체 염기서열 해독은 향후 RNAseq을 통한 gene expression profiling 분석시 reference genome으로 사용 가능함

2) 유전체 연구방향 제시

○ 독도 섬초롱꽃의 genome 크기는 약 1.25Gb로 추정되며 이는 울릉도 섬초롱꽃 genome의 잠정 크기인 1.4G (flow-cytometry 결과) 및 도라지 genome 크기인 700Mb와 차이가 나는 것으로 진화과정 중 초롱꽃과 (Campanulaceae)에서 genome size variation이 존재함이 관측됨. 이러한 genome size variation의 기원으로 배수성 또는 repeat 지역에 의한 크기 차이가 예상됨

○ 본 과제의 결과를 기반으로 향후 한반도, 울릉도, 독도의 초롱꽃과 섬초롱꽃의 comparative genomics를 활용한 유전체 비교연구 분석으로 진화 연구의 토대가 됨

○ 향후 Optical Mapping 기술을 활용하여 독도 섬초롱꽃 genome 3,995개의 scaffold를 수백~수십개에 이르는 scaffold로 축소시켜 genomic rearrangement 및 ploidy event를 연구하는데에 활용 가능함

목 차

표 목 차

그 림 목 차

Ⅰ. 연 구 개 요

Ⅰ. 연구개요

1. 연구 목적과 배경

○ 독도와 울릉도는 우리나라 유일의 대양섬으로 한반도 대륙과 격리되어 섬지역 고유의 식물자원을 다수 보유함

○ 독도는 울릉도에서 약 92km 떨어져 있으며, 독도천연구역(천연기념물 제336호)으로 지정하여 보호하고 있음

○ 특히 독도 주권에 대한 국민적 관심이 높아지는 시점으로 독도 자생식물에 대한 유전자 증빙자료 확보하여 독도 생물의 유전적 기초 자료 확보를 통해 생물주권 확립의 토대 마련하는 데 있음

○ 독도식물목록(국립생물자원관 2015)에는 60종의 관속식물(고사리류 1종, 현화식물 59종)이 알려져 있으나, 핵 유전체에 대한 정보는 없음

○ 본 연구는 생물주권을 위해 독도에 자생하는 **섬초롱꽃 (*Campanula takesimana* Nakai) 1종의 유전체 정보를 확보**할 목적으로 제안된 연구임

※본 연구는 유전체 1G 크기의 homozygous한 유전체를 기준으로 제안되었으며, 실제 발주기관에서 제공한 식물의 유전체의 특성이 상이할 경우, 연구방법을 수정할 수 있고, 이 경우에는 감독관과 협의하여 결정하는 것으로 했음. (과제수행 계획서 참조)

2. 섬초롱꽃과 근연식물의 유전체 정보

○ 섬초롱꽃(*Campanula takesimana* Nakai)은 1922년에 Nakai에 의해, 울릉도에서 발견한 식물로 울릉도(takesima:竹島)초롱꽃이란 뜻이며, 최근에 독도에서도 발견된 관속식물임.

○ 섬초롱꽃은 낮은 숲 속 사면저지대와 산지의 비탈진 풀밭에 나는 여러해살이풀로서 초롱꽃과 유사하나 잎에 광택이 있고 꽃받침의 맥이 뚜렷하여 구별됨. (그림 1)

○ 섬초롱꽃의 형태학적 정보는 아래와 같음.

① 키: 높이 30~155cm 정도로 자람

② 뿌리: 주근으로 길게 신장함

③ 줄기: 곧추서며 두껍고 단생하며 분지하지 않고 자주색임. 줄기잎은 어긋나며 달걀 모양 삼각형 또는 심장형으로 광택이 있고 흰색 털이 있음.

④ 잎은 길이 2.6~11.9cm, 너비 1.2~7.9cm임. 잎끝은 뾰족하고 밑부분은 심장 모양이며 잎 가장자리는 예리한 톱니가 있음

⑤ 꽃: 꽃차례는 총상꽃차례이며 꽃자루는 길이 2.0~11.7cm로 털이 약간 있음. 꽃은 길이 2.3~4.0cm로 연한 자색의 꽃부리에 검은색 반점이 있으며 꽃부리의 끝은 5개로 갈라지며 넓은 삼각형임. 꽃은 6~8월에 피는데 꽃부리의 색깔은 많은 변이를 보임.

⑥ 열매는 삭과로 회갈색의 타원형임.

○ 섬초롱꽃(*Campanula takesimana* Nakai)은 초롱꽃(*Campanula punctata*)와 근연종(박선주 외 2015)으로 알려져 있으며, 학자에 따라 *Campanula punctata* var. *takesimana* 로 분류되기도 함.

그림 1. 섬초롱꽃(*Campanula takesimana* Nakai)
[국립생물자원관 포탈]

○ 섬초롱꽃 염색체 수에 대한 한국연구자료는 없고, 중국에서 초롱꽃(*Campanula punctata*)이 2N=34로 보고되어 있음 (FOC V19). 같은 과의 도라지(*Platycodon grandiflorum*)가 2n = 18, 36으로 배수체가 존재하며, 더덕(*Codonopsis lanceolata*)과 만삼 党参(*Codonopsis pilosula*)도 2n = 16임.

○ Kew Botanical Garden의 Plant DNA C-values를 찾아보면 초롱꽃과의 22개 식물에 대한 정보가 있으며, 1C=0.33~(1.88)~6.38 pg이며, *Campanula*속에서는 9종의 10개의 식물에 대한 정보가 있으며, C=1.98~(1.98)~3.98 pg임. 따라서 *Campanula* 종내에서 배수체가 존재함을 나타냄.

○ NCBI에 등록된 초롱꽃과 식물의 유전체는 중국 Yunnan 농과대학에서 연구하는 더덕의 근연인 만삼(*Codonopsis pilosula,* 党参 2n = 16)이 937.71 Mb로 등록되어 있고, 한국의 농과원의 유전체과에서 다부처유전체 연구로 2017년에 완료된 장백도라지(*Platycodon grandiflorum*: 2n = 18)는 이배체로 염색체를 확인했고, 700Mb로 등록될 예정임(Kim *et al.* unpublished). 이 유전체는 4년에 걸쳐서 1,663 scaffolds로 조립하였으며 비교 대상 식물로 초롱꽃과가 포함되는 국화목(Asterales)에 해바라기와

상치를 기준으로 하는 포함하는 10여개 식물이 등록되어 있음 (그림 2).

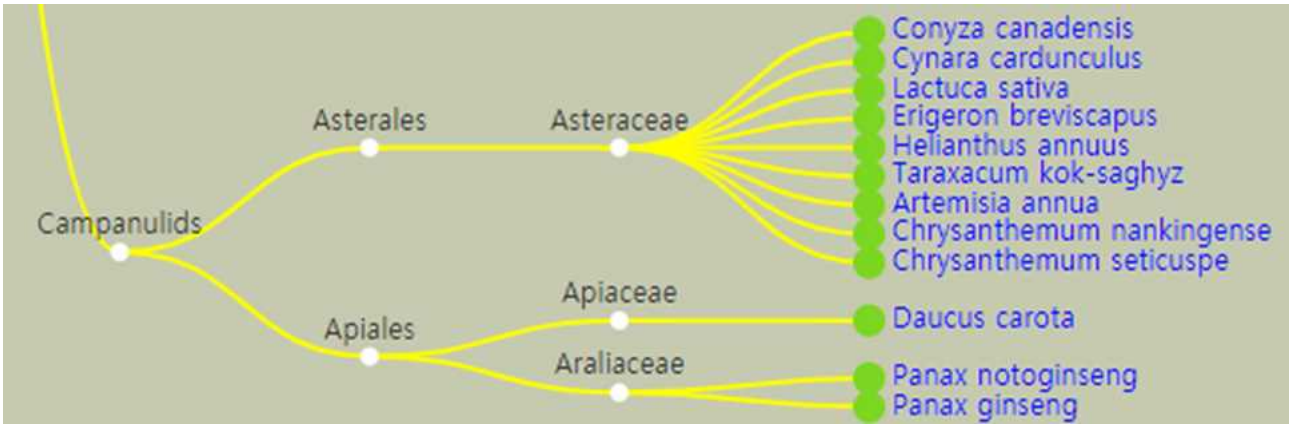

그림 2. 초롱꽃이 속하는 국화목(Astrales)과 근연 식물군의 식물 유전체 계통도 (https://www.plabipd.de/plant_genomes_pa.ep)

○ 그림 2에서 보고된 표준 유전체는 재정적 지원이 원활한 주요 작물 또는 약용식물로써 염색체 연구와 BAC-end seq등 고전 연구와 연계된 연구가 많으며, 보조적으로 다양한 종류의 library를 이용한 다량의 HiSeq data를 이용하거나, 다량의 PacBio data를 이용하였음.

○ 본 연구에서는, 표준유전체 정보가 없는 초롱꽃과 식물인, 독도에 자생하는 섬초롱꽃의 유전체를 제한된 예산으로 짧은 기간에PacBio와 HiSeq의 NGS data의 Combined Analysis를 통한 고품질 표준 유전체 지도확립을 위해 설계됨.

3. 연구범위, 체계 및 진행

○ 본 연구는 독도에 자생하는 섬초롱꽃의 표준게놈지도를 만들고, 섬생물 다양성 및 진화 분야 생물학적 난제 규명을 위한 향후 유전체 연구 방향 제시함.

○ 본 연구를 위하여 주관부서인 국립생물자원관에서 독도 자생식물인 섬초롱꽃 1종의 생체를 본 과제의 수행기관인 서울대 NICEM 유전체분석센터에 제공하여 유전체염기서열 생산 및 유전자 분석을 진행하며 본 과제 수행동안 자문위원회 및 감독관으로부터 과제 진행 및 결과에 대한 자문 및 점검을 받음 (그림 3).

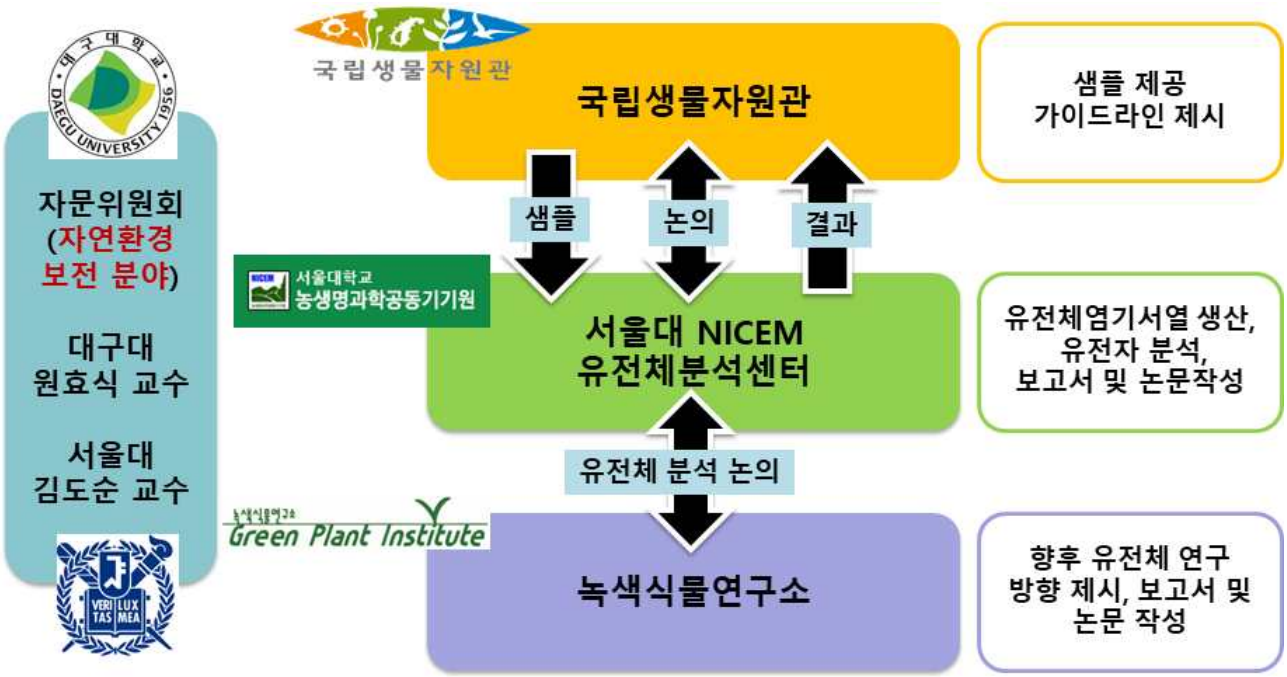

그림 3. 섬초롱꽃 유전체 연구 수행체계

○ 본 연구는 독도에 자생하는 섬초롱꽃의 유전체를 PacBio와 HiSeq의 NGS data를 이용하여 10,000 이하의 고품질 Scaffold를 조립하여, 표준 게놈지도를 만들도록 계획됨.

○ 이 목적을 위하여, Flow-Cytometry를 이용한 유전체 크기 추정하여, NGS

(HiSeq) data 생산 계획을 세우고, HiSeq data를 이용한 K-mer 분석을 통해 유전체 크기를 측정하고, PacBio와 HiSeq의 NGS data를 이용하여 Scaffold를 조립하여 유전체 크기를 측정함 (그림 4).

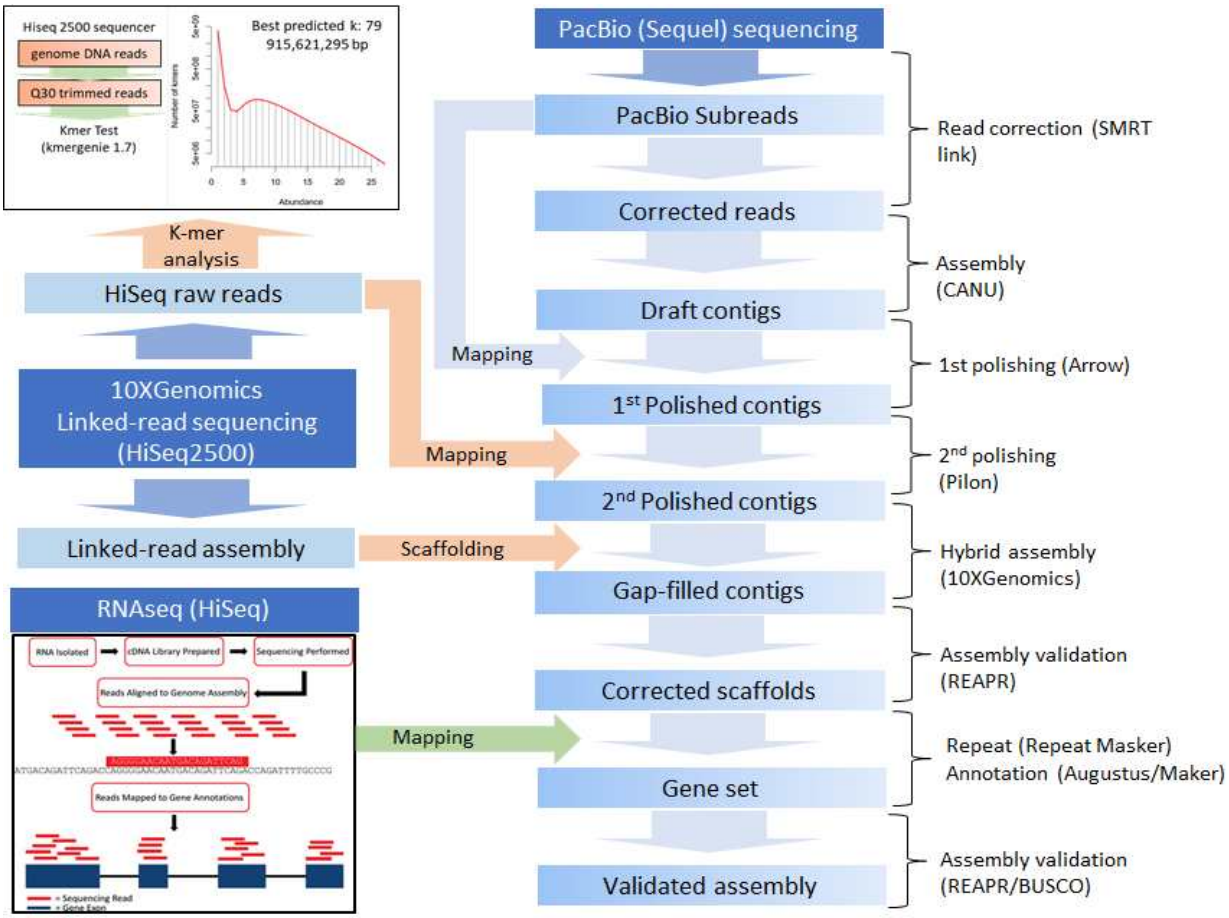

그림 4. 섬초롱꽃 유전체 분석 연구계획 모식도

○ 본 연구팀은 총 8명이며, NICEM 유전체분석팀의 7인과 녹색식물연구소 1인으로 구성되어 있음 (그림 5).

○ 유전체 분석은 NICEM의 전문인력이 투입되었으며, Biogreen 유전체사업단의 평가위원장을 6년간 수행경험이 있는 녹색식물연구소의 이정호 소장을 유전체 정보의 활용 및 분류/진화 전문가로 연구진에 포함하였음.

○ 유전체 분석을 위하여 NICEM에서는 실험팀, NGS(Next Generation Sequencing) 분

석팀과 정보분석팀으로 이루어져 있으며, '섬생물 다양성의 진화 양상 규명을 위한 향후 유전체 연구 전략 제시 및 연구결과의 활용' 에 대하여 연구총괄-녹색연-정보분석팀의 연계연구로 계획되었음.

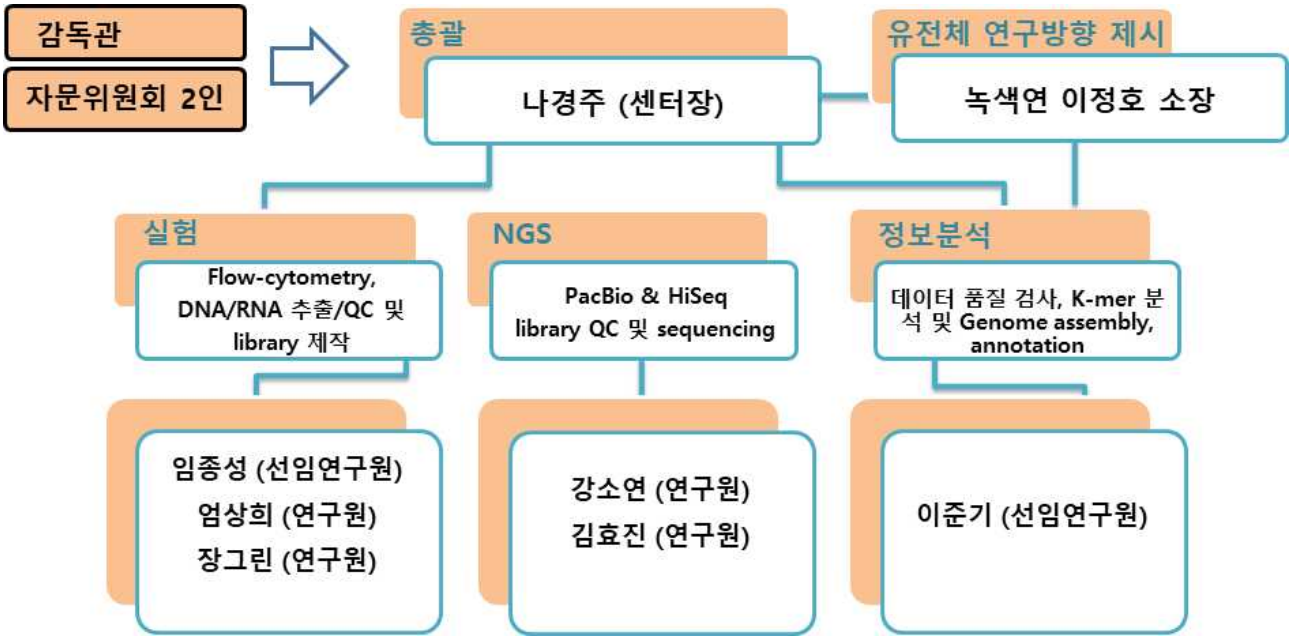

그림 5. 섬초롱꽃 유전체 연구 과제 참여 인력 및 업무 분장

○ 유전체 분석을 수행계획 일정표 (그림 6)에서 RNA Seq이 3달 늦게 되었고, 이에 따른 Annotation이 지연된 것을 제외하고 계획대로 진행되었음. 붉은색 box는 실제 진행된 일정을 보여 줌.

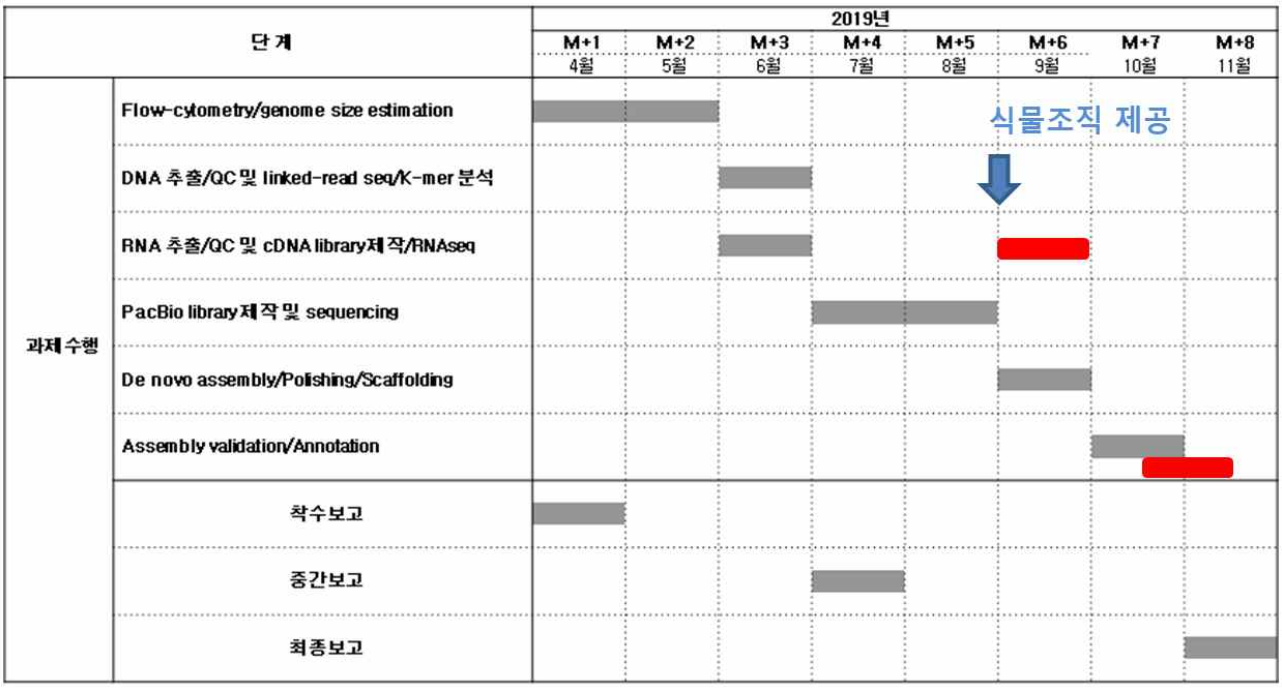

그림 6. 섬초롱꽃 유전체 연구 과제 수행계획 및 실제 수행 일정표

Ⅱ. 연 구 방 법

Ⅱ. 연구방법

1. 연구재료 확보

○ 독도 섬초롱꽃 식물체는 2019년 5월 21일에 제공되었고, 이 재료는 Flow-Cytometry 분석과 DNA 분석(HiSeq과 PacBio)에 사용되었음. RNA분석을 위해 조직별 시료가 필요하였지만, 제공된 식물체가 꽃을 피우지 않아, 잎과 뿌리의 RNA를 추출하여 사용하였음. 수차례 독도채집을 시도하였으나 잦은 기후변화로 용이하지 않았고, 결국 2019년 8월 1일에 독도 채집을 성공하였으나, 현지에서 꽃이 시든 상황이어서 이 식물체는 표본으로 만들어졌음. 이러한 사유에서 줄기, 꽃, 열매 등의 RNA 시료는 8월 말에 국립생물자원관 온실에서 자라는 울릉도 섬초롱꽃에서 채취해서 사용하였음. Voucher정보는 표 1과 같음.

표 1. 본 연구에서 사용된 식물재료

	채집번호	Voucher No.	GPI Plant/DNA No.	Site
독도 섬초롱꽃	Lim&Lee Dokdo3 (2019. 8. 1.)	NIBRVP0000752857 (KB)	GPI2019_CAMP12D	경상북도 울릉군 독도 서도 야생
울릉도 섬초롱꽃	Lim190725	NIBRVP0000752859 (KB)	GPI2019_CAMP12A	자원관 온실 재배

그림 7. 본 연구에서 사용된 독도 섬초롱꽃(CAMP12D)과 울릉도 섬초롱꽃(CAMP12A)

2. 유전체 크기 측정

가. Flow-cytometry를 이용한 유전체 크기 측정방법

1) 장비 및 시약

○ 장비: CyFlow Cube 6 (Sysmex Partex) (그림 8)

○ 시약: CyStain UV Precise P (Sysmex Partex)

○ 소모품: 50 ㎛ CellTrics disposable filter (Sysmex Partex)

2) 실험방법

○ CyStain UV Precise P는 식물 조직으로부터 핵 추출과 DNA 염색을 할 때 필요한 시약으로서 Extraction buffer 125 ml와 Staining buffer 500 ml로 구성되어 있으며, flow cytometer의 420 nm UV 형광에 의해 excitation되어 435 nm에서 500 nm의 형광 emission 파장을 내보내는 방식으로 진행됨.

○ 약 0.5 cm^2 정도나 그 이하 크기의 잎 조직을 55 mm 페트리디시에 넣은 후 400 ㎕의 Extraction buffer를 넣어준 후 날카로운 면도날을 이용하여 30초에서 60초 정도동안 샘플을 자름. 면도날은 5개에서 10개 정도 샘플에 사용한 후에는 무뎌지므로 새것으로 교체해 주어야 함.

○ 샘플을 Extraction buffer에서 자른 후 30초 기다린 후, 샘플을 50 ㎛ CellTrics disposable filter 에 통과시켜 3.5 ml 시험관에 담음.

○ 여기에 1.6 ml의 Staining buffer를 넣어주고 30초 기다린 후, flow cytometer (CyFlow Cube 6)에 넣어 파랑색 형광 채널에서 분석함 (그림 9).

그림 8. Flow-cytometry 장비 사진

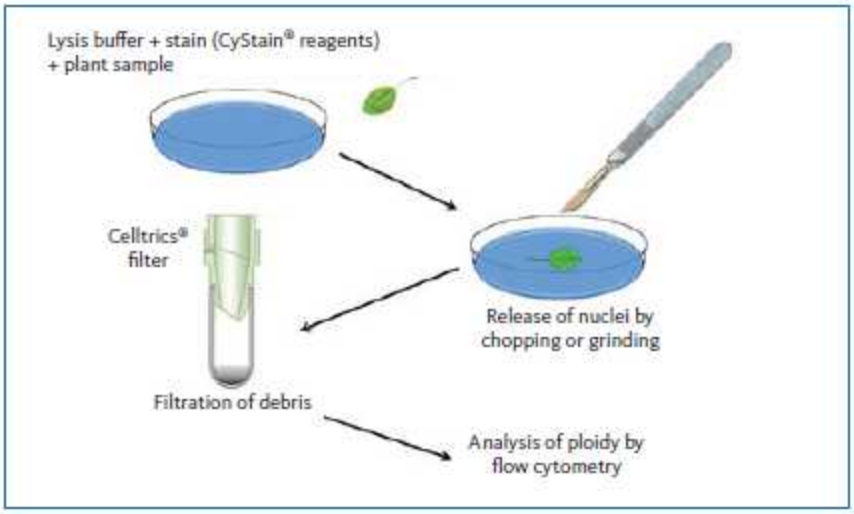

그림 9. Flow-cytometry를 위한 실험 절차

나. K-mer 분석을 이용한 유전체 크기 측정방법

○ 본 연구에서는 HiSeq(151PE) data를 활용하여 k-mer frequency plot을 얻어 유전체의 사이즈를 예측하였으며, kmergenie (ver. 1.7048, http://kmergenie.bx.psu.edu/.) 프로그램을 활용함 (Chikhi R. *et al.* 2014).

○ Kmergenie는 자동으로 kmer plot을 그리고 유전체 사이즈를 예측해주는 프로그램으로 본 분석에서는 최소 10mer에서 최대 70mer로 kmer를 설정하고 분석을 진행함.

3. 유전체분석을 위한 NGS data 생산 방법

가. DNA 및 RNA 추출

1) DNA 추출 및 QC

가) 장비 및 시약

○ 장비: Micro17TR 원심분리기(Hanil Scientific Inc., Korea), Victor3™ Plate Reader (PerkinElmer, U.S.A.)

○ 시약: 2X Extraction buffer (100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2% (w/v) CTAB , 20 μl/ml β-mercaptoethanol), Precipitation buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% (w/v) CTAB), 1.5 M NaCl, chloroform:isoamyl alcohol (24:1) (Sigma-Aldrich, U.S.A.), RNaseA

(MGmed, Korea)

○ 키트/소모품: Quant-iTTM PicoGreenTM dsDNA Assay Kit (ThermoFisher Scientific. U.S.A.)

나) 실험방법

○ 독도 섬초롱꽃의 genomic DNA (gDNA)는 CTAB method를 이용하여 추출하였음.

○ 실험 시작하기 전 2X Extraction buffer는 65℃ water bath에 pre-warming 시켜둠.

○ 독도 섬초롱꽃 꽃 600 mg을 액체질소와 함께 막자에서 이용하여 곱게 파쇄함.

○ 곱게 간 샘플을 50 ml tube에 옮겨담은 후 2X extraction buffer를 3.5 ml 넣고 여러 번 inverting 후 실온에서 2시간 incubation하는데, 중간중간에 inverting을 시행함.

○ Incubation이 끝난 후 2ml tube에 700 ㎕씩 분주 후 각각의 분주된 tube에 4 ℃의 chloroform:isoamyl alcohol (24:1)을 700 ㎕씩 넣고 여러 번 inverting후 4 ℃ 에서 6000 rpm으로 10 분 원심분리를 해 줌.

○ 원심분리가 끝난 후 상층액을 350 ㎕씩 조심히 1.5 ml tube에 옮긴 후 900㎕의 Precipitation buffer를 넣고 총 10개의 tube가 준비하여 이들을 실온에서 5시간 동안 반응시킴.

○ 반응이 끝난 후 4℃에서 6000 rpm으로 15분 동안 원심분리시키고, pipet을 이용해 pellet을 제외한 용액을 제거함.

○ 펠렛에 1.5 M NaCl 600 ㎕와 10 mg/ml RNaseA 6 ㎕를 첨가한 후 pipet을 이용하여 천천히 pellet을 잘 녹여줌.

○ 실온에서 1시간 동안 incubation한 후, 4 ℃의 chloroform:isoamyl alcohol (24:1)을 600 ㎕를 넣고 여러 번 inverting 함.

○ 4℃에서 6000 rpm으로 5분 원심분리를 한 후, 상등액 450㎕를 새로운 1.5ml tube에 조심히 옮겨줌.

○ 여기에 -20 ℃의 100% 에탄올을 900 ㎕씩 넣고 여러 번 inverting 해준 후 -20℃에서 overnight incubation 함.

○ 다음날 13000 rpm에서 15분 원심분리한 후 pipet을 이용해 pellet을 제외하고 모두 제거해 줌.

○ 80% 에탄올을 1 ml씩 넣고 천천히 여러 번 invert 후 4℃에서 13,000

rpm으로 10분동안 원심분리를 해 줌.

○ Pipet으로 상등액을 모두 제거한 후 pellet을 10분간 건조시키고, 증류수로 10개 tube 합이 100 ㎕가 되도록 gDNA를 녹여 합쳐 줌.

○ 추출된 gDNA는 agarose gel에서 integrity를 확인하며, dsDNA만을 측정할 수 있는 Quant-iT™ PicoGreen™ dsDNA Assay Kit의 picogreen 형광 기법을 이용하여 picogreen과 결합한 gDNA의 농도를 plate reader를 통해 측정한 후 박테리오파지 람다 DNA (100 ㎍/ml)를 1/50로 희석한 후 최종 200 ㎕의 부피로 0~1㎍/ml 사이의 농도를 5개 이상 만들고 이들을 picogreen과 섞고 2-5분 반응 후 plate reader에서 480 nm로 읽어 control curve를 만듬. 샘플을 picogreen과 섞어 형광 값을 측정한 값을 standard curve에 대입하여 샘플의 농도를 알아냄.

2) RNA 추출 및 QC

가) 장비 및 시약

○ 장비: LabChipGX (Caliper LifeSciences, U.S.A.), Centrifuge 5424 (Eppendorf, Germany)

○ 키트/소모품: Hybrid-RT™ kit (GeneAll, Korea), RNase-Free DNase Set (QIAGEN, Germany) RNeasy MinElute Cleanup kit (QIAGEN, Germany), HT RNA Labchip Kit (Caliper LifeSciences, U.S.A.), HT RNA Reagent Kit (PerkinElmer, U.S.A.)

나) 실험방법

○ 독도 섬초롱꽃의 잎과 뿌리, 울릉도 섬초롱꽃의 줄기, 꽃 그리고 열매에서 Hybrid-RT™ kit을 이용하여 total RNA를 추출하였음. 아래에 나오는 시약들과 mini column type F는 이 키트에 속함.

○ 각 샘플 100 mg을 액체질소와 막자사발을 이용하여 곱게 파쇄함. 곱게 갈린 sample에 RiboEx™ 을 1 ml 넣고 잘 풀어준 후 실온에서 5분동안 incubation함.

○ 4℃에서 12,000 g로 10분간 원심분리한 후 침천물이 딸려오지 않게 조심하면서 상층액만 새로운 1.5 ml tube로 옮김.

○ 1 ml의 RiboEx™ 당 chloroform 200 ㎕를 넣고 15초동안 여러 번 inverting해 줌.

○ 2분 동안 방치한 후, 4℃에서 12,000 g로 15분 원심분리함.

○ 상층액 400 μl를 새로운 1.5 ml tube에 옮겨주고 동량(400 μl)의 Buffer RB1을 첨가한 후 여러 번 inverting 해 줌.

○ 이 중 700 μl를 mini column type F에 옮겨 상온에서 10,000 g로 30초 원심분리 해 줌.

○ 컬럼 워싱을 위해, 500 μl Buffer SW1을 mini column type F에 넣고 상온에서 10,000 g로 30초 원심분리한 후, 이어서 500 μl Buffer RNW를 mini column type F에 넣고 역시 상온에서 10,000 g로 30초 원심분리 해 줌.

○ 남아 있는 washing buffer 제거를 위해 상온에서 10,000 g로 1분 원심분리 함.

○ Column을 새 1.5ml 튜브로 옮긴 후 50 μl nuclease-free water를 mini column membrane에 넣고 1 min 기다림. 상온에서 10,000 g로 1분 원심분리 하여 total RNA를 elution함.

○ 추출한 total RNA에서 DNA를 제거하는 과정은 RNase-Free DNase Set과 RNeasy MinElute Cleanup kit을 사용하였는데, 최대 45 μg의 RNA까지 처리할 수 있음. 아래에 나오는 시약들과 RNeasy MinElute Spin Column은 이 키트에 속함.

○ 먼저 87.5 μl 이하의 RNA solution에 10 μl Buffer RDD와 2.5 μl DNase I stock solution을 혼합하고 RNase-free water를 이용하여 100 μl가 되도록 맞춰줌.

○ 상온에서 10분 incubation한 후, 350 μl Buffer RLT를 넣고 잘 섞어줌. 여기에 250 μl의 100% 에탄올을 넣고 pipetting으로 잘 섞어줌.

○ 샘플 700 μl를 RNeasy MinElute Spin Column에 옮긴 후 상온에서 8,000 g로 15초 원심분리를 시행함.

○ 워싱을 위해, 용액 제거 후 500 μl Buffer RPE을 넣고 상온에서 8,000 g로 15초 원심분리하고, 이어서 500 μl의 80% 에탄올을 넣어 상온에서 8,000 g로 2분 원심분리를 시행함.

○ 남아 있는 washing buffer 제거를 위해 column 뚜껑을 열고 상온에서 최대 속도로 5분 동안 원심분리를 시행함.

○ Column을 새 1.5ml tube로 옮긴 후 25 μl nuclease-free water를 column의 membrane에 정확히 떨어뜨려 줌. 1분 기다린 후 상온에서 최대 속도로 1분 원심분리하여 DNA가 제거된 total RNA를 elution함.

○ 추출된 RNA의 integrity와 농도는 HT RNA Labchip Kit와 LabChipGX를

시행함.

나. HiSeq 기반 data 생성방법

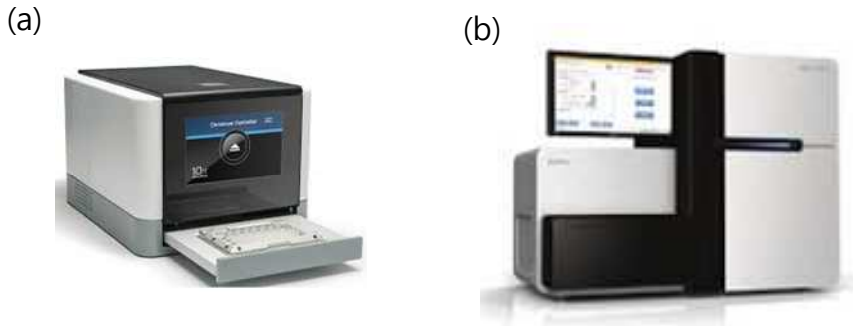

그림 10. Linked-read sequencing에 사용된 장비
(a) 10XChromium (b) Illumina HiSeq2500

1) 10X Genomics library 제작 및 sequencing

가) 장비 및 시약

○ 장비: 10x Chromium (10x genomics, U.S.A.) (그림 10a), BluePippin (Sage Science, U.S.A.), Qubit Fluorometer (ThermoFisher Scientific. U.S.A.), Applied Biosystems Veriti Thermal Cycler (ThermoFisher Scientific. U.S.A.), Agilent Bioanalyzer (Agilent, U.S.A.), Pippin Pulse (Sage Science, U.S.A.), LabChipGX (Caliper LifeSciences, U.S.A.)

○ 시약: Dynabeads MyOne Silane beads (ThermoFisher Scientific. U.S.A.), SPRIselect Reagent (Beckman Coulter, U.S.A.), High Sensitivity DNA Reagents (Agilent, U.S.A.)

○ 키트/소모품: ChromiumTM Genome Library Kit & Gel Bead Kit v2 (10x genomics, U.S.A.), ChromiumTM i7 Multiplex Kit (10x genomics, U.S.A.), Chromium Genome Chip kit v2 (10x genomics, U.S.A.), High Sensitivity DNA chip (Agilent, U.S.A.), Agilent DNA 1000 Kit (Agilent, U.S.A.), HT DNA High Sensitivity LabChip Kit (Caliper LifeSciences, U.S.A.)

나) 실험방법

○ 10x chromium로 라이브러리를 제작하는 경우, DNA quality가 sequencing quality에 매우 큰 영향을 미치기 때문에 샘플에 맞는 DNA 추출 방법을 선택하여 큰 사이즈의 DNA가 intact하게 존재하도록 샘플을 준비하여야

함. 10x의 경우 library QC로 sequencing data quality를 파악할 수 없기 때문에 DNA를 최상으로 준비하는 것이 매우 중요함.

○ Size-selection: 40kb 이상의 gDNA가 확인되면 40kb 아래 사이즈를 제거하고 큰 사이즈만 얻기 위하여 BluePippin으로 size-selection을 수행함.

○ Qubit을 이용하여 40kb 이상만 추출된 gDNA DNA 농도를 3반복 측정하고, 측정값의 평균값으로부터 샘플 농도가 1 ng/㎕가 되는 희석값을 계산함.

○ Dilution of DNA: 1 ng/㎕이 20 ㎕ 이상의 볼륨이 되도록 샘플을 희석함 이때 희석은 EB buffer를 사용하며 mixing 할 때에는 wide-bore tips 으로 pipeting 해줌.

○ Qubit을 이용하여 희석된 DNA를 3반복으로 측정하고, 값이 0.8-1.2 ng/㎕ 범위 내에 들면 다음단계로 넘어가고, 편차가 큰 경우에는 다시 희석해서 3반복값이 0.8-1.2 ng/㎕ 내에 들 때까지 농도값을 측정함.

○ ChromiumTM Genome Library Kit & Gel Bead Kit v2의 프로토콜에 따라 denaturing한 gDNA를 Master Mix에 섞은 후, 이 용액과 Gel bead 그리고 Partitioning Oil을 순차적으로 Chromium Genome Chip kit v2에 들어 있는 microfluidics chip에 분주함 (그림 11).

그림 11. Microfluidics chip 및 샘플 로딩 위치와 순서

○ Chromium Genome Chip kit v2에 포함되어 있는 가스켓을 칩 위에 장착한 후, 가스켓이 장착된 칩을 10x Chromium의 트레이에 넣고 20분간 기기를 작동시켜 Gel Bead와 gDNA가 들어간 GEMs (Gel Bead-In EMulsions)를 만듬 (그림 12).

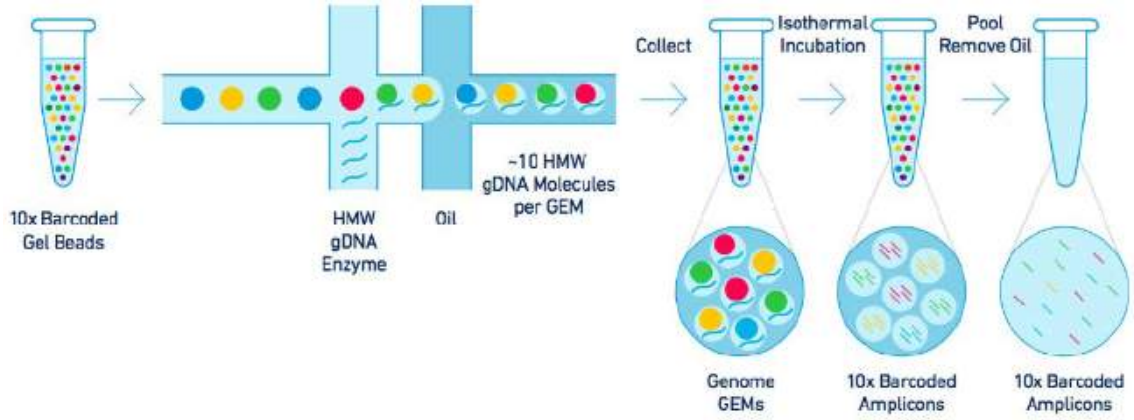

그림 12. 10X genome assay workflow

○ 10x Chromium의 작동이 끝난 후 125 ㎕를 회수함. GEMs가 제대로 만들어지면 희뿌연 불투명 용액임.

○ Thermal cycler를 사용하여, 30 ℃에서 3시간 반응한 후 65 ℃ 10분 반응시키고 이 단계에서 random hexamer가 genomic DNA에 붙어 합성함.

○ Post GEM Incubation Cleanup 단계를 들어가기 전에 Dynabeads를 상온에 두고, Buffer Sample cleanup 1 시약의 경우 65 ℃에 10-15분 정도 두어 완전히 녹게 함.

○ PCR plate의 Foil seal을 제거하고 125 ㎕ Recovery Agent를 추가하여 피펫팅으로 섞어주고, 각 tube에서 250 ㎕를 일반 8-tube stip에 옮기고 뚜껑을 닫아 15초 정도 voltexing 해 줌.

○ spin down 시켜주면 아래의 Recovery Agent/Partitioning Oil (pink)과 상층액 (clear)으로 층이 분리되는데, 상층액만 남기기 위해 135 ㎕ 의 Recovery Agent/Partitioning Oil (pink) 을 제거함. 이 과정에서 약간의 pink 층이 바닥에 남을 수 있음.

○ Dynabeads Cleanup Mix를 프로토콜에 따라 만들어준 후 바로 150 ㎕를 각 tube에 분주해주고 pipeting으로 섞은 후 상온에서 10분 방치함.

○ Tube를 10xTM Magnetic Separator의 High position에 두고 용액이 맑아질 때까지 2분 이상 기다린 후 상층액을 제거함.

○ 250㎕ 80% 에탄올을 넣고 섞어준 후 위와 같은 방법으로 상층액을 제거 후 200㎕ 80% 에탄올로 이 단계를 반복함.

○ Washing이 끝난 후 spin-down한 tube strip을 10xTM Magnetic Separator의 Low position으로 옮기고, 용액이 맑아질 때까지 기다림.

○ 잔여 에탄올 제거 후, 10xTM Magnetic Separator으로부터 tube strip을 다

른 랙으로 옮긴 후, 바로 미리 프로토콜에 따라 만들어둔 Elution Solution I을 51 ml 넣어줌.

○ 30초 후 pipeting으로 섞어주고, 상온에서 5분 방치함. Spin-down한 tube strip을 10xTM Magnetic Separator의 Low position으로 옮기고 용액이 맑아진 후 DNA가 elution된 상등액 50㎕를 새 tube strip으로 옮겨줌.

○ 이렇게 DNA가 회수된 후 아래와 같이 150-800 bp 크기의 DNA만 선택적으로 회수할 수 있는 SPRIselect 단계를 거침.

○ SPRI select Reagent를 vortexing한 후 35㎕ (0.7X)를 DNA가 회수된 각 tube에 넣음.

○ 상온에서 5분 반응 후 tube strip을 10xTM Magnetic Separator의 High position에 두고 2분 이상 용액이 맑아질 때까지 기다림.

○ 상층액을 제거하고 80% 에탄올을 125 ㎕ 이용하여 위에서 언급된 것과 같은 방법으로 총 3번 washing해 줌.

○ Tube strip을 10xTM Magnetic Separator의 Low position으로 옮기고 남아 있는 에탄올을 제거한 후, tube strip을 다른 랙으로 옮겨준 후 바로 프로토콜에 따라 미리 만들어둔 52.5 ㎕의 Elution Solution II를 넣어줌.

○ Pipeting으로 섞어주고, 상온에서 5분 방치함. Spin-down한 tube strip을 10xTM Magnetic Separator의 Low position으로 옮기고 용액이 맑아진 후 short length DNA가 elution된 상등액 52 ㎕를 새 tube strip으로 옮김.

○ 1㎕의 샘플을 Agilent Bioanalyzer의 High Sensitivity DNA chip에 loading하여 yield와 fragment size를 살펴본 후 농도는 Qubit으로 측정함.

○ Library construction: P5, P7 프라이머 및 인덱스를 삽입시키고 라이브러리 형태로 제작하는 단계로 Chromium Genome Library kit v2의 프로토콜에 따라 아래와 같은 흐름으로 제작됨 (그림 13). 이 과정에서 index를 부여하기 위해서는 Chromiu i7 Multiplex Kit을 사용함.

그림 13. Chromium Genome Library 각 fragment의 최종 구조

○ 라이브러리 제작 후 농도 및 사이즈는 일반 일루미나 라이브러리 QC와 마찬가지로 농도는 Qubit으로, 크기 분포는 LabChipGX로 측정함.

○ Sequencing은 HiSeq2500으로 진행함.

2) RNA-Seq library 제작 및 sequencing

가) 장비 및 시약

○ 장비: LabChipGX (Caliper LifeSciences, U.S.A.), Qubit Fluorometer (ThermoFisher Scientific. U.S.A.), Centrifuge 5424 (Eppendorf, Germany)

○ 키트/소모품: NuGEN Universal Plus mRNA-Seq kit (Tecan Genomics, Inc., U.S.A.), HT DNA High Sensitivity LabChip Kit (Caliper LifeSciences, U.S.A.), Sensitivity Reagent Kit (PerkinElmer, U.S.A.)

○ RNA-Seq library를 제작에는 NuGEN Universal Plus mRNA-Seq kit을 이용하였음 (그림 14).

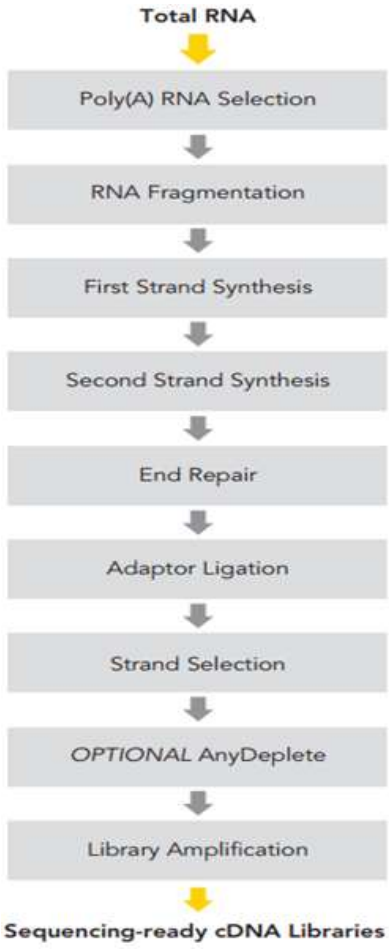

그림 14. RNA-Seq library 제작 모식도

나) 실험방법

○ 울릉도 섬초롱꽃의 꽃, 줄기, 열매는 각각 400 ng으로 pooling하여 total 1.2 μg을 사용하였으며, 독도 섬초롱꽃의 뿌리, 잎은 각 600 ng으로 pooling하여 total 1.2 μg을 library 제작에 사용하였음.

○ Total RNA로부터 mRNA만 추출: Total RNA 50 μl (1.2 μg)을 60 μl의 Oligo dT Bead Master Mix 와 섞어 65 ℃에서 5분간 반응시켜 RNA 이차 구조를 편 후, 상온에서 중간에 pipeting으로 섞어주며 총 10분간 반응시켜 beads의 polyT에 mRNA의 polyA가 붙도록 함. 이후 프로토콜에 따라 magnet을 이용해 여러 번의 wahsing을 해 주고 마지막 washing buffer가 tube에 담겨 있는 채로 바로 다음 단계로 들어감.

○ RNA Fragmentation: magnet에 튜브가 올려져 있는 채로 washing buffer를 제거 하고, tube를 새로운 랙에 옮긴 다음, 1X Fragmentation buffer 20 μl를 beads에 넣고 94 ℃에서 8분간 반응시켜 RNA를 조각낸 후, tube를 다시 magnet에 거치 후 20 μl를 새 tube로 옮겨 줌.

○ First Strand cDNA Synthesis: 프로토콜에 따라 준비한 First Strand Master Mix 5 μl를 각 tube에 넣어 총 25 μl를 만든 후 25 ℃에서 5분, 42 ℃에서 15분, 70 ℃에서 15분을 순차적으로 반응시켜 줌.

○ Second Strand cDNA Synthesis: 프로토콜에 따라 준비한 Second Strand Master Mix 50 μl를 각 tube에 넣어 총 75 μl를 만든 후 16 ℃에서 60분 반응시켜 줌.

○ cDNA Purification: 상온에서 135 μl (시료의 약 1.8배 volume)의 Agencourt beads를 각 샘플에 넣고 10분간 반응시킨 후 tube를 magnet에 부착시켜 70% 에탄올로 washing해준 후, elution을 위해 11 μl의 Nuclease-free water을 넣어 최종 10 μl를 새 tube로 옮겨 줌.

○ End Repair: 프로토콜에 따라 준비한 End Repair Master Mix 5 μl를 각 tube에 넣어 총 15 μl를 만든 후 25 ℃에서 30분, 70 ℃에서 10분을 순차적으로 반응시켜 줌.

○ Adapter Ligation: 프로토콜에 따라 준비한 Ligation master mix 12 μl를 각 tube에 넣고, 각 샘플에 적합한 Barcode adapter mix를 3 μl씩 넣어 총 30 μl를 만든 후 25 ℃에서 30분 반응시켜 줌.

○ Strand Selection: 프로토콜에 따라 준비한 Strand Selection master mix 70 μl를 각 tube에 넣어 총 100 μl를 만든 후 72 ℃에서 10분을 반응시켜 줌.

○ Strand Selection Purification: 80 ㎕ (샘플의 약 0.8배 부피)의 Agencourt bead suspension을 넣고 상온에서 10분간 반응시켜 DNA가 bead에 달라붙게 함. 이후 magnet을 이용하여 70% 에탄올로 washing 해주고 elution을 위해 16 ㎕의 Nuclease-free water을 넣어 최종 15 ㎕를 새 tube로 옮겨 줌.

○ Library Amplification: 프로토콜에 따라 준비한 Library Amplification A Master Mix 85 ㎕를 각 tube에 넣어 총 100 ㎕를 만든 후 37 ℃에서 10분, 95 ℃에서 2분, (95 ℃에서 30초, 60 ℃에서 90초) X 2번, (95 ℃에서 30초, 65 ℃에서 90초) X 15번, 그리고 65 ℃에서 5분을 순차적으로 반응시켜줌.

○ Amplification Library purification: 100 ㎕ (샘플의 약 1배 부피)의 Agencourt bead suspension을 넣고 상온에서 10분간 반응시켜 DNA가 bead에 달라붙게 함. 이후 magnet을 이용하여 70% 에탄올로 두 번 washing 해주고 elution을 위해 31 ㎕의 Nuclease-free water을 넣어 최종 30 ㎕를 새 tube로 옮겨 줌.

○ 제작이 완료된 library는 LabChipGX DNA High Sensitivity chip과 DNA High Sensitivity Reagent Kit을 사용하여 library 농도 및 크기 분포를 확인하였음.

다. PacBio 기반 data 생성방법

그림 15. PacBio Sequel 장비

가) 장비 및 시약

○ 장비: PacBio Sequel (PacificBio Science, USA) (그림 15), BluePippin

(Sage Science, U.S.A.), Qubit Fluorometer (ThermoFisher Scientific. U.S.A.), Centrifuge 5424 (Eppendorf, Germany), Agilent Bioanalyzer (Agilent, U.S.A.)

○ 시약: AMPure XP bead (Beckman Coulter, U.S.A.)

○ 키트/소모품: SMRTbell™ Template Prep Kit 1.0, g-TUBE (Covaris, U.S.A.), High Sensitivity DNA chip (Agilent, U.S.A.)

○ PacBio library 제작의 전체적인 scheme은 그림 16에 보이는 바와 같음.

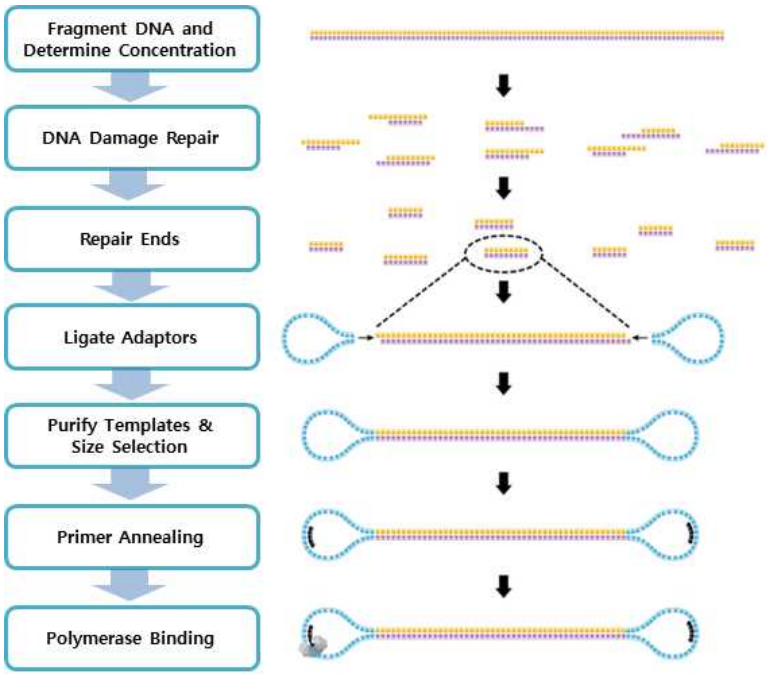

그림 16. PacBio library 제작 process

나) 실험방법

○ Genomic DNA shearing: 10 μg의 intact한 high quality genomic DNA는 Covaris g-TUBE에 넣고 48000 rpm에서 1분 원심분리하여 무작위적으로 20 kb로 shearing 해줌.

○ g-TUBE의 사용 프로토콜에 따라 샘플을 회수 후 15분 이내에 AMPure Beads를 이용하여 제조사 프로토콜에 따라 DNA concentrate를 수행하고 Bioanalyzer로 품질을 확인함.

○ DNA damage repair: Sheared DNA는 DNA damage repair 단계를 통해 abasic site, nicks, blocked 3'-ends, oxidized guanines/pyrimidines이

repair됨. Repair 되지 않은 damage site는 exonuclease 처리 단계에서 library yield가 낮은 결과를 보이게 되며, 차후 sequencing 결과 짧은 read length가 더 많은 data를 생성하게 됨.

○ Repair ends: blunt-end ligation 단계를 위한 fragments를 수선하는 과정으로 T4 DNA polymerase에 의해 fragments의 5’ 부분은 채워지고, 3’ overhang은 제거되어 blunt end를 가진 fragment를 만든 후 T4 Polynucleotide Kinase는 5’ 말단에 인산기를 붙이고 3’ 말단에서는 인산기를 떼어 수산기로 만드는 것으로 damage을 입었을 수 있는 fragment들도 다음 단계인 ligation에 적합한 DNA fragment로 만들어 줌.

○ Ligation adapter: Hairpin 구조의 adapter를 fragments 말단에 붙어, SMRT bell template 구조를 만듬. 반응 시간은 insert size에 따라 달라지며, 17 kb 이상의 SMRT bell template제작은 25 ℃에서 15~20시간 반응함.

○ Purify templates: Ligation 과정 후 생긴 다양한 templates 중에 완벽한 SMRT bell 구조의 templates를 제외한 ligation에 실패한 template들을 제거하는 과정임.

- Exonuclease Ⅲ: double-stranded DNA에 특이적으로 작동하며 3’→5’ exodeoxyribonuclease activity를 보임.
- Exonuclease Ⅶ: single-stranded에 특이적으로 작동하며 3’→5’와 5’→3’ exodeoxyribonuclease activity를 보임.

○ Size selection: 본 과정은 20 kb이상의 large insert library 제작시 반드시 필요한 과정이며, BluePippin system 등을 이용하여 우선적으로 load되는 짧은 template를 제거하고, 20kb 이상의 SMRT bell templates를 분리함. Long read lengths는 *de novo* assembly에 매우 필수이며, large insert libraries로 더 긴 subread들을 얻을 수 있게 됨. 분리된 20kb 이상의 SMRT bell templates들은 AMPure XP beads로 5번의 washing 후 농축함.

○ Primer annealing and polymerase binding: PacBio에서 제공된 binding calculator에 최종 제작된 SMRT bell library 농도와 insert size, SMRT cell수 등을 입력하여 sequencing에 적합한 standard condition을 만든 후 Annealing sequencing primer와 P6 polymerase binding을 진행하여 complex” 를 만듬. 이후 Mag bead에 complex를 binding하여 기기에 loading 준비하고 총 6개의 cell을 이용해 시퀀싱함.

4. 생물정보분석

가. PacBio 기반 *de novo* assembly

○ Sequel에서 생성된 BAM data를 PacBio의 SMRTLINK (ver. 6.0.0.47841)에 포함되어있는 HGAP4 protocol에 투입함.

○ HGAP4 protocol은 크게 subread filtering, pre-assemble (read correction), draft assembly 그리고 polishing의 4 단계로 이뤄지는 과정임 (그림 17).

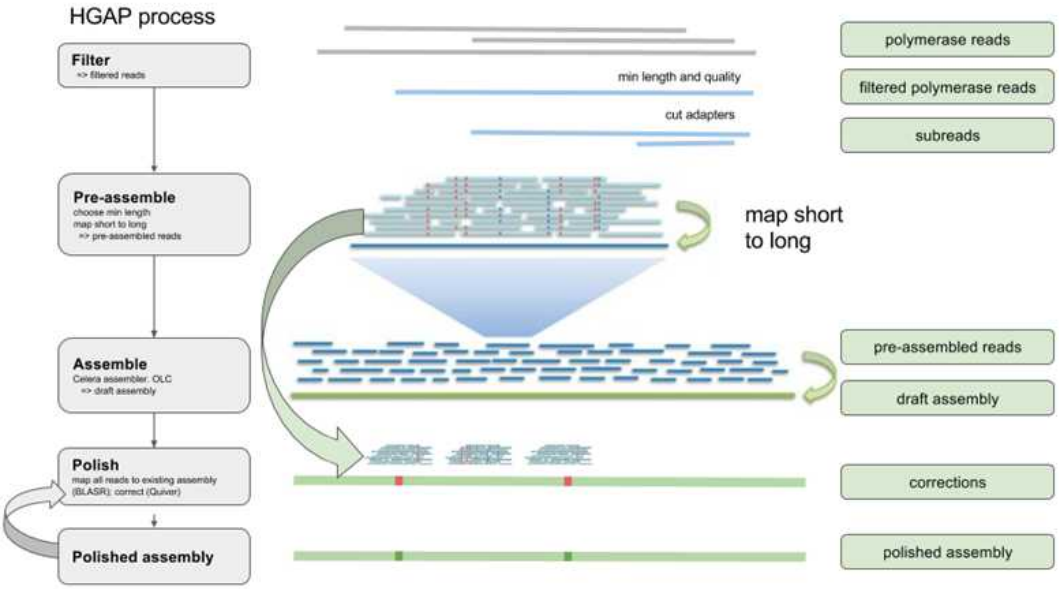

그림 17. HGAP4 process 개념도
http://sepsis-omics.github.io/tutorials/modules/pacbio/

○ HGAP4 protocol에 투입된 BAM data는 minimum length 500 및 Read Quality ≥ 0.8을 포함한 조건으로 수행된 quality & adapter filtering을 거쳐 Filtered Subreads로 변환함.

○ Filtered Subreads는 약 20% 정도의 random error를 포함하고 있으며 HGAP4 protocol의 preassembly process (30X 정도의 긴 PacBio서열에 짧은 PacBio서열을 alignment하여 랜덤하게 발생하는 Pacbio 서열상의 에러를 교정하는 방법)에 투입되어 약 2% 정도의 random error를 포함한 Corrected Subreads로 변환함.

○ Corrected Subreads는 HGAP4 protocol의 Falcon assembler를 통하여 draft contigs로 조립됨.

○ HGAP4 외에 CANU로도 draft contig 조립을 시도하였으나 적합한 결과를 얻지 못하여 HGAP4로 형성된 결과물을 활용하여 이후 절차를 진행함.

나. Contig Polishing (1차)

○ HGAP4 protocol에서 산출된 draft contigs는 같은 protocol에 포함된 Arrow process에 투입됨 (그림 18).

○ Arrow process에서는 draft contigs에 contig 조립 시 활용된 PacBio read를 다시 mapping 하여 consensus contig를 생성하는 방식 즉 Resequencing 방식으로 남아있는 Random error를 교정하여 1차 polished contigs를 산출함.

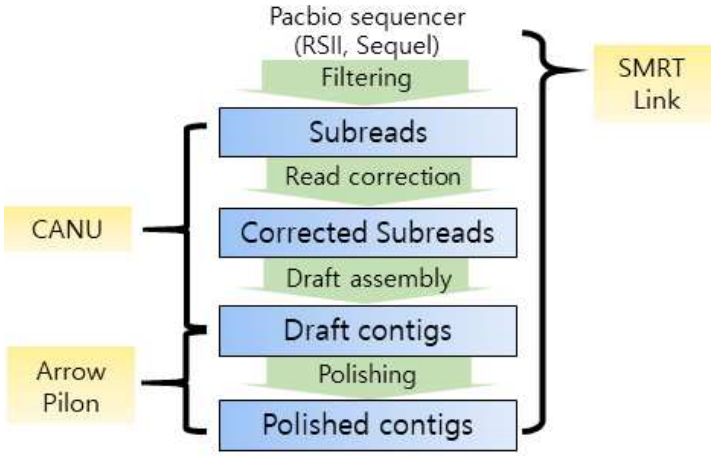

그림 18. PacBio read를 활용한 draft contigs 및 1차 polished contig 조립 개념도

다. Contig Polishing (2차)

○ 10X linked DNA reads는 quality_trim (ver. 4.010.83648)에 투입되어 Q30이상으로 trimming 되고 trimmed 10X reads는 bwa (ver. 0.7.8-r455)로 1차 polished contigs에 mapping하여 sam file 산출함.

○ sam file은 samtools (ver. 1.9)에 투입되어 sorted bam file로 변환함.

○ sorted bam file과 1차 polished contigs는 pilon(ver. 1.22)에 투입되어 fasta 형태의 HiSeq polished contigs를 산출함.

○ hiseq polished contigs에 다시 trimmed 10X reads를 mapping 하고 polishing 하는 과정을 12번 더 반복하여 2차 polished contigs를 산출함 (그림 19).

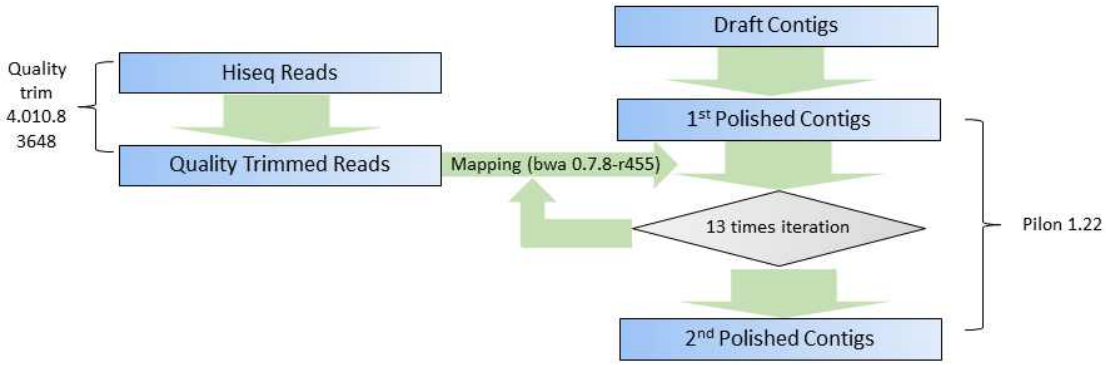

그림 19. HiSeq DNA read를 활용한 2차 polishing 개념도

라. Scaffolding

◯ Scaffolding은 assembly 된 contigs를 원거리 정보에 근거하여 ordering 하는 작업이며 (그림 20) 본 연구에서는 10X linked reads를 contigs에 mapping 하여 linked reads의 원거리 정보로 scaffolding을 수행하는 ARCS (ver. 1.0.3) 를 활용함 (Yeo S. *et al.*, 2017) (그림 21).

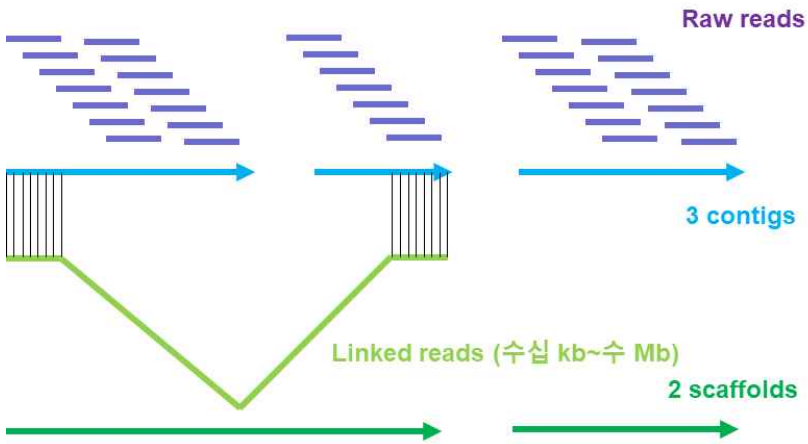

그림 20. Scaffolding 개념도

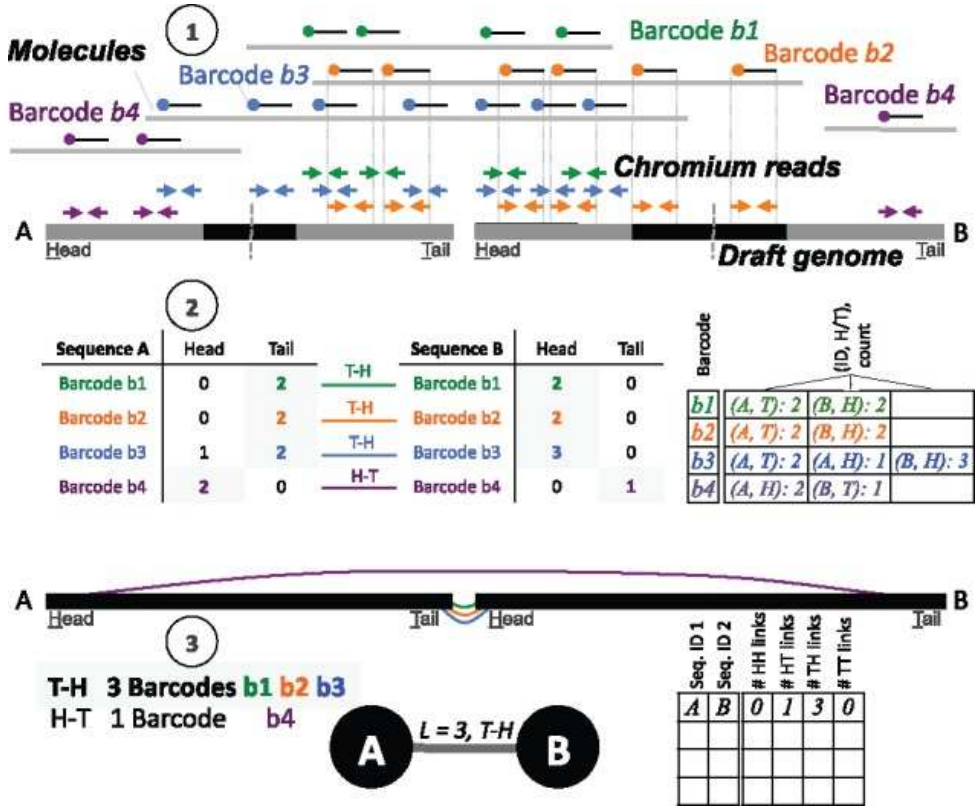

그림 21. ARCS 원리 개념도 (Yeo S. *et al.*, 2017)

○ Scaffolding을 위한 첫 번째 사전작업으로 10X linked reads를 longranger basic (2.1.6)에 투입하여 barcode processed (read trimming, barcode error correction, barcode whitelisting, and attaching barcodes to reads를 포함) 10X linked reads를 산출하고 script를 활용하여 CHROMIUM_interleaved.fastq 파일로 전환함.

○ Scaffolding을 위한 두 번째 사전작업으로 2차 polished contigs를 script를 활용하여 대문자로 전환한 후 header를 일련번호로 변환함.

표 2. Scaffolding 수행과정에서 활용된 script 예시

Function	Script	Ref.
CHROMIUM_interleaved.fastq 생성	unpigz -c barcoded.fastq.gz ¦ perl -ne 'chomp;$ct++;$ct=1 if($ct>4);if($ct==1){if(/(\@\S+)\sBX\:Z\:(\S{16})/){$flag=1;$head=$1."_".$2;print "$head\n";}else{$flag=0;}}else{print "$_\n" if($flag);}' > CHROMIUM_interleaved.fastq	https://github.com/bcgsc/arcs/issues/42, https://github.com/bcgsc/arcs/issues/9, https://162.38.181.155/reservebenefit/genome_assemblies_collection/blob/6548c501c4d3b0663dc35fe9b306205a5adceb97/arcs/pipeline.sh
대문자 전환	awk '{ print toupper($0) }' contigs.fasta > contigs_upper.fasta	http://biofeed.tumblr.com/post/45747795087/how-to-convert-text-files-to-all-upper-or-lower
header 일련번호 전환	cat contigs_upper.fasta ¦ perl -ne 'chomp;if(/^>/){$ct++;print ">$ct\n";}else{print "$_\n";}' > contigs_upper_1.fasta	https://github.com/bcgsc/arcs/issues/42, https://github.com/bcgsc/arcs/issues/9, https://162.38.181.155/reservebenefit/genome_assemblies_collection/blob/6548c501c4d3b0663dc35fe9b306205a5adceb97/arcs/pipeline.sh

○ CHROMIUM_interleaved.fastq를 bwa를 활용해 2차 polished contig에 mapping하여 sam file을 산출하고 산출된 sam file을 samtools에 투입하여 sorted bamfile을 생성함.

○ sorted bamfile과 2차 polished contig을 arcs에 투입하여 ARCS graph file 및 tigpair_checkpoint file (ARCS에 포함된 makeTSVfile.py script를 통해 생성) 생성함.

○ tigpair_checkpoint file을 LINKS(ver. 1.8.7)에 투입하여 fasta 형태의 1차 scaffold 생성함.

○ 1차 scaffold에 다시 CHROMIUM_interleaved.fastq를 mapping 하고 ARCS 및 LINKS를 통해 scaffolding 하는 과정을 5번 더 수행함 (수행과정에서 활용된 linked reads의 양은 과정에 따라 변동이 있었으며 약 30Gb 정도였음) (그림 22, 표 2).

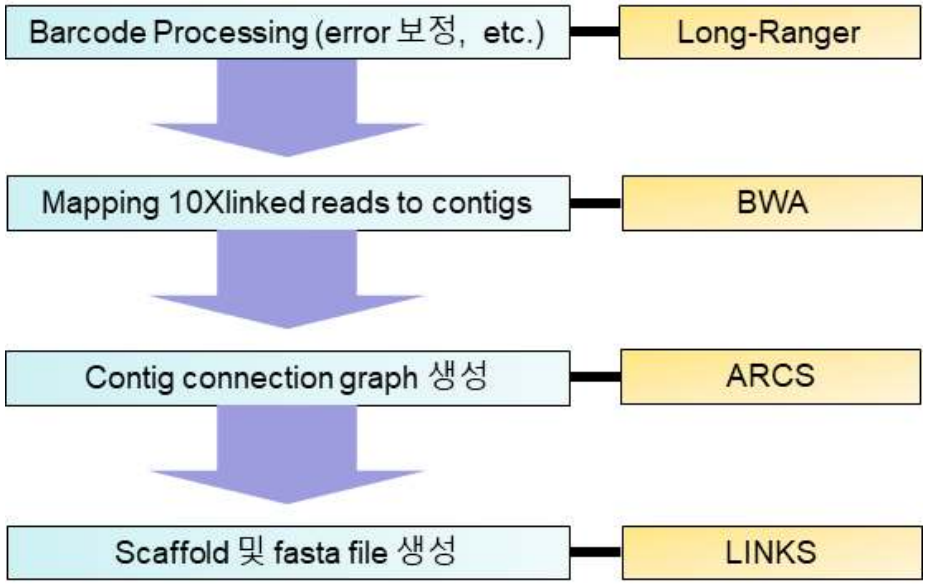

그림 22. Scaffolding process 모식도

마. Assembly validation

○ 최종 결과물인 Scaffold의 길이를 Flow-cytometry 및 k-mer를 통해 예상된 결과와 비교함.
○ Scaffold의 개수 및 길이를 통해 전반적인 DNA의 연속성을 평가함.
○ Scaffold를 BUSCO(ver. 3.0.2)에 투입하여 해당 근연종에 포함된 core gene의 검출여부 조사함 (본 연구에서는 eudicotyledons_obd10 DB를 활용하였음).
○ Assembly 도식화는 DNAPlotter (https://www.sanger.ac.uk/science/tools/dnaplotter)를 사용하였음.

바. RepeatMasker

○ RepeatMasking을 위한 사전작업으로 repeat library를 제작하였으며 본 연구에서는 Repbase(20170127) library 및 Repeat Modeler (ver. 1.0.11)와 Scaffold를 활용해 제작한 *de novo* repeat library를 통합한 Combined repeat library를 활용함.
○ Combined repeat library와 Scaffold를 Repeatmasker (ver. 4.0.7)에 투입하여 Repeat masked Scaffold (hard masked)를 제작 하고 Scaffold 내 Repeat 현황 분석결과물 도출함 (그림 23).

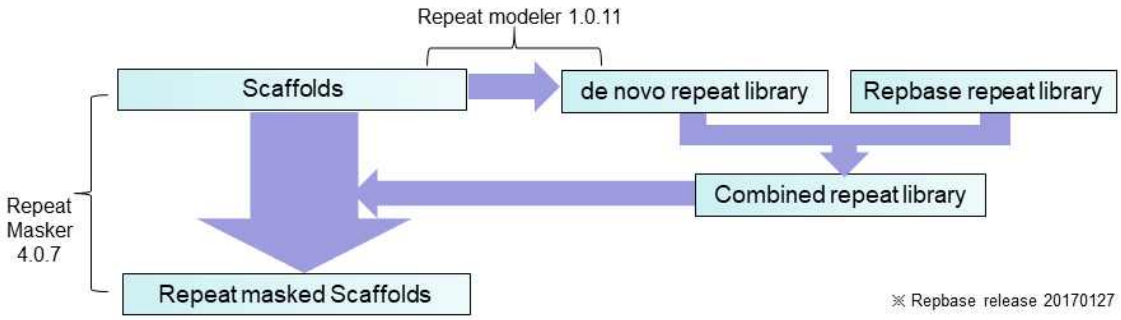

그림 23. Repeat masking process 개념도

○ RepeatMasker 외에도 Scaffolds를 misa (ver. 1.0)에 투입하여 Genome 내 SSR의 존재 여부를 확인하였음 [unit size / minimum number of repeats : (2/10) (3/4) (4/4) (5/4) (6/4)].

사. Annotation

○ 본 연구에서는 Hiden Markov Model과 같은 순 이론적(ab-initio) 방법으로 수행되는 Gene prediction의 오류를 최소화하기 위해 RNAseq의 Genome mapping 결과를 hint 로 활용하는 방식을 채택하였음.

○ RNAseq read의 첫 번째 전처리 과정으로 trim_galore (ver. 0.4.2)에 투입되어 1차 adapter 제거 및 quality trimming 수행하였으며 Q20 이상의 trimming 된 RNAseq read를 23.1Gb 산출함 (독도 샘플 : 11,945,566,410bp, 울릉도 샘플 11,263,648,433bp).

○ RNAseq read의 두 번째 전처리 과정으로 trim_galore (ver. 0.4.2)에를 통해 산출된 RNAseq short read data는 Trimmomatic (ver. 0.36)에 투입되어 2차 adapter 제거 및 quality trimming 수행함.

○ adapter 제거 및 quality trimming 된 paired RNAseq read를 STAR(ver. 2.6.0)에 투입하여 hard masking 된 Scaffold에 mapping 하여 결과물로 sorted bam file 생성함.

○ sorted bam file과 Scaffold를 WebAugustus (Hoff KJ *et al.*, 2013)에 투입하여 Gene prediction을 수행하여 GFF 파일과 fasta 형태의 transcript sequence 결과물을 도출 함.

○ transcript fasta file를 Blastall(ver. 2.2.26, blastx)을 활용한 (e valve : 10^{-5}) NCBI의 Plant reference DB (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plant/) 기반의 function

annotation을 수행하여 각 CDS sequence에 관한 1차 function annotation을 결과물을 도출함.

○ 1차 function annotation을 결과물을 Blast2GO (ver. 5.2.5) 에 NCBI의 Plant reference DB에 기반하여 function annotation된 transcript에 관한 Gene Ontology annotation을 수행하여 GO annotation 결과물 및 KEGG 결과물 도출함 (그림 24).

그림 24. Annotation Process 개념도

Ⅲ. 연 구 결 과

Ⅲ . 연구결과

1. 유전체 크기 측정

가. Flow-cytometry를 이용한 유전체 크기 측정

○ Flow cytometer를 이용하여, 콩, 울릉도 섬초롱 그리고 독도 섬초롱의 gDNA 크기를 측정하였음. 콩의 gDNA는 1.12Gb라는 것을 알고 있기 때문에, 이것을 참조하여, 울릉도 섬초롱과 독도 섬초롱의 gDNA 크기를 계산하였음 (그림 25-27, 표 3).

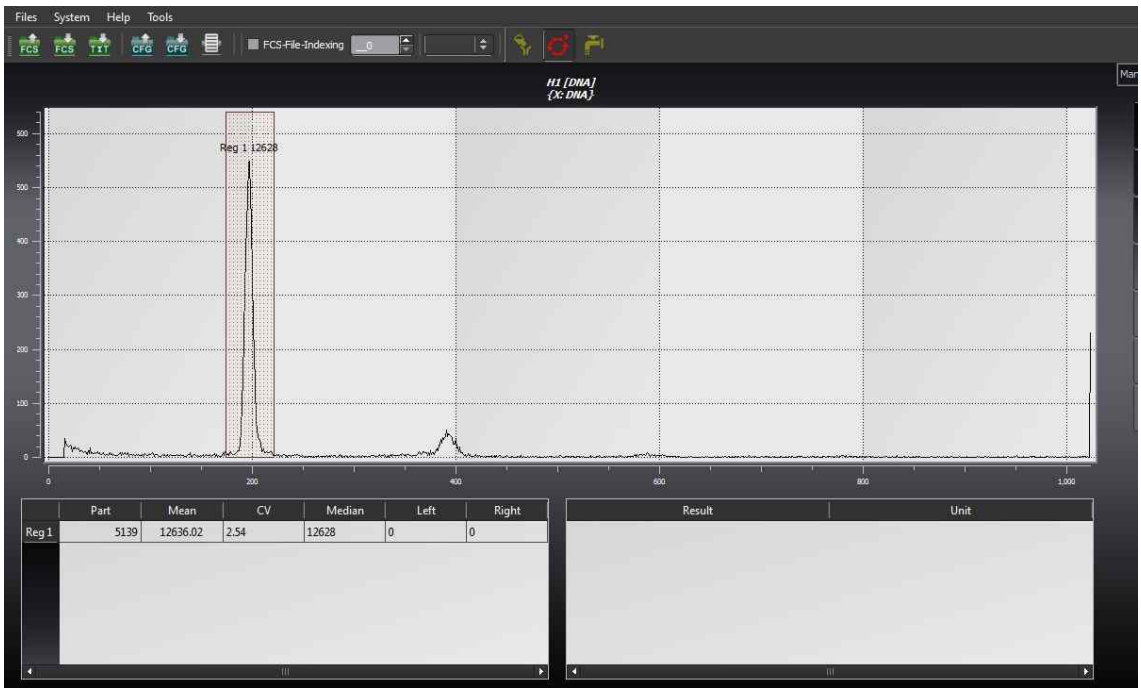

그림 25. 콩의 flow-cytometry 결과

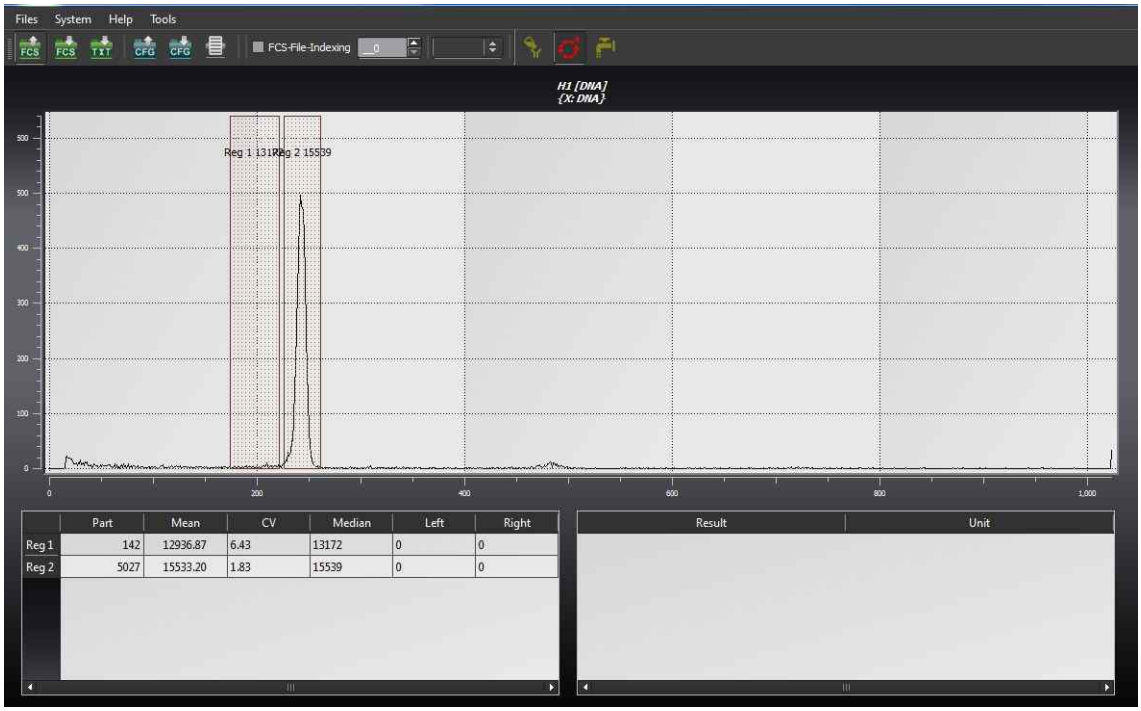

그림 26. 울릉도 섬초롱의 flow-cytometry 결과

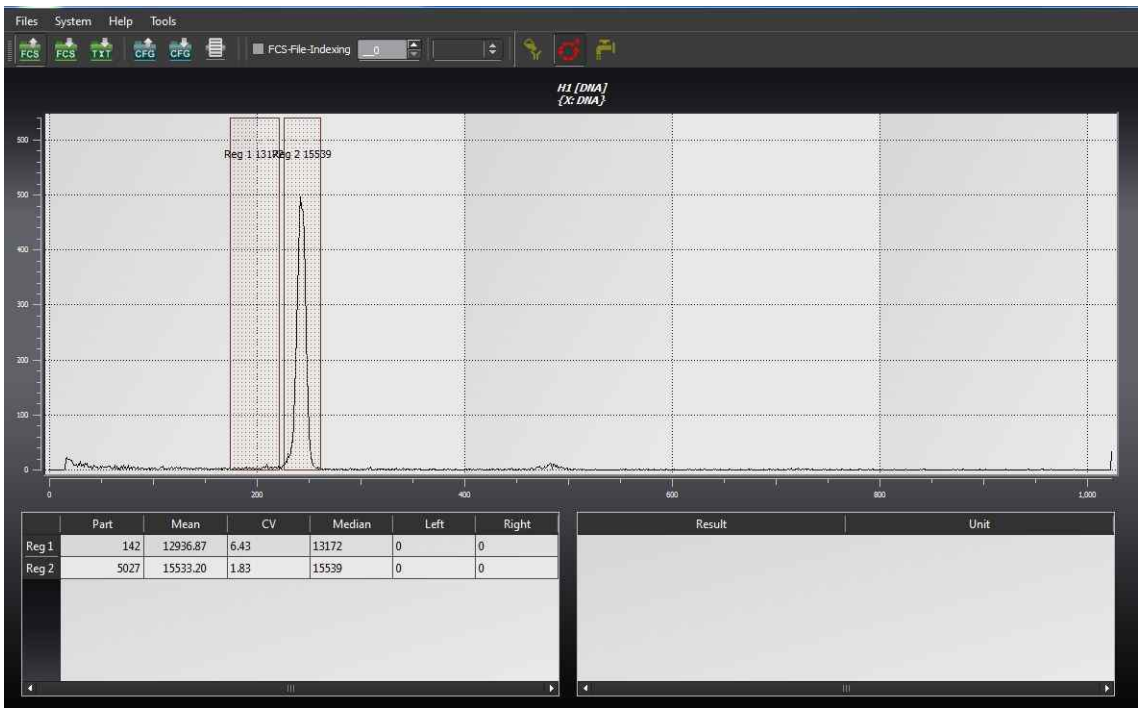

그림 27. 독도 섬초롱의 flow-cytometry 결과

표 3. Flow-cytometry를 통한 지놈 크기 결정

시료	Part	Median	Genome Size
콩	5,139	12,628	1.12Gb
울릉도섬초롱꽃	5,027	15,539	1.38Gb
독도섬초롱꽃	5,223	10,941	0.97Gb

나. K-mer 분석을 이용한 유전체 크기 측정

◯ 약 69G의 short read data를 사용하여 최소 10mer에서 최대 70mer 까지 k-mer size를 변경하여 최적의 kmer graph를 찾은 결과 36mer에서 약 0.84Gb로 게놈 사이즈를 예측하였음 (그림 28).

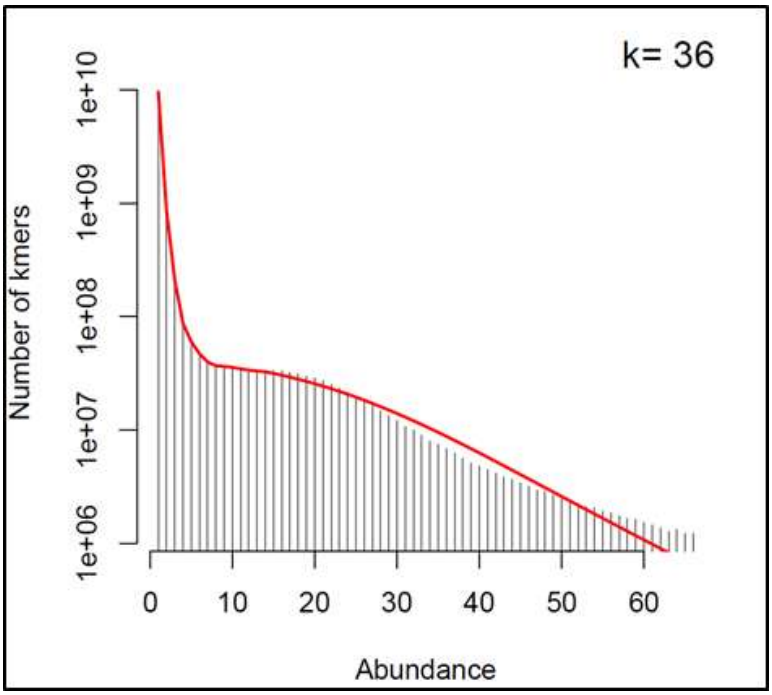

그림 28. K-mer 분석 graph

2. 유전체분석을 위한 NGS data 생산

가. DNA 및 RNA QC

1) DNA QC

가) 전기영동

○ 1% agarose gel을 이용하여 추출된 DNA를 확인하였음 (그림 29).

○ 추출된 독도 섬초롱꽃 gDNA를 10배로 희석한 후 1 μl를 loading하였음 (그림 10).

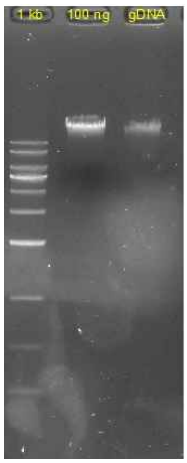

그림 29. 독도 섬초롱꽃
gDNA electrophoresis

나) Picogreen의 형광(fluorescence-based quantification) 기반 농도 측정

* 모든 재료 희석할 때에는 1X TE buffer를 이용하였음.

○ λDNA를 50배 희석하여 사용하여 standard curve를 만들었음 (그림 30).

○ gDNA를 10배 희석한 후 1 μl를 이용하여 농도 측정하였으며 picogreen은 200배 희석하여 사용하였음 (표 4).

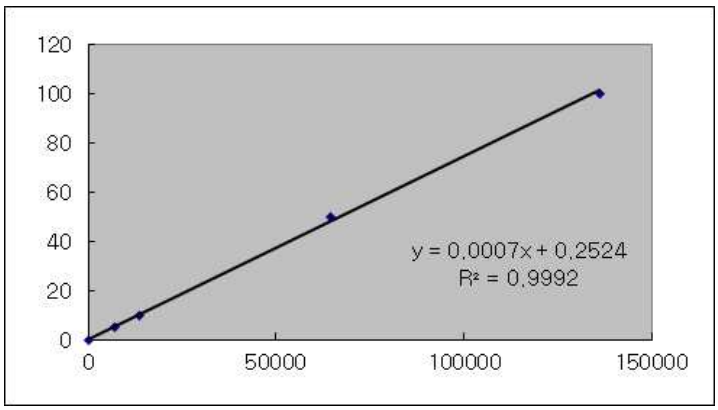

그림 30. λDNA의 Standard curve

표 4. Picogreen을 이용하여 측정한 독도 섬초롱꽃 gDNA 농도

시료명	20배 희석된 시료 농도(ng/μl)	Total DNA (μg) in 100 μl
독도 섬초롱꽃	28.9608	57.9216

2) RNA QC

○ 각 조직에서 추출된 RNA는 HT RNA Labchip Kit와 LabChipGX (Caliper LifeSciences)를 이용하여 28S rRNA와 18S rRNA의 비율을 관찰하였으며 (그림 31), 각 RNA 샘플의 농도 또한 측정하였음 (표 5).

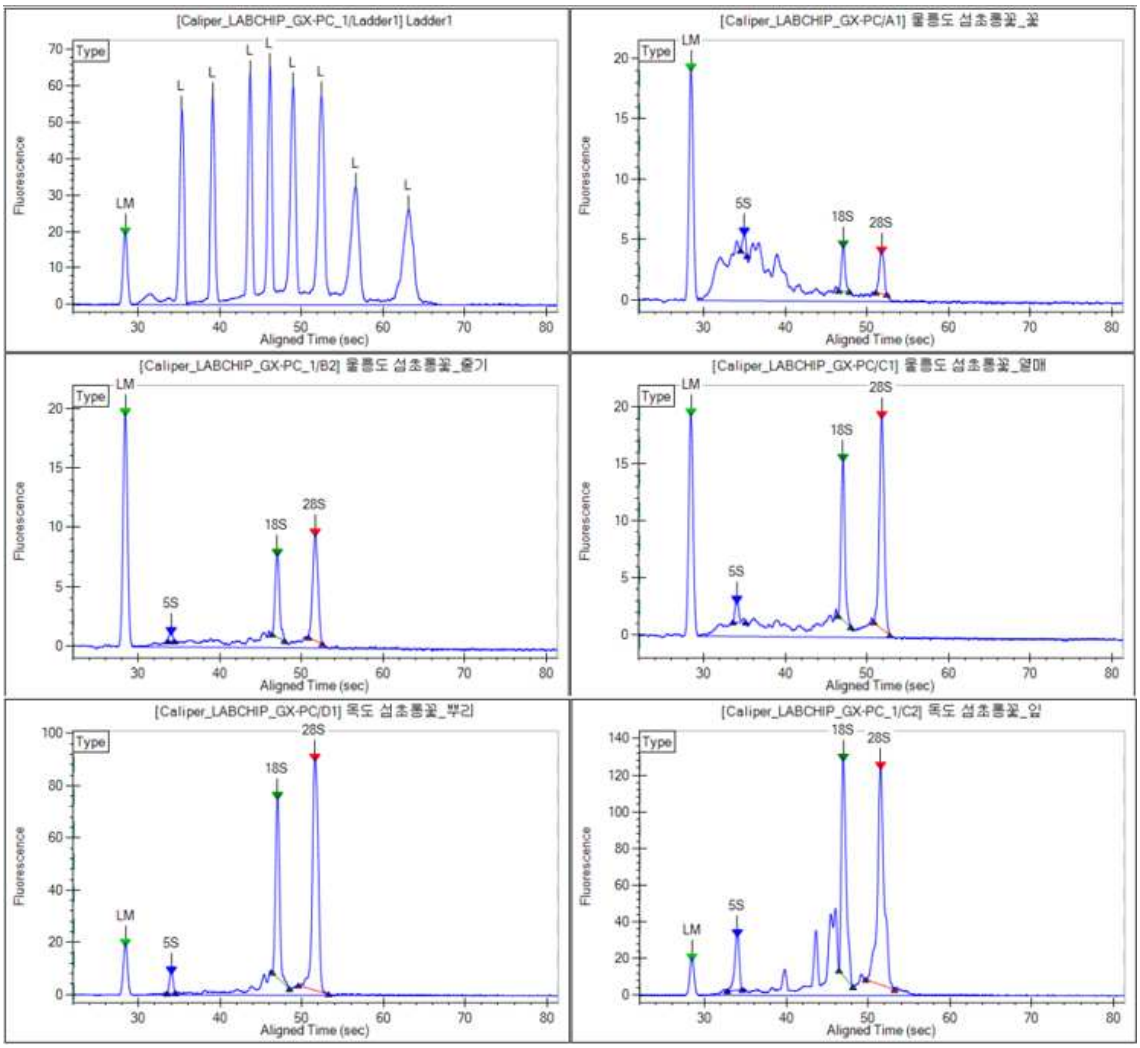

그림 31. 여러 조직에서 추출된 RNA의 Capiler LabChipGX에서의 running 결과

표 5. Capiler LabChipGX를 통한 RNA의 품질 확인

Sample Name	Well Label	Total Conc. (ng/μl)	RNA Quality Score	Peak Count	RNA Area	rRNA Area Ratio [28S/18S]	rRNA Height Ratio [28S/18S]	rRNA Fast Area Ratio	5S Area	5S % Total	18S Area	18S % Total	28S Area	28S % Total
Ladder1	Ladder 01	480.00			364.70									
울릉도 섬초롱꽃_꽃	A01	57.89	4.2	8	46.33	1.23	0.97	0.47	0.8	1.70	1.94	4.20	2.39	5.20
울릉도 섬초롱꽃_줄기	B02	29.96	7.8	5	23.22	1.57	1.29	0.29	0.39	1.70	4.32	18.60	6.78	29.20
울릉도 섬초롱꽃_열매	C01	53.35	7.5	6	44.18	1.73	1.31	0.28	1.18	2.70	7.51	17.00	12.96	29.30
독도 섬초롱꽃_뿌리	D01	203.88	9.3	9	165.55	1.88	1.21	0.14	4.32	2.60	39.17	23.70	73.72	44.50
독도 섬초롱꽃_잎	C02	490.56	7.8	12	385.79	1.39	0.98	0.28	19.99	5.20	79.99	20.70	111.42	28.90

나. HiSeq 기반 data 생성방법

1) 10X Linked-read sequencing

○ BluePippin으로 40 kb 이상 size-selection 후 Qubit 측정 결과는 표 6에 나타남.

표 6. 섬초롱꽃 DNA 정량

Sample	conc. (ng/μl)	Total vol. (μl)	Total amount (ng)
섬초롱꽃	24.6	30.0	738.0

○ BluePippin으로 40 kb 이상으로 size-selection된 gDNA의 확인. Pippin Pulse (Pulsed-field Power for Gel Boxes)로 75V로 16시간 전기영동하였음 (그림 32). Size marker는 Invitrogen 사의 1kb extension ladder임.

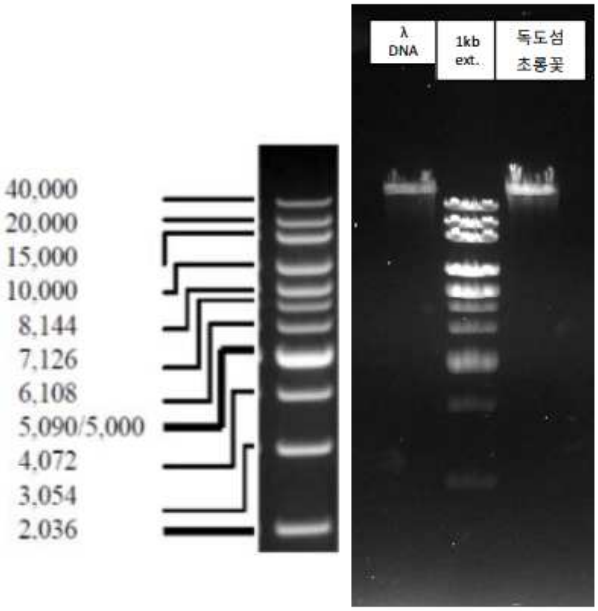

그림 32. BluePippin으로 40 kb 이상으로 size-selection된 gDNA의 확인

○ Post GEM QC: GEM 안에서 random hexamer가 genomic DNA에 붙어 합성한 fragment들의 분포를 Agilent Bioanalyzer에서 High Sensitivity DNA Chips를 이용해서 살펴보았으며, 그 결과는 전형적 Post GEM 분포를 보여줌 (그림 33).

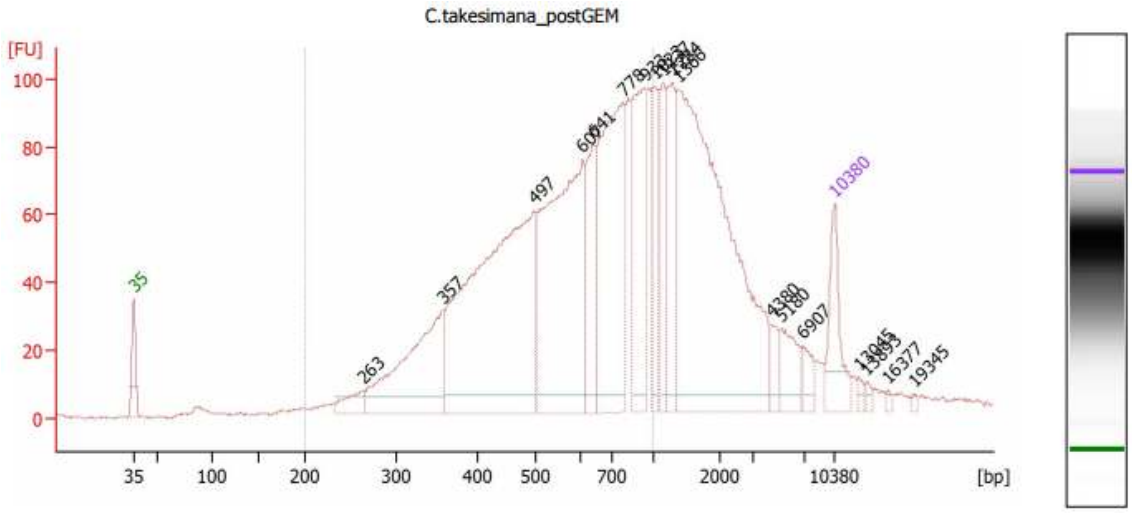

그림 33. Post GEM QC

○ 10X library 제작 후 QC: Qubit을 이용하여 측정한 농도 및 adapter 정보 (표 7)

표 7. Library adapter 정보

Sample lib.	conc. (ng/㎕)	10x index	Index sequence			
독도 섬초롱꽃	14.5	B11	GTTCCTCA	AGGTACGC	TAAGTATG	CCCAGGAT

○ DNA library는 LabChipGX (Caliper LifeSciences)를 이용하여 크기 분포를 관찰하였음 (그림 34).

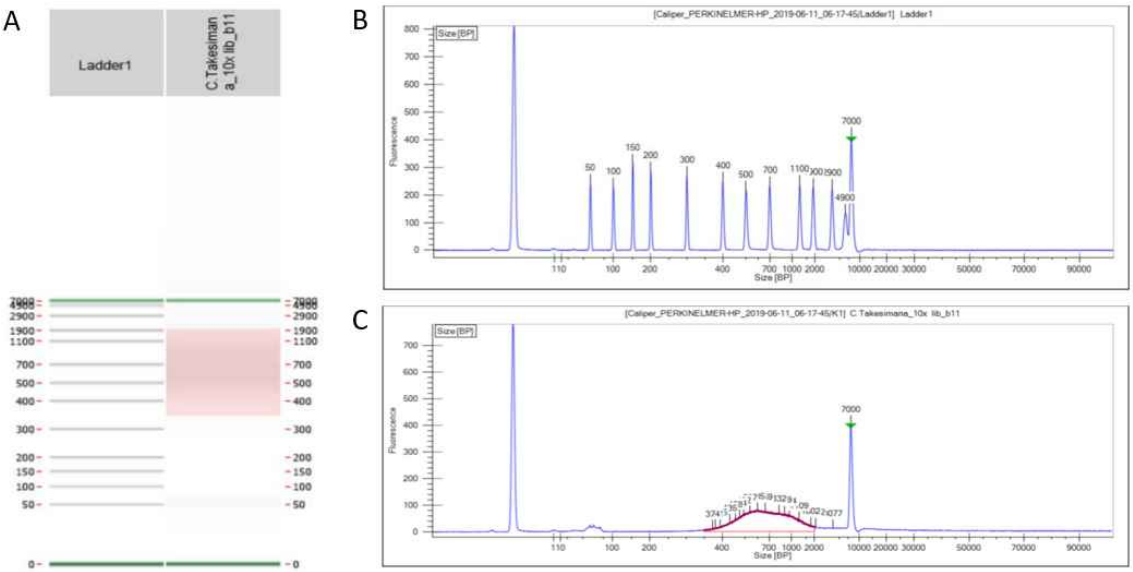

그림 34. 10X DNA library의 크기 분포. A. Size marker (B)와 DNA libarary (C)로부터 재구성한 digital capillary electrophoresis gel image B. Size marker의 capillary electrophoresis 결과 C. 10X DNA library의 capillary electrophoresis 결과

2) RNAseq

○ 제작된 RNA-Seq library 농도

- 울릉도 섬초롱꽃: 1.682 ng/㎕
- 독도 섬초롱꽃: 0.517 ng/㎕

○ RNA-Seq library QC: 완성된 RNA-Seq library의 read 분포를 LabChipGX 장비를 이용해 측정하였음 (그림 35).

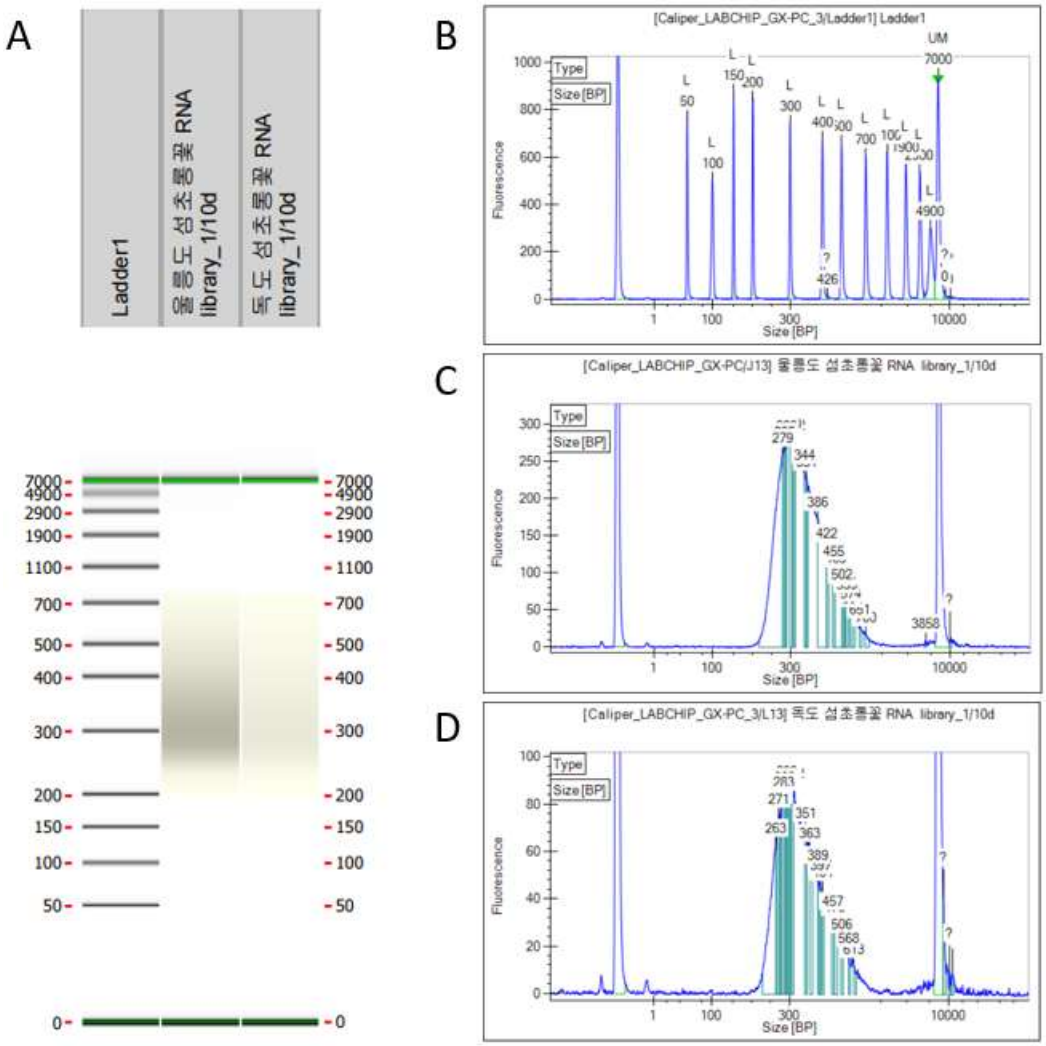

그림 35. RNA-Seq library의 크기 분포. **A**. Size marker (**B**)와 두 libarary (**C, D**)로부터 재구성한 digital capillary electrophoresis gel image **B**. Size marker의 capillary electrophoresis 결과 **C**. 울릉도 섬초롱꽃 RNA-Seq library의 capillary electrophoresis 결과 **D**. 독도 섬초롱꽃 RNA-Seq library의 capillary electrophoresis 결과

다. PacBio 기반 생산 방법

1) Genomic DNA shearing 후 QC

○ Genomic DNA shearing 후 크기는 10>kb 인 fragment가 많아야 하는데, Bioanalyzer 결과를 보면 이 조건을 충족시킴 (그림 36). Qubit Fluorometer로 측정한 농도는 157 ng/μl 임.

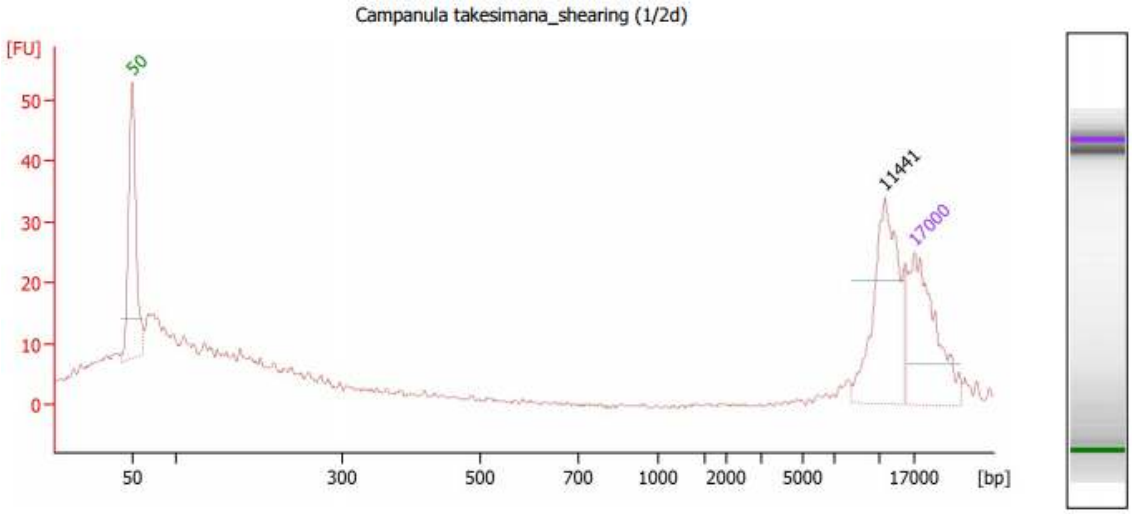

그림 36. gDNA shearing 후 Bioanalyzer를 통한 electropherogram 결과

2) Adapter ligation 후 QC

○ Shearing된 gDNA에 adapter를 ligation한 후 크기는 10>kb 인 fragment가 많아야 하는데, Bioanalyzer 결과를 보면 이 조건을 충족시킴 (그림 37). Qubit Fluorometer로 측정한 농도는 97.8 ng/μl 임.

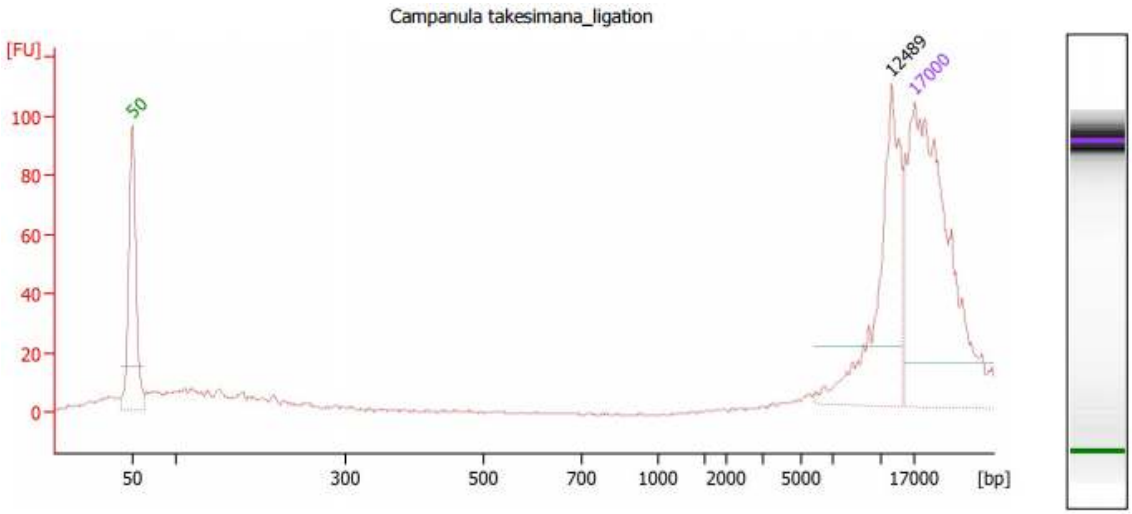

그림 37. Adapter ligation 후 Bioanalyzer를 통한 electropherogram 결과

3) 최종 library 제작 후 QC

○ BluePippin에서 20 kb 이상인 fragment들을 선별한 것으로 Bioanalyzer로 확인한 결과 20 kb 이상인 fragment가 많음 (그림 38). Qubit Fluorometer로 측정한 농도는 49.6 ng/μl 임.

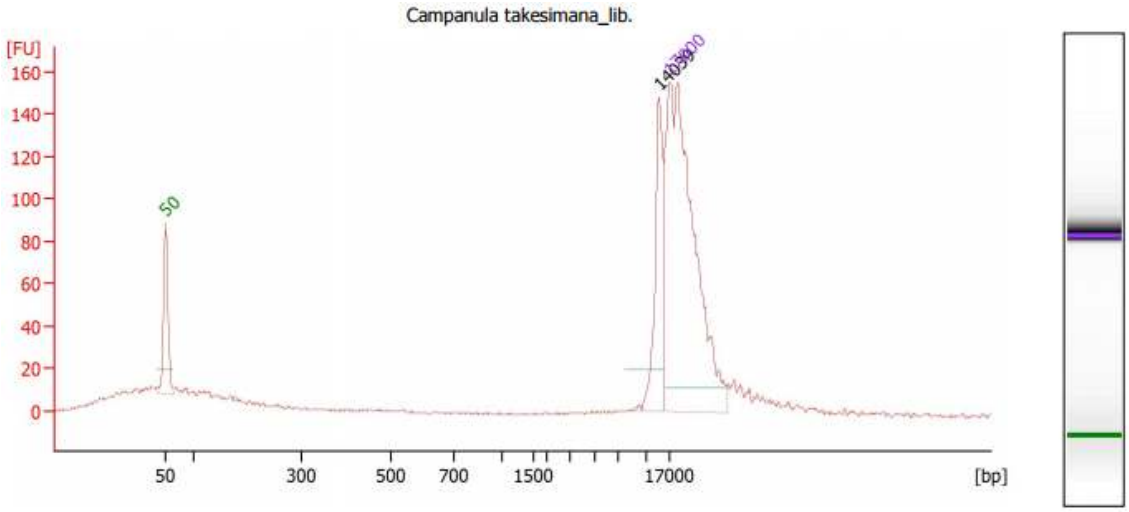

그림 38. Pacbio gDNA library 제작 후 Bioanalyzer를 통한 electropherogram 결과

3. 생물정보분석 결과

가. Raw data QC

○ 총 54Gb Q30≥80% 에 해당하는 250 PE fastq short read data 획득 (표 8)

표 8. HiSeq (Linked-read seq) raw read 생성표

10XChromium	Raw reads	
	No. of reads	Length (bp)
1	196,989,262	29,745,378,562
2	57,996,524	8,757,475,124
3	79,550,882	12,012,183,182
4	47,693,974	7,201,790,074
5	74,413,976	11,236,510,376
Total	456,644,618	68,953,337,318

○ Annotation을 위한 각 조직별 pooloing 한 RNA로 제작한 RNAseq data 생성함 (표 9)

표 9. HiSeq (RNAseq) raw read 생성표

RNA sequencing	Raw reads	
	No. of reads	Length (bp)
울릉도	78,549,162	11,860,923,462
독도	83,165,208	12,557,946,408

○ 총 3회에 6cell의 running을 거쳐 Read quality 0.86의 데이터 76.44Gb (약 76X) 산출함 (표 10).

표 10. PacBio raw read 생성표

SMRT Cell	Polymerase Read		Reads of Insert		Control Reads			Total Bases (Gb)
	Length	N50	Length	N50	No.	Length	Quality	
1	14,857	24,924	11,272	17,459	19,429	36,213	0.85	10.57
2-1	18,060	34,333	11,934	18,122	15,015	31,117	0.86	13.61
2-2	17,200	32,546	11,639	17,950	19,254	30,622	0.86	12.35
2-3	17,890	34,720	11,811	18,134	17,677	31,221	0.85	13.31
3-1	17,144	32,449	11,615	17,939	14,685	30,290	0.86	12.78
3-2	18,903	36,375	12,285	18,571	13,148	32,809	0.86	13.82
Total	17,342	32,558	11,759	18,029	16,535	32,045	0.86	76.44

나. *De novo* assembly

○ SMRT LINK의 HGAP4를 통해 도출된 1차 Polished contig는 최대 길이 1,594,050 bp, N50 230,362 bp, 총 길이 1,249,981,272 bp의 contig 개수 8,870개로 나타남.

○ HiSeq Polishing을 통해 도출된 2차 Polished contig는 최대 길이 1,583,234 bp, N50 229,849 bp, 총 길이 1,244,669,265 bp의 contig 개수 8,870개로 나타남.

○ 14번 반복된 HiSeq Polishing 과정에서 교정횟수가 4,100회 정도에서 수렴함을 확인함 (그림 39).

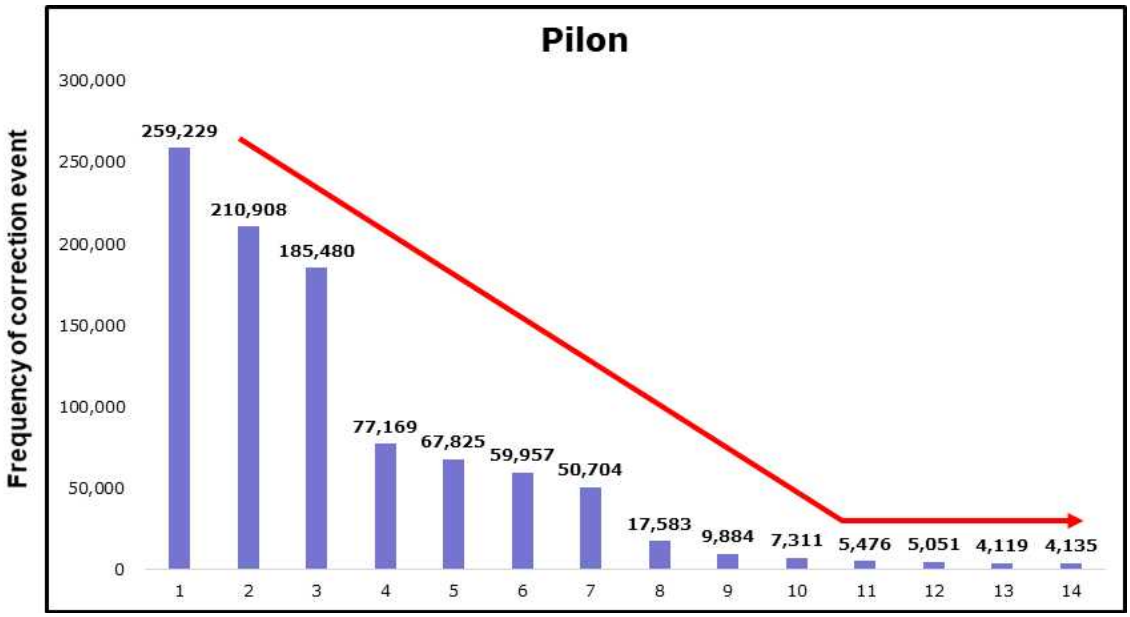

그림 39. Pilon process 결과

○ Linked Reads Scaffolding을 통해 도출된 Scaffold는 최대길이 4,264,097 bp, N50 731,187 bp, 총 길이 1,245,156,765 bp, Scaffold 개수 3,995개로 나타남 (표 11).

○ 3,995개(총 1.25G)의 scaffold는 아래 circular map으로 도식화 되었으며 각 scaffold에 위치하는 GC plot [Count(G + C)/Count(A + T + G + C)X100%] 및 GC skew [(G-C)/(G+C)] 분포로 나타내었음 (그림 40).

표 11. *De novo* assembly 및 scaffolding 결과

Statistics	1st Polished Contigs	2nd Polished Contigs	Scaffolds
Total number	8,870	8,870	3,995
Total length	1,249,981,272 bp	1,244,669,265 bp	1,245,156,765 bp
N50 length	230,362 bp	229,849 bp	731,187 bp
Average length	140922.35	140,323 bp	311,678.79 bp
Maximum length	1,594,050 bp	1,583,234 bp	4,264,097 bp
Gap number	0	0	4,875(1gap:100bp)

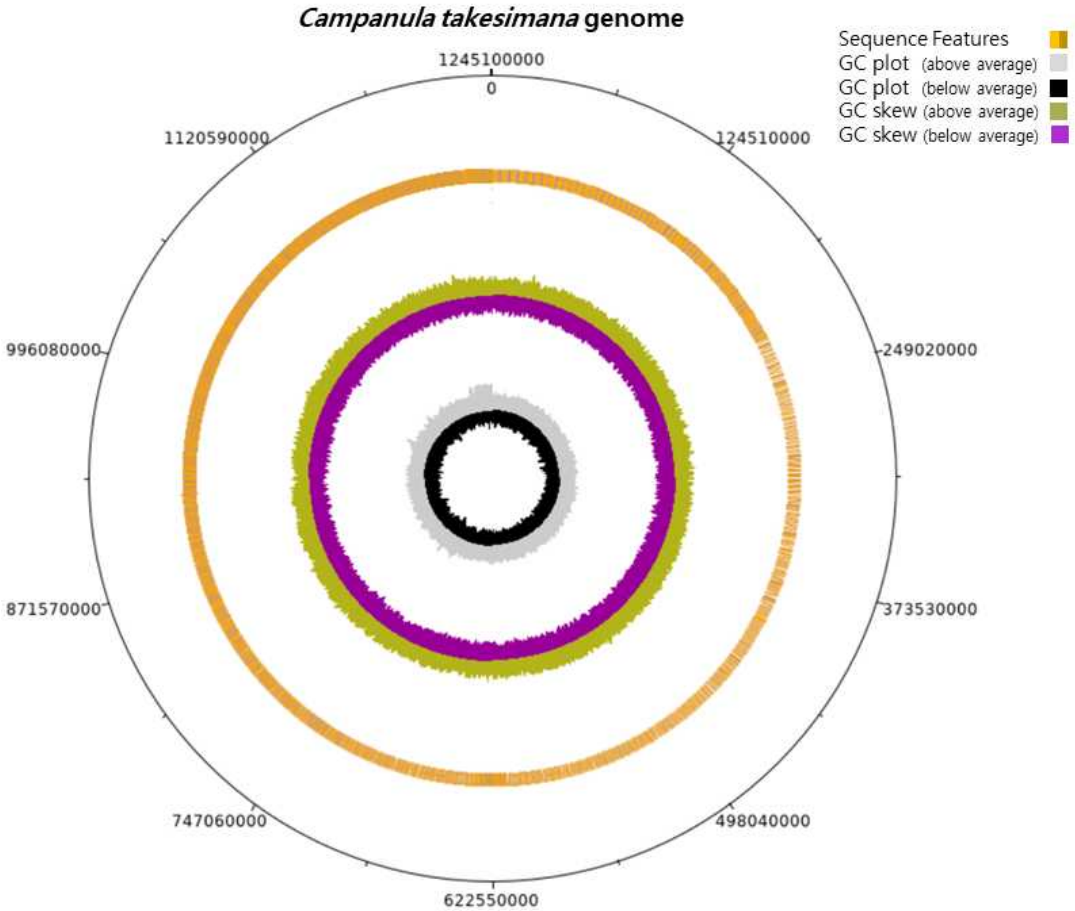

그림 40. 독도 섬초롱꽃의 Genome Map

○ Scaffold를 BUSCO에 투입하여 해당 근연종에 포함된 core gene의 (본 연구에서는 eudicotyledons_obd10 DB를 활용하였음) 검출여부를 조사한 결과 Complete BUSCO값이 90%에 가깝게 도출된 것으로 파악됨 (표 12).

표 12. BUSCO 결과

Categories	No. of genes	Percentage
Complete BUSCOs	1,784	84.1
Complete and single-copy BUSCOs	1,596	75.2
Complete and duplicated BUSCOs (D)	188	8.9
Fragmented BUSCOs (F)	47	2.2
Missing BUSCOs (M)	290	13.7
Total BUSCO groups searched	2,121	100%

다. Repeat 분석

○ RepeatMasker를 활용하여 도출된 Repeat의 총 길이는 896,102,186 bp 로서 전체 genome의 71.97% 였고 이중 LTR elements가 633,544,859 bp (50.88%)로 가장 많은 비율을 차지하였으며 별도로 분류되지 않은 Repeat은 198,877,753 bp (15.97%)로 나타남 (표 13).

표 13. Repeat 분석 및 종류

Type	No. of elements	Length occupied (bp)	% of sequences
SINEs	2,337	488,910	0.04
LINEs	16,253	9,369,397	0.75
LTR elements	342,780	633,544,859	50.88
DNA elements	126,860	53,821,267	4.32
Unclassified	386,812	198,877,753	15.97
Total	875,042	896,102,186	72.59

○ misa를 활용하여 도출된 SSR은 196,182개 였으며 3,903개의 scaffold에서 발견되었음 (표 14).

표 14. SSR 분석 및 종류

Total number of SSR	Unit size (repeat #)	Number of SSR
196,182	2 (10)	43,204
	3 (4)	127,999
	4 (4)	15,675
	5 (4)	5,205
	6 (4)	4,099

라. Annotation

○ 본 연구에서는 WebAugustus에 RNAseq의 Genome mapping 결과를 hint로 활용하는 방식으로 gene prediction을 수행하였으며 Plant Refseq Database로 총 135,438 개의 gene을 도출하였음 (표 15).

○ Interproscan으로는 총 154,209개의 gene이 annotation되었고 Blast2GO를 이용한 방식으로는 총 76,676개의 gene이 annotation됨.

○ KEGG 분석결과 총 2797개의 pathway로 분류됨.

○ Gene Ontology 분석에 의하면 (그림 40) Biological Process의 경우 metabolic process 및 cellular process가 각각 1, 2위를 나타내었고, Molecular Function에서는 binding 및 catalytic activity가, Cellular Component에서는 cellular anatomical entity 및 intracellular component가 1, 2위를 차지하였음.

표 15. Annotation 및 KEGG 분석 결과

Type	Number
Plant Refseq (NCBI)	135,438
Interproscan	154,209
B2G Annotation (GO annotated)	76,676
KEGG	2,797
Total transcripts	173,615

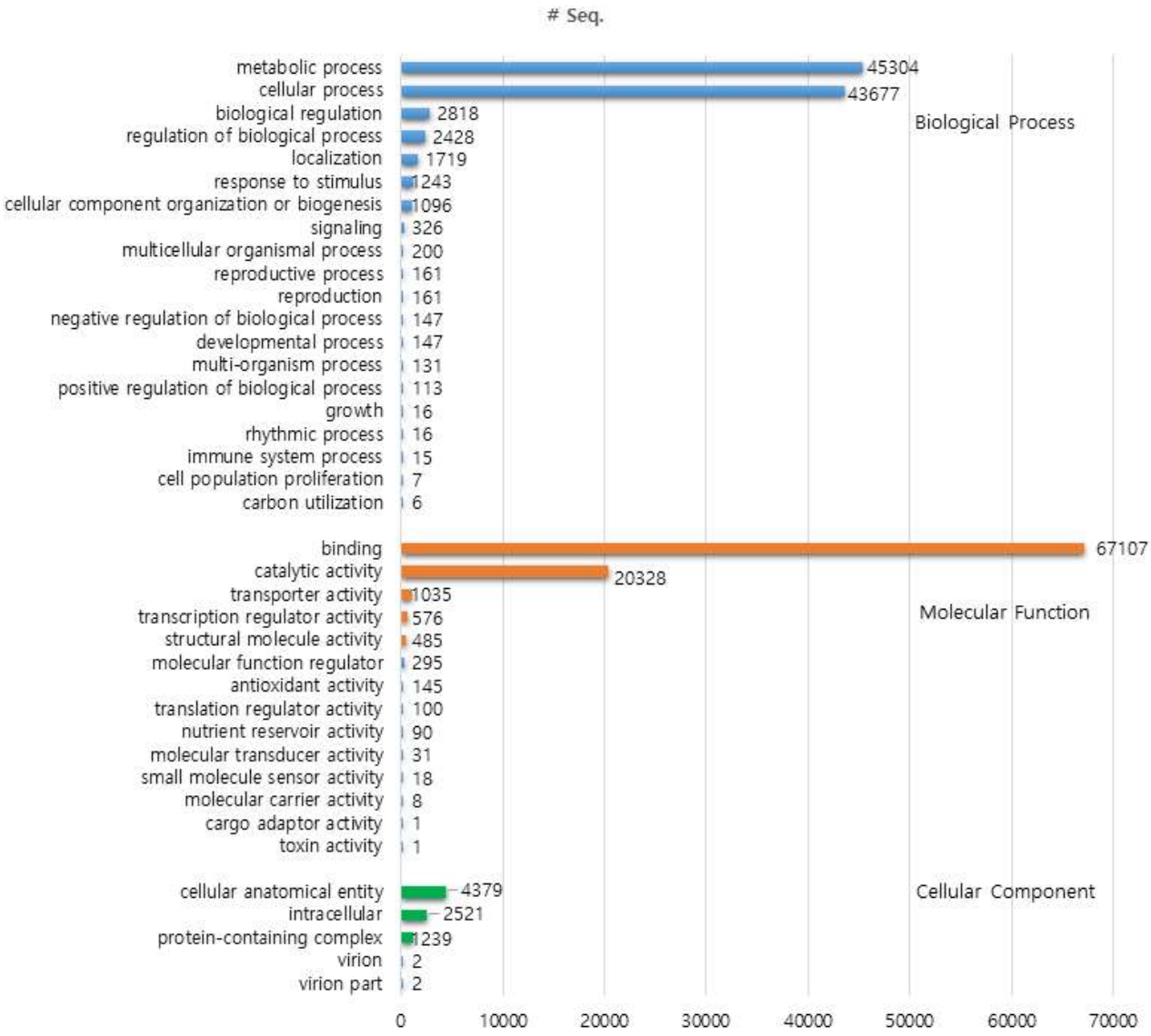

그림 41. 3개 category에 의한 Gene Ontology 분포

Ⅳ. 고 찰 및 결 론

Ⅳ. 고찰 및 결론

1. 독도 섬초롱꽃의 유전체분석 결과 및 고찰

가. Genome size

○ Flow-Cytometry 분석에 근거하여 울릉도산 섬초롱꽃은 1.4Gb의 유전체 크기를 가지며, 독도산 섬초롱꽃은 0.97Gb로 그 유전체 크기가 작은 것으로 나타냈다.

○ 독도산 섬초롱꽃의 유전체는 HiSeq data를 이용한 K-mer 분석으로 0.84Gb로 Flow-Cytometry 분석보다 130Mb 정도 작은 것으로 나타났고, PacBio와 HiSeq Combined Analysis을 이용해 각각 1.25Gb로 측정되었다.

○ 독도 섬초롱꽃 1.25Gb의 genome은 총 8,870개의 contig 및 3,995개의 scaffold로 조립되었고 이는 매우 고품질의 유전체 조립의 결과이며 본 연구는 PacBio 및 Linked-read sequencing 방식의 hybrid scaffolding 방식으로 단기간에 효율적인 유전체 염기서열 분석이 가능함을 보여주었다.

○ 독도 섬초롱꽃 genome의 크기는 울릉도 섬초롱꽃 genome의 잠정 크기인 1.4G (flow-cytometry 결과) 및 도라지 genome 크기인 700Mb (Kim *et al.* unpublished)와 상당한 차이가 나며 이는 진화과정 중 초롱꽃과 식물 (Campanulaceae)에서 genome size variation이 존재함을 예상할 수 있다.

○ 이러한 genome size variation의 기원으로 배수성 또는 repeat 지역에 의한 차이로 사료된다.

나. Genome content

○ 독도 섬초롱꽃 genome에서 Plant RefSeq DB를 사용할 경우 총 135,438개의 gene이 annotation 되었으며 비슷한 크기의 genome을 지닌 Soybean (Glycine max)가 46,430개의 protein-coding gene을 지닌 것을 비교하면 약 3배 많은 수이다.

○ 독도 섬초롱꽃 genome의 repeat 지역의 경우 genome size expansion에 중요한 역할을 하는 LTR이 전체 genome의 약 53%를 차지하는데 LTR은 soybean genome의 42%를, maize genome의 63%를 차지한다고 알려져 있다 (Schmutz *et al.* 2010).

○ 독도 섬초롱꽃 genome의 SSR은 총 196,182개가 발굴되었으며 이는 향후 집단유전분

석 및 진화적 유연관계 분석을 위한 마커개발의 source로 활용 가능하다.

○ 또한 유전체 염기서열 데이터는 향후 비교종의 re-sequencing을 통한 SNP 발굴 source로 활용 가능한다.

○ 유전자 해독 정보는 향후 유용 유전자 기능분석의 source로 활용 가능할 뿐 아니라 조절인자 발굴의 주요 source가 된다.

○ 유전체 해독 정보는 향후 RNAseq 분석시 mapping reference genome로서 사용가능하다.

다. 향후 추가 연구 방법

○ 향후 Optical Mapping (BioNano) 기술을 활용하여 독도 섬초롱꽃 genome 3,995개의 scaffold를 수백~수십개에 이르는 초고도화된 scaffold로 축소시켜 genomic rearrangement 및 ploidy event를 연구하는 데에 활용할 수 있다.

○ 향후 한반도, 울릉도, 독도의 초롱꽃과 섬초롱꽃의 comparative genomics를 활용한 유전체비교 진화연구가 가능하다.

○ Annotation된 gene 수에 대해 이미 알려진 식물들의 gene의 개수에 비해 많은 양이 도출된 것으로 파악되며 (Sterck L. *et al.*, 2007) 향후 cd-hit이나 transdecoder와 같은 tool로 redundant한 sequence를 제거하고 ORF가 확실히 나타나는 protein-coding region을 파악해야 할 것이다.

○ Conserved orthologous gene 분석으로 고등식물간에 보존되어 온 gene set 및 독도 섬초롱꽃 특이적인 gene의 존재를 조사하는 분석도 진행되어야 할 것이다.

2. 독도산 섬초롱꽃, 울릉도 섬초롱꽃과 초롱꽃

○ 초롱꽃(*Campanula punctata*)는 한국, 일본, 중국 및 시베리아에 널리 퍼져있는 종이며, 섬초롱꽃(*Campanula takesimana*)은 바다에 떨어져 있는 종으로 인식되어 있다. 독도의 섬초롱꽃은 울릉도에서 가까워 최근에 이식되었거나 울릉도에서 새를 통해 넘어 왔을 것을 고려해 섬초롱꽃으로 추정된다. 그러나, 본 종이 서식하는 서도의 물골에는 인위적 식재가 불가능한 암벽으로 이

루어져 있다. 이러한 관점에서 독도의 섬초롱꽃은 섬초롱꽃과 초롱꽃과의 유연관계를 추적할 필요가 있다.

○ 현재, 초롱꽃(*Campanula punctata*)과 섬초롱꽃(*Campanula takesimana*)은 엽록체 마커를 이용해 구분하고 있으나, 독도산 섬초롱꽃과 이들 종간의 구분이 용이 하지 않아 엽록체 마커이외의 다른 유전체 정보의 이용이 필요하다. 이러한 관점에서, 본 유전체 연구자료를 이용한 기술개발이 필요하다 하겠다. 이는 인근 지역의 같은 종과 근연종의 수집과 더불어 유용마커의 개발이 필요하다.

3. 섬생물 다양성 및 진화 분야 생물학적 난제 규명을 위한 향후 유전체 연구 방향

○ 독도를 비롯한 섬에서는 Genetic drift, 배수체 현상, 종의 합성이 흔히 일어난다 (Park *et al.* 2018). 초롱꽃(*Campanula punctata*)의 예로 알 수 있듯이, 타식성 식물이 생존을 위하여, 섬에서는 자식성 식물과 같은 유전현상을 나타낸다 (Inoue and Kawahara 1990). 이는 제한된 gene pool 유전자 손실을 대비 할 수 있는 수단으로 중요한 특징이고, 섬이란 특수 조건에서 제한된 gene pool에서 생기는 현상이 흔히 일어난다. 위와 같은 현상과 더불어 섬생물의 기원 추적이 용이하지 않다.

○ 이러한 현상을 추적하기 위하여, 식물간의 조직별 RNA의 유전적 변이를 비교하면 찾을 수 있으나, ① 조직별로 RNA를 추출할 수 있게 야생 근연식물을 모두 키우거나 개화기를 맞추어 모두 현지 채취(RNA later 이용)을 해야 하는 어려움이 있고, ② 개발된 RNA 마커를 추출이 용이한 DNA에서 확인된다는 보장이 없다. 이러한 사유에서 Low depth Genome sequence를 통한 SSR 마커를 개발하여, 유전다양성 분석에 이용한다. 이럴 경우, 유용마커 선발에 많은 시간과 노력이 필요로 한다.

○ 주요 작물의 경우, 표준 유전체 정보가 확립되어, 벼의 경우 수천개 벼계통의 GWAS를 이용하여 MAS(Molecular Assisted Selection)을 수행하여 2년내에 원하는 신품종을 만들고 있고, 고추를 비롯한 많은 원예작물은 대상 유전자를 근거로 한 마커가 실용화되어 있다.

○ 현실적으로 야생식물에 대한 농작물의 표준 유전체 수준의 유전체 분석에는 많은 부담이 뒤따른다. 본 연구진이 제안하는 섬생물 다양성 및 진화 분야 생물학적 난제 규명을 위한 향후 유전체 연구 방향은 섬자생식물과 근연 육지식물의 2종의 유전체 비교분석을 통한 유용마커 개발(유전자 정보를 가진 Targeted Marker)을 통한 섬자생식물의 진화를 추적하는 것이다. 이는 나노 포어 씨퀀싱(NanoPore Sequencing)을 포함한 새로운 기계의 도입으로 생각보다 적은 예산으로 가능할 것이다. 그러나, 이러한 연구는 유용 정보분석 자료 확보에 연구팀이 시간을 많이 투여해야 하므로, 단년 계약에 의한 연구를 진행한다면 그 효율과 완성도가 낮을 것으로 예상한다.

○ 본 연구에서 독도 섬초롱꽃의 유전체를 분석했으나, 다른 초롱꽃과 식물과의 비교 분석이 포함되지 않아 그 자세한 비교 분석이 필요하다. 독도초롱꽃의 기원을 추적한다는 관점에서 연구한다는 가정에서 보면, ① 한국, 일본, (러시아, 중국)의 초롱꽃(*C. punctata*) 식물체 확보, ② 울릉도 섬초롱꽃과 한반도와 일본의 초롱꽃의 Flow-Cytometry 측정을 통한 유전체 크기 추정, ③ 1~2종류 유전체 분석, ④ 독도산 섬초롱꽃, 울릉도 섬초롱꽃, 초롱꽃 유전체 비교분석 및 마커개발, ⑤ 마커를 이용한 독도초롱꽃의 진화 규명(유전체와 Targeted Marker를 통한) 의 단계적 연구를 진행하면, 엽록체 유전체 및 기존에 개발된 핵 유전자의 적용에서 해결되지 않는 섬생물 다양성 및 진화 분야 생물학적 난제인 독도 섬초롱꽃의 기원을 추적할 수 있을 것으로 판단된다.

Ⅴ. 참 고 문 헌

Ⅴ. 참고 문헌

Anthony Rhoads, Kin Fai Au, PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics. 2015. Vol. 13, pp. 278-289.

Chen-Shan Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 2013. Vol. 10, pp. 563-569.

Lauren Coombe *et al.*, Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data. PLOS. 2016. Vol. 11, e0163059.

de novo Assembly Solution, Product Brochure, https://www.10xgenomics.com/ 2017.

Robert M *et al.*, Using BUSCO to assess insect genomic resources, Methods in Molecular Biology, Insect Genomics, Humana Press, New York, NY 2019, Pages 59-74 (published online November 10, 2018)

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 2017. Vol. 27, pp. 722-736.

Chikhi R. *et al.*, Informed and automated k-mer size selection for genome assembly. Bioinformatics, 2014. Vol 30, pp. 31-37.

Yeo S. *et al.*, ARCS: scaffolding genome drafts with linked reads. Bioinformatics, 2018. Vol 34, pp. 725-731.

Hoff KJ, Stanke M, WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Research, 41, 2013, W123-W128.

Sterck L. *et al.* How many genes are there in plants (and why are they there)? Curr Opin Plant Biol. 2007. Apr;10(2) Epub 2007. pp. 199-203.

박선주 외. 2018. 독도 자생식물 보전 및 관리를 위한 유전자 분석 연구(4차년도)-2018년도 최종 보고서. 국립생물자원관

이정현 외. 2017. 독도 자생식물 보전 및 관리를 위한 유전자 분석 연구(3차년도)-2017년도 최종 보고서. 국립생물자원관

Park, Chong-Wook *et al.* Polyploidy and introgression in invasive giant knotweed (*Fallopia sachalinensis*) during the colonization of remote volcanic islands. Scientific Reports 2018. 8(1): 16021.

Ken Inoue and Takayuki Kawahara 1990. Allozyme Differentiation and Genetic Structure in Island and Mainland Japanese Populations of C*ampanula punctata* (Campanulaceae). American Journal of Botany, 1990. Vol. 77, No. 11 pp. 1440-1448.

Kim, Jungeun *et al.* (unpublished) Whole-Genome, Transcriptome, and Methylome Analyses of a Medicinal Plant *Platycodon grandiflorus* Provide Insights into the Evolution and Regulation of Platycoside, a triterpenoid saponin.

Schmutz *et al.* Genome sequence of the palaeopolyploid soybean. 2010. Nature Vol. 463 pp. 178-183.