

차세대 염기서열 결정장치

나노포어를 이용한 유전체 결정

3GGCAATAACGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACC



DATA FOR ILLUSTRATIVE PURPOSES ONLY

성신여자대학교 바이오생명공학과

전공기반 비교과 프로그램

**현대 생물학과 그 응용분야에 있어서
유전체 결정은
모든 다른 분야 연구와 기술개발의
기초자료가 됨!**

- 인간 유전체: 진단, 의약품개발, 맞춤형 의료, 단백질 기능연구...
- 식물, 동물, 미생물 유전체: 천연물의약품, 검역, 분류진화 연구...

생물학의 발달 과정

(현대적 의미의) 분류학 (Modern Taxonomy)

계통학 (Phylogenetics) 진화학 (Evolutionary Biology) 생물정보학 (Bioinformatics) 유전체학 (Genomics) ...

INTEGRATIVE BIOLOGY

유전공학, 생명공학 (Genetic Engineering Biotechnology) 면역학 (Immunology) 단백질체학 (Proteomics) ...

발생학 (Developmental Biology)
- 개체에서의 형태 형성 과정에 대한 연구

분자생물학 (Molecular Biology)
- 분자수준에서 생명현상을 이해함

Watson and Click DNA 구조 해독

유전학 (Genetics)
- 생물의 유전정보의 전달을 연구

생태학 (Ecology)
- 생물간의 상호작용을 연구

Mendel의 유전법칙 발견

생리학 (Physiology)
- 한 생물 내에서의 기능을 연구

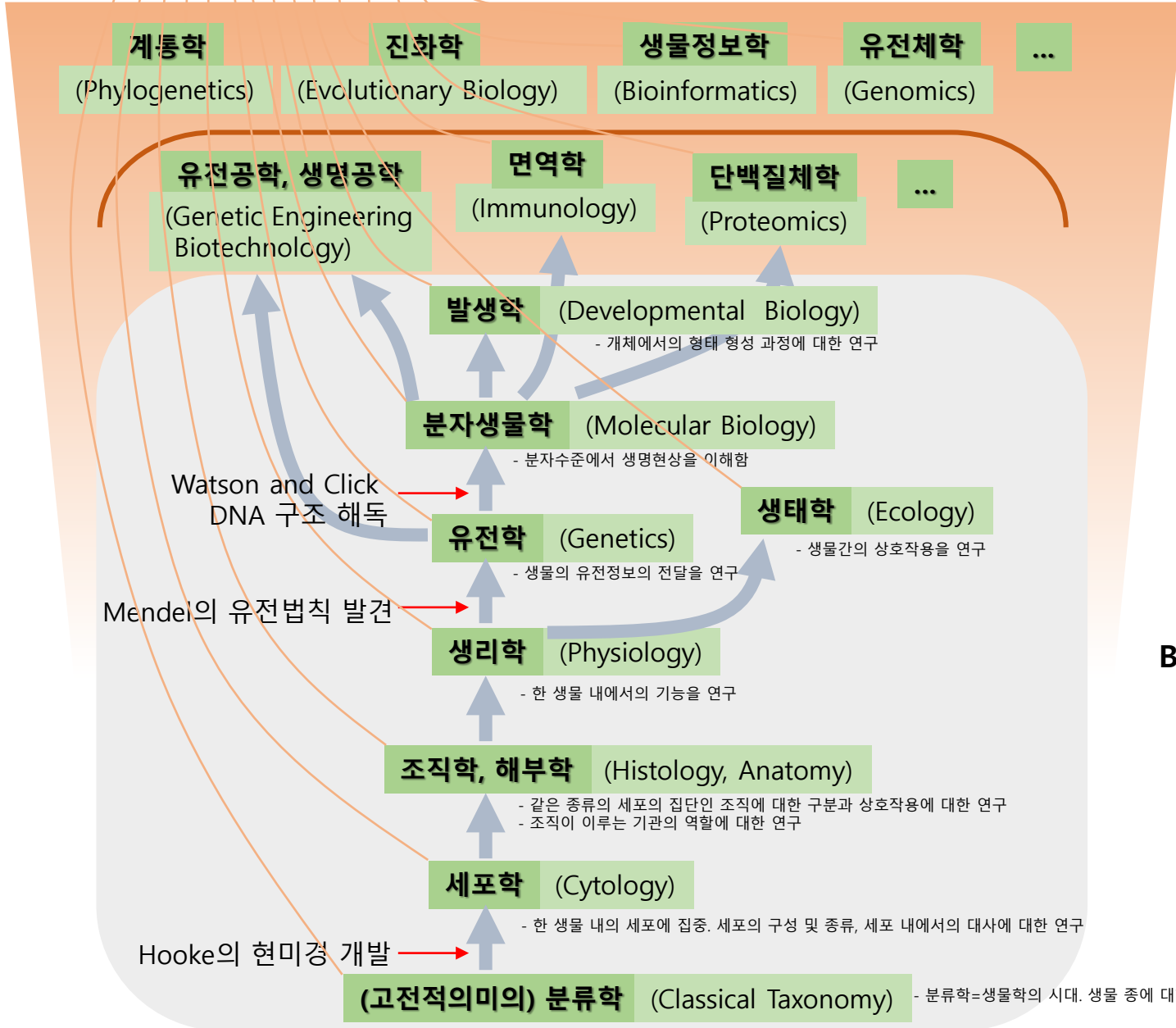
CORE BIOLOGY

조직학, 해부학 (Histology, Anatomy)
- 같은 종류의 세포의 집단인 조직에 대한 구분과 상호작용에 대한 연구
- 조직이 이루는 기관의 역할에 대한 연구

세포학 (Cytology)
- 한 생물 내의 세포에 집중. 세포의 구성 및 종류, 세포 내에서의 대사에 대한 연구

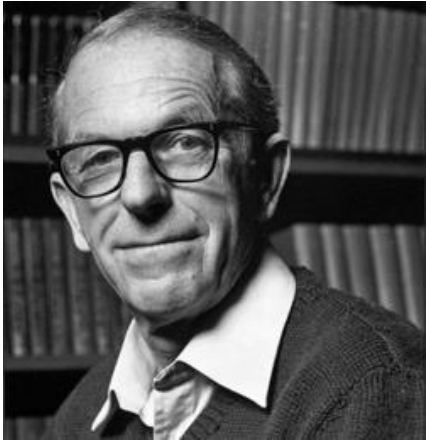
Hooke의 현미경 개발

(고전적의미의) 분류학 (Classical Taxonomy) - 분류학=생물학의 시대. 생물 종에 대한 탐험, 발견, 기재



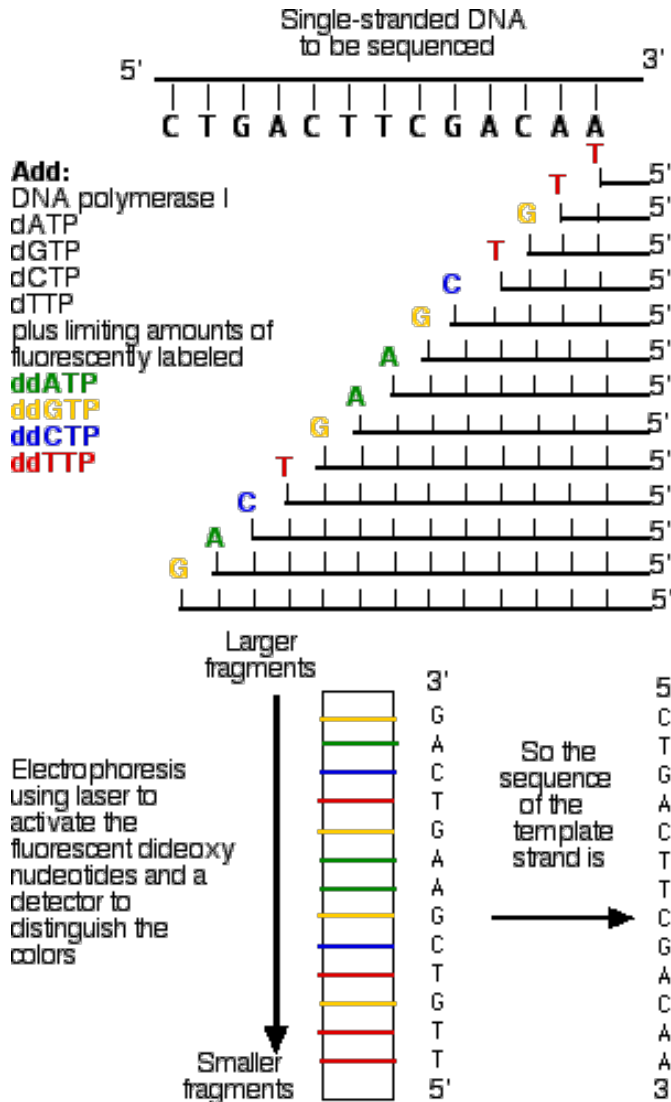
I. Introduction: 염기서열 결정의 진화

The first generation sequencing: Sanger sequencing



Frederick Sanger (영국)

- 두 번의 노벨상 수상자
- “termination” method
- One-dye four lane system에서 four-dye one lane system으로 발전.



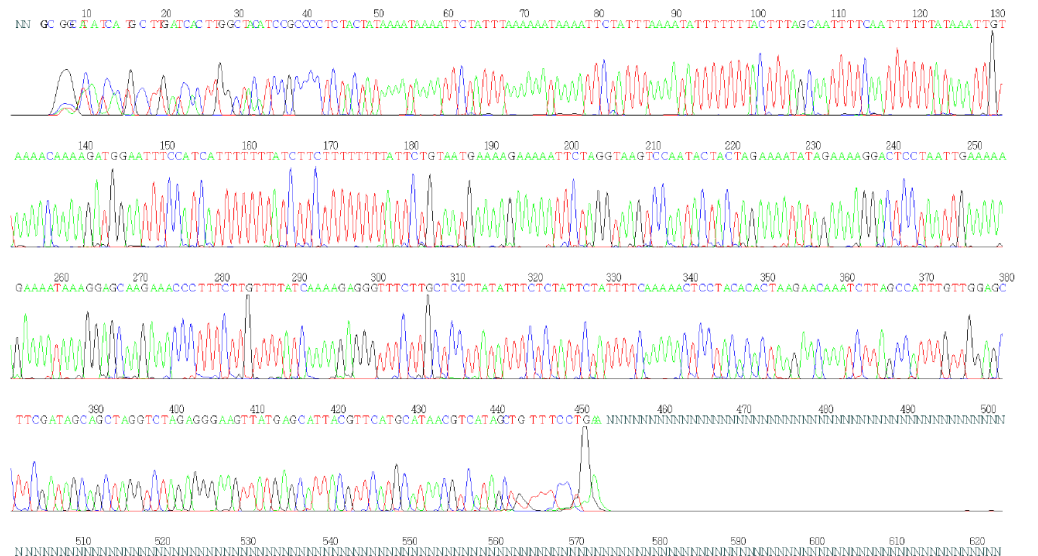
현재 가장 많이 쓰는 모델: ABI 3730 (Thermo Fisher)

- Applied Biosystems Co. → Thermo Fisher Co.
- Ca. 900 bp/capillary
- 384 capillary/run
- 900 X 384 = ca. 350 kbp



File: 102-M13F-pUC.ab1 Run Ended: 2008/7/2 5:13:25 Signal G:379 A:800 C:608 T:1045
Sample: 102_M13F-pUC Lane: 65 Base spacing: 14.219999 951 bases in 11477 scans Page 1 of 2

MACROGEN
Advancing through Genomics



Next Generation Sequencing (차세대 염기서열 결정)

1) The second generation sequencing:

- 1) emulsion based clonal amplification (**emPCR**)의 기술,
- 2) DNA 분자가 합성될 때 형광을 발하는 염기서열 결정기술 (**pyrosequencing**)
- 3) 광섬유들을 평행하게 붙여 만든 **pico-titer plate**
등의 신기술을 이용하여 염기서열 결정 용량은 획기적으로 증가.

대표적 기업 / 기술:

Roche (454) / 454

Solexa / Illumina → MGI (Illumina 유사 기술에 의한 중국 기업/제품)

ABI / SOLiD

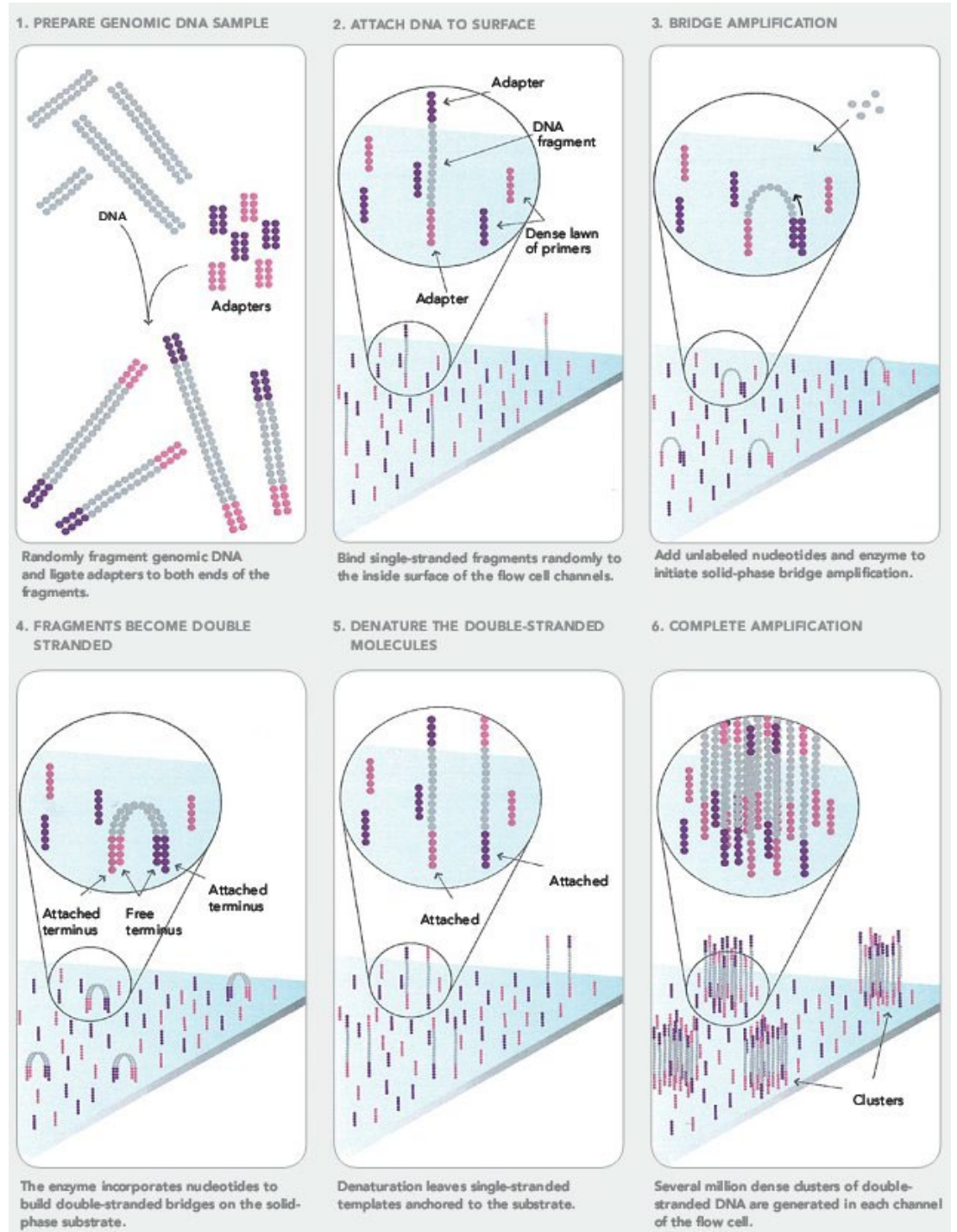
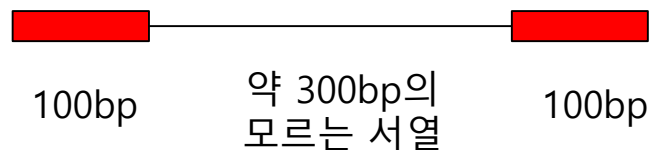
Hilicos / Ion Torrent

→ 대부분 없어지고, 현재 Illumina (and MGI)가 가장 널리 쓰임.

Solexa / Illumina Technology

Illumina를 이용한 결과의 특징:

- 한 가닥의 DNA로 부터 염기서열을 결정할 때 순방향과 역방향으로 각각 약 100~150 bp 정도 읽게 됨.
- 그러므로 DNA를 일정크기로 잘라 만든 조각이 500bp일 경우 좌우로 100bp씩을 얻게 되고,中间的 300bp는 모르는 서열로 연결되게 됨.

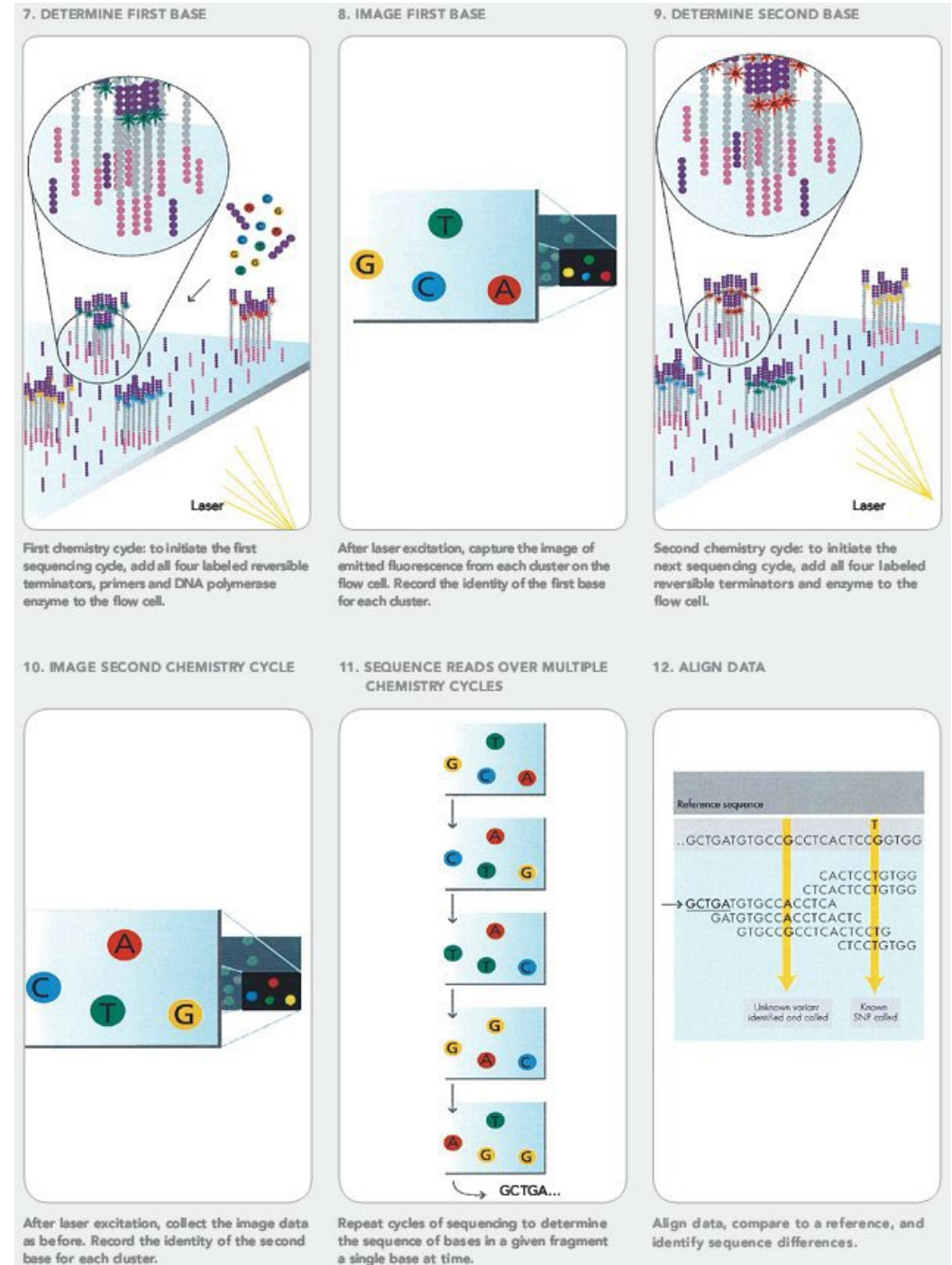


Solexa / Illumina Technology

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

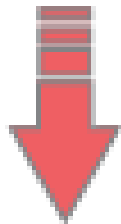
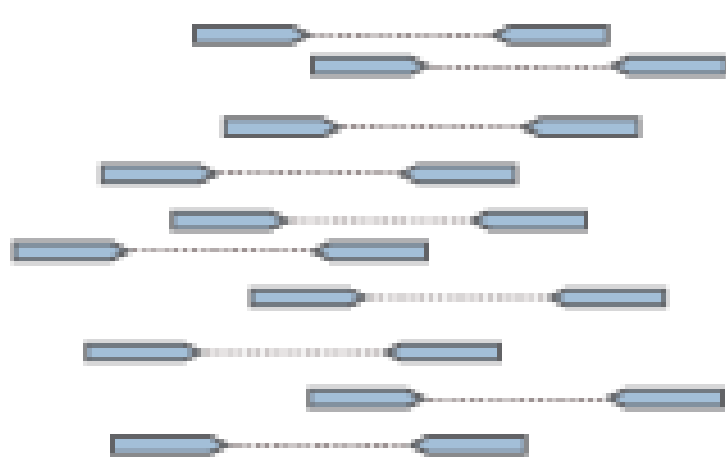
※ Illumina 기술에 의한 염기서열 결정은 “매우” 많은 염기서열을 한번에 얻을 수 있음. 그러므로 종종 많은 시료를 섞어서 염기서열을 결정하기도 한다. 이때 섞은 시료들을 구분하기 위하여 시료 각각을 구분하는 index sequence 를 각각의 시료에 붙임. 전체 염기서열 결정 후 index sequence로 시료들은 구분한 후 각각 정렬하여 결과를 얻게 됨.

※ **index sequence**: 여러 시료를 섞어 실험할 때 각각의 시료를 구분할 수 있는 짧은 염기서열. 실험의 첫 단계인 adaptor를 붙이는 과정에 index sequence를 삽입한다.

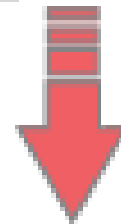
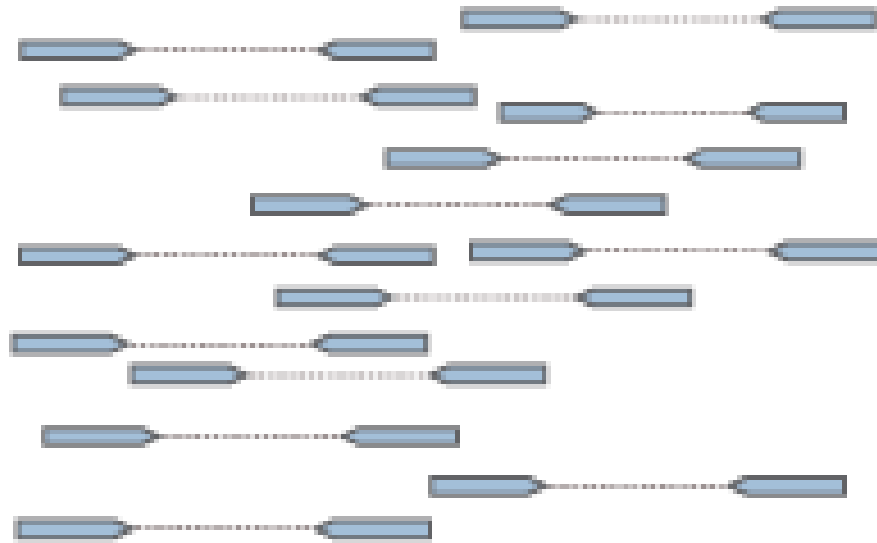


Illumina data 의 assemble 과정

Paired-End Reads



Assembled Contig



Assembled Contig

- Illumina sequence는 최근까지 유전체 결정에 있어서 주된 data로 사용되어 옴.
- Macrogen Co. service:
HiSeq X ten, HiSeq4000, NextSeq
- 진정한 인간유전체 결정 \$1,000 시대



HiSeq X System Performance Parameters	
Key Application	Large Whole-Genome Sequencing (human, plant, animal)
Output per Run	Dual flow cell: 1.6-1.8 Tb
Single Reads Passing Filter	Dual flow cell: 5.3-6 billion
Maximum Read Length	2 x 150 bp
Run Time	< 3 days
Quality	≥75% of bases above Q30 at 2 x 150 bp

하지만... 무엇이 문제인가?

→ 100~150 bp의 너무 짧은(short-read) 서열을 제공하고 있기 때문에, 이것을 이어 붙여 유전체를 조립하면 구조적인 에러를 발생할 확률이 높아짐.

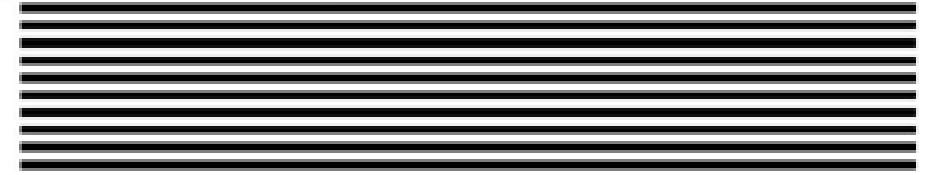
1) Mate-pair sequencing에 의한 해결

2) **Long-read data**를 생산하는 것이 궁극적 해결 방법임.

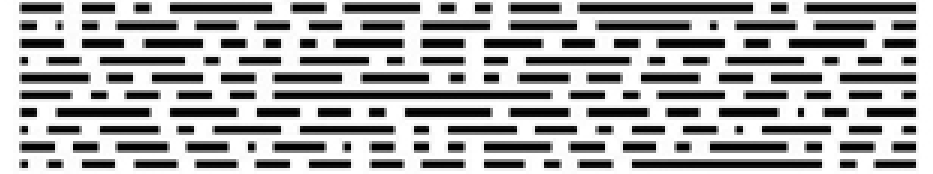
Genome assembly의 과정 (Illumina)

- 1) 추출된 DNA를 적당한 크기로 자른다.
- 2) 잘린 DNA 절편들 중 약 500 bp 의 크기를 갖는 절편들만 정제한다.
- 3) Illumina **paired-end sequencing**에서는 잘린 절편의 양쪽 끝의 각각 약 100bp 정도의 데이터를 얻을 수 있다(reads).
- 4) Assembly program을 이용하여 각각의 read들을 정렬하여 **contig**들을 생성해 낸다.
- 5) 이와는 별도로 추출된 DNA를 보다 긴 길이의 절편으로 잘라 (1K, 2K, 5K...) 이들 절편의 양쪽 끝 100bp 의 염기서열들을 결정한다 (**mate-pair library**).
- 6) 일반 paired-end sequencing을 통해 생성된 contig들과 mate-pair 결과를 합쳐 assembly한다. 이를 통해 결과적으로 전체 유전체는 부분적으로 염기서열이 완성된 contig들이 mate-pair 에 의해 연결되어 많은 gap을 포함하지만 서로의 위치관계가 명확해진 긴 염기서열을 얻게 된다. 이를 **scaffold** 또는 **super contig**라 한다

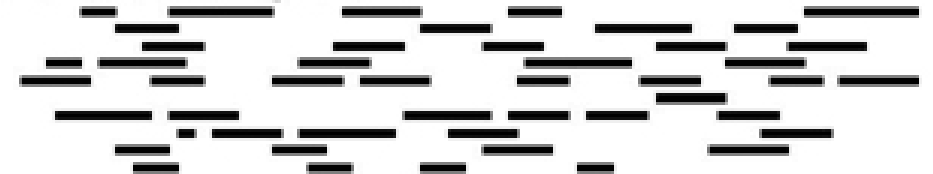
a) Multiple copies of genome



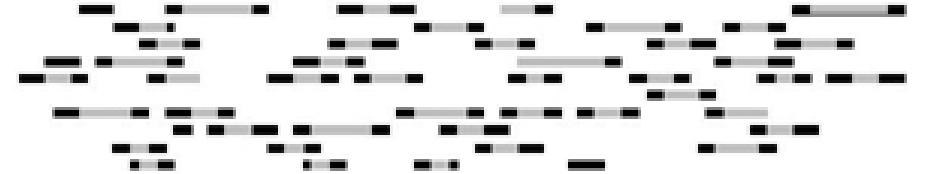
b) Sheared random fragments



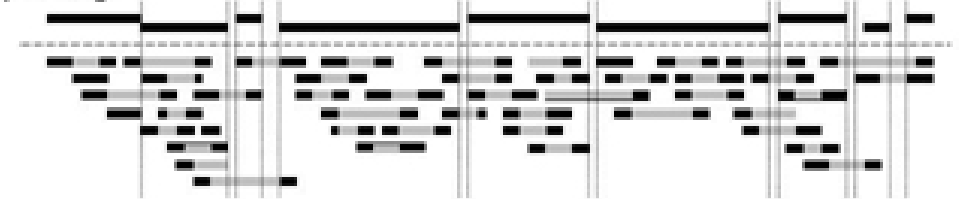
c) Size fractionated fragments



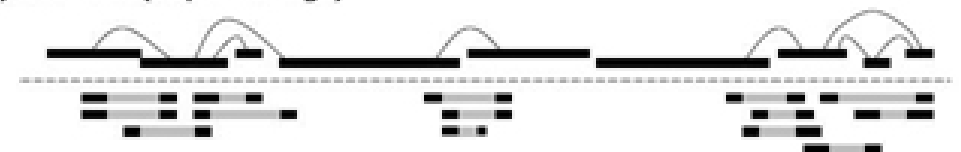
d) Reads



e) Contigs



f) Scaffolds(Super contigs)



Next Generation Sequencing (차세대 염기서열 결정)

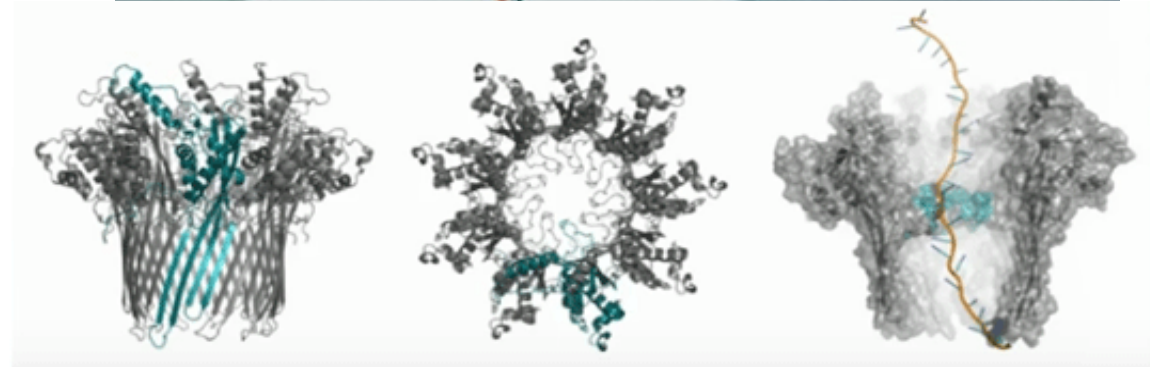
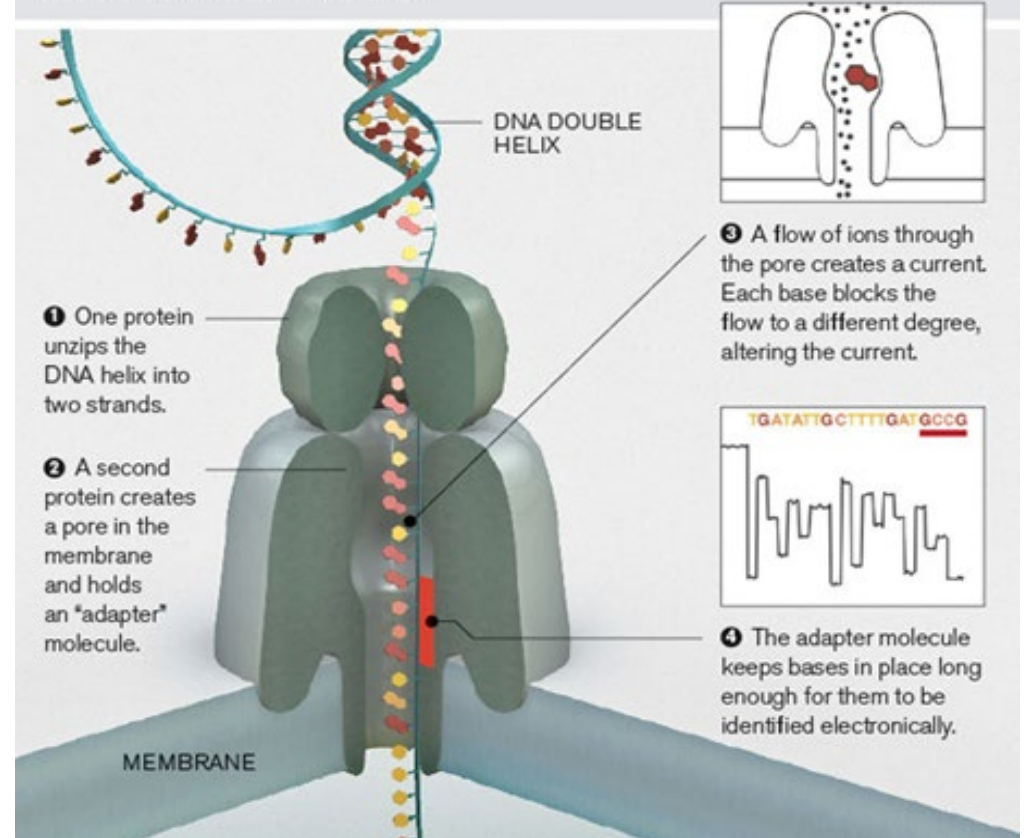
2) The Third generation sequencing:

- 100-150 bp 길이의 short-read를 생성해 내는 2세대 염기서열결정장치와는 달리 수십 kbp 에 달하는 long-read를 생성해 내는 기술
- 대표적 기업 / 기술:
 - 1) PacBIO / PacBioHiFi
 - 2) **Oxford Nanopore Technology (ONT)** / Nanopore
- **PacBioHiFi** 주로 10~15 kbp (up to 25 kbp)의 제한된 길이의 대용량 데이터를 생산하여 고가의 장비로 일반적으로 sequencing service 업체에 의해 운영됨.
- **Nanopore**는 연구실 단위의 실험이 가능하고, 추출된 DNA의 길이만큼 훨씬 긴 서열의 생성이 가능함.

ONT / Nanopore technology

- **Oxford Nanopore Technologies** (ONT): 2005년에 University of 로부터 설립.
- “Nanopore (나노포어)”라는 막단백질을 이용한 획기적인 NGS.
- 막단백질의 직경이 수 나노미터를 이루기 때문에 이런 이름이 붙여짐.
- 세포에서는 막에 위치하는 나노포어를 통해 이온들이 막 내외로 이동하는데, 이온이 지나가면서 막에 전류를 발생시키게 됨. 그런데 나노포어로 DNA나 RNA를 통과시키면, 이들이 이온의 흐름을 방해하여 여기서 발생하는 전류에도 변화가 생기게 되어 이런 막 단백질의 전류의 변화를 측정하는 것임.
- 즉, A, C, G, T 각 염기서열마다 이온의 흐름을 방해시키는 정도가 다르고 이에 의해 전류가 변화하는 정도도 달라지게 되므로, 이를 분석하여 서열을 파악할 수 있음.

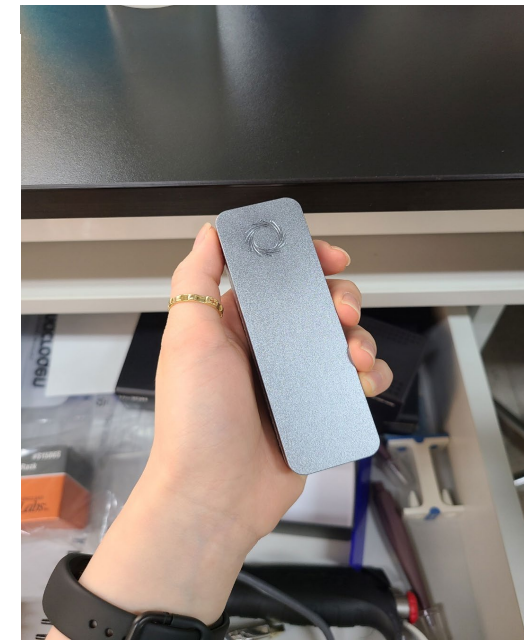
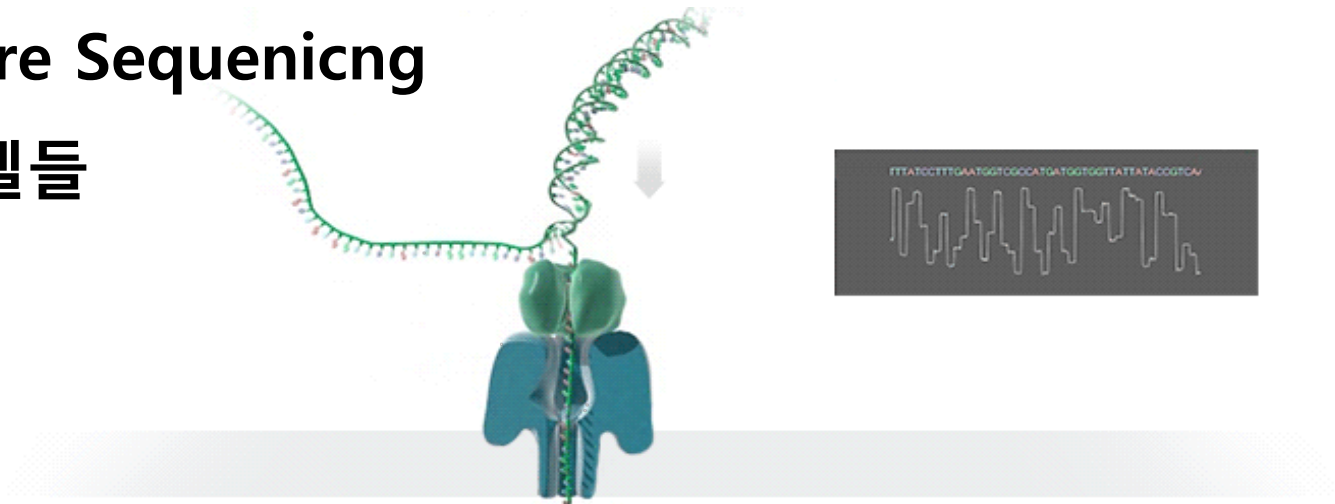
DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



ONT 시퀀싱에 사용되는 나노포어인 CsgG의 구조

Nanopore Sequencing

기기 모델들



가장 작은 model인 MinION은 노트북에 연결하여 현장사용도 가능하다.



MinION
Mk1C

Available to pre-order

Complete sequencing, analysis, and viewing device

Up to 30 Gb data / flow cell
512 channels*



MinION
Mk1B

Commercially available

Portable, USB powered biological analysis

Up to 30 Gb data / flow cell
512 channels*



GridION
Mk1

Commercially available

Five flow cell capacity and integrated computing

Up to 150 Gb (5 x 30 Gb) data / device
5 x 512 channels*



PromethION
P24 P48

Commercially available

High-throughput, versatile benchtop system (P24 or P48)

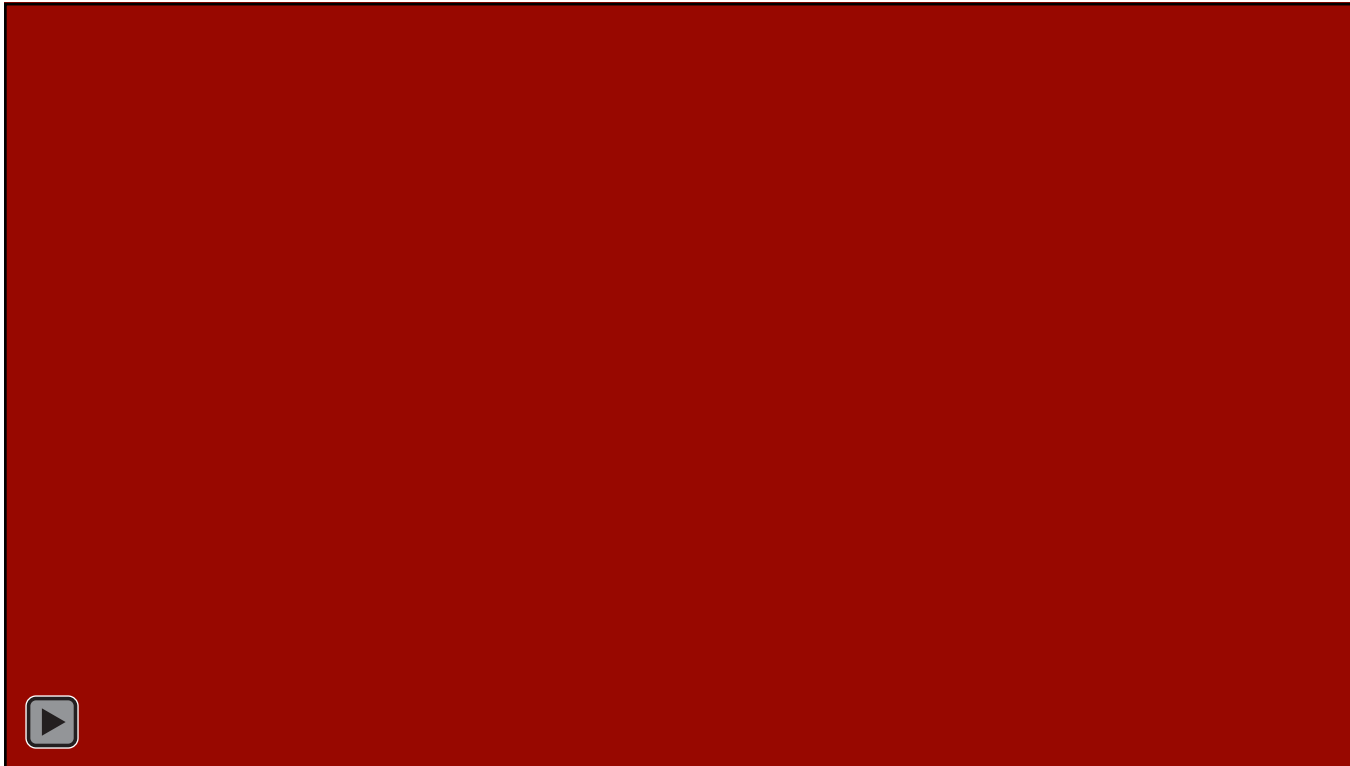
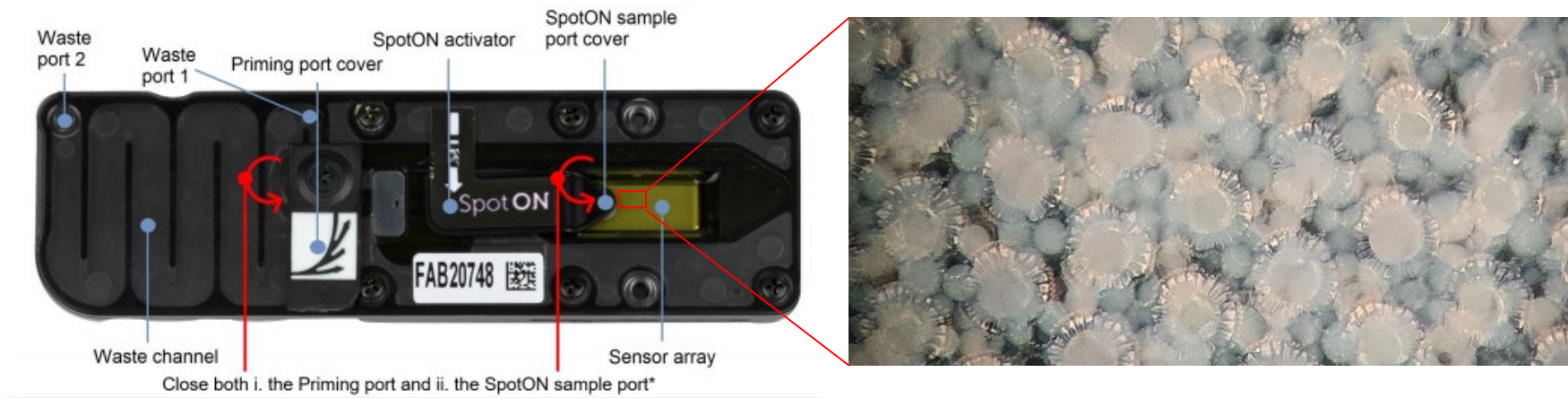
P24
>3.5 Tb data / device
24 x 3,000 channels*

P48
>7 Tb data / device
48 x 3,000 channels*

Flongle

Adapter for MinION/GridION, supports smaller single-use flow cells. Up to 1.8 Gb currently; towards 3 Gb.





Molecular weight
T: 126.11 g/mol
A: 135.13 g/mol
G: 151.13 g/mol
C: 190.24 g/mol

Summary:

Comparison Sequencing Platforms

	Platform	Amplification Method	Sequencing Method	Detection Method	Average read length	Errors
○	Illumina (HiSeq, MiSeq etc)	bridge PCR	sequencing by synthesis	Light	100-200 bp	~ 0.1 %
×	Life Tech Ion Torrent / Proton	emulsion PCR	Ion semiconductor sequencing	pH	200-400 bp	~ 1 %
×	Roche 454	emulsion PCR	Pyrosequencing, cleavage of released pyrophosphate	light	700 bp	~1 % High error rate in homopolymer
×	Life Tech SOLiD	emulsion PCR	sequencing by ligation of hybridizing labeled oligos	light	100 bp	~ 0.1 %
○	Pacific Biosciences PacBio	No amplification, single-molecule sequencing	polymerase incorporating colored NTPs	light	4.5 – 8 kb	>10 %
○	Oxford Nanopore MinION	No amplification, single molecule nanopore sequencing	DNA molecule traverses pore	current	> 5.4 kb	< 5 %

Further reading, great lecture: Sequencing technology - Past, Present and Future, http://www.molgen.mpg.de/899148/OWS2013_NGS.pdf

× 현재 사장되어가는 기술들

II. 실험 스케줄

주차별 주제	강사	날자	시간	내용	참여예상 학생수
High-molecular weight DNA 추출	김상태	8월 1일 (목)	9:00~12:00	- DNA 추출 이론 강의 - DNA 추출 1단계	10
			13:00~16:00	- DNA 추출 2단계	10
		8월 2일 (금)	9:00~12:00	- 전기영동 및 Qubit 정량 - Nanopore 이론 강의 - 길이측정 분석의뢰	10
Nanopore running	(주)필코리아	8월 8일 (목)	9:00~12:00	- 라이브러리 작성 실습	10
			13:00~16:00	- Nanopore running	10
		8월 14일 (수)	9:00~12:00	- 결과 생성 및 결과 분석	10

(주) 필코리아: Oxford Nanopore의 Korean distributor

- 조교: 식물분자계통학실 서정우 (석박통합과정), 이지은 (석사과정), 서지예 (4학년)
- DNA 추출 실험: 3개 조로 진행.

- 실험 재료

1조: 목련 (*Magnolia kobus* DC.)

- Basal angiosperm
- 한국, 일본에 분포하지만, 한국에서는 제주도에 약 1,000여 개체 만이 존재하는 멸종위기종임.
- 현재 성신여대 식물분자계통학실에서 유전체결정 프로젝트 진행 중

2조: 통보리사초 (*Carex kobomugi* Ohwi)

3조: 나도별사초 (*Carex gibba* Wahlenb.)

- 단자엽식물. 특수한 꽃구조를 갖는 식물군.
- 현재 성신여대 식물분자계통학실에서 Carex chloroplast genome project 진행 중

	이름	학년
1조	박도율	4-1
	박수현	3-1
	서예원	1-1
2조	원유진	3-2
	윤수민	4-1
	이가은	3-1
3조	이지나	3-2
	정예림	4-1
	최소윤	4-1

- 성신여대 식물분자계통학실에서 지속적으로 분류, 계통, 진화, Evo-Devo에 대한 연구를 진행하고 있는 중임.

- 2016년 Illumina sequencin에 의해서 ver. 1.0 genome이 완성되었지만, 완성도가 너무 낮아 논문으로 발표되지 못한 상태임.

- 예상 유전체 크기: 1.44 Gbp (flowcytometry)
- Ver. 1.0의 assembly: 2.6 Gbp

# of scaffolds	5,462,029
Total length (bp)	2,579,648,040
Average length (bp)	472.29
N20 (bp)	20,081
N50 (bp)	2,595
N90 (bp)	146
Maximum length of scaffolds (bp)	65,505

→ long-read data의 추가가 필요함!

Genome of *Magnolia kobus* (Magnoliaceae), an Additional Key Reference for the Studies of Angiosperm Evolution: the first version

Jongsun Park^{1,2}, Jong Bhak³, and Sangtae Kim^{1*}
¹Sungshin Univ., Seoul, Rep. of Korea; ²InfoBoss. Co., Ltd., Seoul, Korea; ³UNIST, Ulsan, Rep. of Korea

InfoBoss UNIST
Information Systems

Why Magnolia?

- Magnolia has received keen interests from many botanists because it is believed one of early-diverging angiosperms in the evolutionary history of angiosperms.
- Although recent molecular phylogenetic studies showed that Magnolia is placed at the "Magnoliids" (sister to eudicots + Ceratophyllales/Chloranthales (below tree), clarification of the feature of genome in Magnolia will provide a key to understand the evolutionary gap between basal angiosperms and eudicots.

Summary of angiosperm phylogeny based on recent molecular phylogenetic studies (Soltis et al., 2012; Zhang et al., 2014) and the placement of Magnolia in the phylogenetic tree (a red arrow). Blue letters indicate genes of 92 angiosperm genomes which is published or is redownloadable on the public web sites until now.

InterPro annotation

- Predicted genes were annotated using InterPro Scan (Quinlan et al., 2005).
- Detection of high number of genes are expected because the filtering process of repetitive sequences has not been done yet.

AP2/ERF family

Number of putative AP2/ERF genes from three genomes

Species	# of genes	Proportion (%)
M. kobus	130	0.25%
Am. trichopoda	57	0.21%
Ar. thaliana	107	0.44%

- M. kobus has around four times of AP2 TFs in comparison to Am. trichopoda.

Length distribution of AP2/ERF genes from three genomes

- 6 AP2 TFs (5%) from M. kobus are shorter than those of Am. trichopoda and Ar. thaliana, which may be caused by partially predicted genes.
- Am. trichopoda and M. kobus have long AP2 genes (>500 aa).

Comparison of gene contents among Magnolia, Arabidopsis, and Amborella

- Three genomes showed quite different pattern of gene contents based on an analysis of functional domains.

Cytochrome P450 family

Number of putative P450s from three genomes

Species	# of genes	Proportion (%)
M. kobus	198	0.33%
Am. trichopoda	266	0.25%
Ar. thaliana	248	0.20%

- M. kobus has more cytochrome P450s than those of the two genomes.

Length distribution of cytochrome P450s from three genomes

- M. kobus AP450s of which length is less than 100 aa seem not to be complete cytochrome P450 genes in comparison to those of Ar. thaliana.
- However, Am. trichopoda also have small P450 genes like M. kobus.

Overall process for understanding M. kobus genome

Gene cluster analyses of three genomes

- Most of genes from M. kobus are M. kobus-specific (428,571; 90%).
- About half genes in the common cluster are from M. kobus.

Sequencing method and statistics

Library	Total size (bp)	Total reads	Total size (bp)	Total reads
150bp PE library	8,832,295,460	87,349,202	8,655,609,472	86,353,047
100bp PE library #1	9,322,078,376	90,317,576	8,511,178,262	86,187,890
100bp PE library #2	35,776,776,772	348,020,572	34,867,209,018	343,392,694
100bp PE library #3	39,139,076,180	387,348,795	38,772,283,213	386,779,817
100bp PE library #4	41,058,344,750	406,528,158	38,954,617,377	384,029,243
50 bp MP library	6,849,090,776	65,792,578	5,794,097,048	58,328,845
50 bp MP library	6,857,042,916	65,712,376	5,792,646,219	58,325,575
50 bp MP library	6,095,094,644	60,005,444	5,205,325,423	54,521,556
Total	157,098,753,800	1,584,794,930	141,195,896,524	1,492,424,589

- We sequenced genomic DNA of M. kobus using Illumina HiSeq2000; 92.46% raw sequences were filtered out around 1.44Gbp filtered data were used for genome assembly.
- Additional libraries are included: 1) 180bp pair-end and 2) 10k and 15k mate pair libraries.

Analyses of selected gene families

- MADS-box genes**
 - encodes the DNA-binding MADS domain, generally transcription factors
 - involved in controlling all major aspects of development, including gametophyte development, embryo/seed/oot/fruit development
- AP2/ERF genes**
 - encodes the DNA-binding ERF/AP2 domain, generally transcription factors
 - one class of AP2/ERF genes has one domain and another class has two domain
 - play a central role in the establishment of the floral meristem, the specification of floral organ identity, the regulation of floral homeotic gene expression, and etc. in plants
- Cytochrome P450 (CYP)**
 - genes related to plant secondary metabolites
 - related to the superfamily of proteins containing a heme cofactor
 - terminal oxidase enzymes in electron transfer chains
 - Plant CYPs are related to biosynthesis of secondary metabolites including flower color

Conclusion and further researches

- As a taxon included in one of major lineage for understanding the evolution and diversification of angiosperms, we started a genome project of Magnolia (M. kobus).
- Preliminary version of an assembly is >2.5 Gbp and the size of the genome is higher than our expectation.
- Although it is the first version of assembly, we may speculate the following things: 1) genome size is larger than that of Am. trichopoda, 2) number of M. kobus genes in common clusters are two times than that of Ar. trichopoda, 3) number of genes in MADS-Box, Cytochrome P450, and AP2/ERF families are higher than those in Am. trichopoda.
- In the future version of assembly which containing more data, we may expect solutions for the key questions to understand the evolution and diversification from basal angiosperms to eudicots such as 1) location of whole genome duplication (WGD) events during the angiosperm evolution and 2) specification of novel genes for eudicots.

Gene prediction

- Total 476,013 genes were predicted by complete gene model of AUGUSTUS (Stanke et al., 2008), although current prediction is based on a preliminary version (ver. 0.1) of the assembly.

Species	# of genes	Average length (aa)
M. kobus	476,013	182.9
Am. trichopoda	314,466	230
Ar. thaliana	466,466	230

MADS-box family

Number of putative MADS-box genes from three genomes

Species	# of genes	Proportion (%)
M. kobus	79	0.16%
Am. trichopoda	34	0.13%
Ar. thaliana	108	0.23%

Length distribution of MADS-box genes from three genomes

- There are 29 MADS-box genes of M. kobus of which length is less than 100 aa, while Am. trichopoda has only 5, indicating that these MADS-box genes are probably partially predicted.
- Long MADS-box gene was found in M. kobus and Ar. thaliana.

REFERENCES

Loew et al., ScallopMDS: an empirically improved memory-efficient short-read de novo assembler. *Bioinformatics* 11 (2012): 18.
 Quinlan et al., InterProScan: protein domain identifier. *Nucleic acid research* 33 (2005): W118-W120.
 Soltis et al., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704-750.
 Stanke et al., Using native and symmetrically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24 (2008): 837-844.
 Zhang et al., 2014. Reconstruction of tree angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature communications* 5: 1-10. DOI: 10.1038/ncom08556.

N50

- 일반적으로 유전체 수준에서 assembly가 잘 되었는지 안되었는지는 **N50** 값으로 제시함.
- 전체 contig 들을 크기순으로 배열하여 큰 것으로부터 크기를 차례로 더하여 더한 값이 전체 유전체 크기의 50%를 넘는 순간의 contig의 크기를 N50라 함.
- N50 는 평균값 또는 중간값과는 다른 의미의 수치임!!!
- 3 3 4 6 7 8 8 9 9 9 10 11 13 25 의 길이의 contig들이 있을 때

$$\text{Mean} = 125/14 = 8.93 \quad (\text{sum}=125)$$

$$\text{Median} = (3+25)/2 = 14$$

125/2= 62.5 sum of contig lengths reach to 62.5 (from the largest to the smallest)

$$25+13+11+10=59$$

$$25+13+11+10+9=68$$

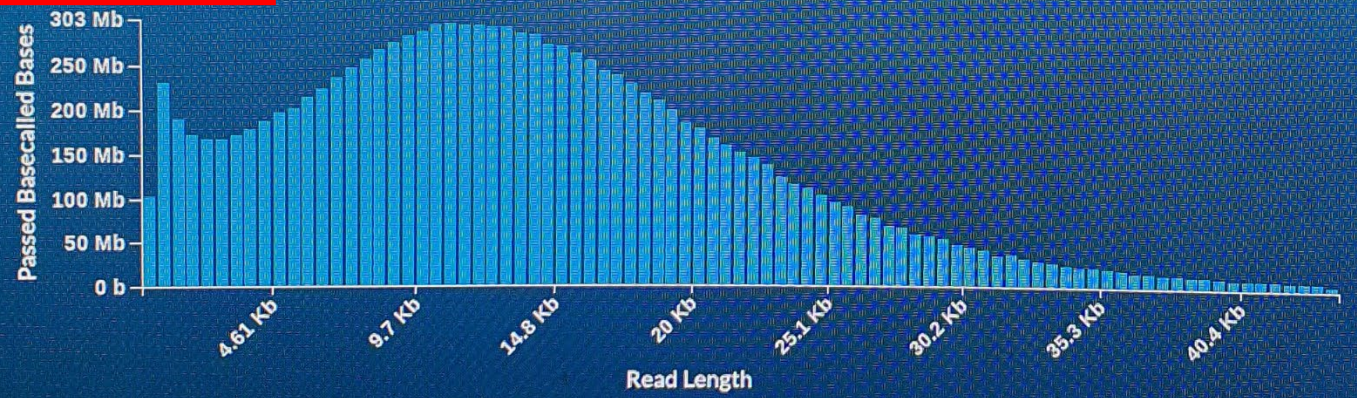
Therefore, **N50** = 9

Position	Flow cell ID	Sample ID	Health	Run time	Run state	Reads	Bases	Baseca	
<input type="checkbox"/>	MN35716	FAO37153	no_sample		44.8 H / 72 H	Active	2.52 M	14.9 Gb basecalled 16.32 Gb estimated	100%

Scroll right >

Read length histogram

Estimated N50: 13.15 Kb



Read Bases

Lengths Counts Estimated Basecalled

Hide outliers Split by read end reason



- 사용할 Oxford Nanopore의 flowcell 및 시약: **MinION Starter Pack**

store.nanoporetech.com/minion-basic.html

OXFORD NANOPORE Technologies

PRODUCTS APPLICATIONS **STORE** RESOURCES SUPPORT ABOUT


HOME / STORE / DEVICES

< **Configure your package**


SELECT PACKAGE **FLOW CELLS** SEQUENCING KITS TRAINING CONFIRM

Select your flow cell type Continue >

1x Flow Cell (R10.4.1)


 R10 is our nanopore chemistry designed to deliver highest consensus accuracy. Paired with the Kit 14 chemistry, R10.4.1 generates data at a modal accuracy above 99%.
Note: R10.4.1 flow cells currently require Kit 14 chemistry.
Product lead time: 1 week Select

1x Flow Cell (R9.4.1)

 The MinION and GridION Flow Cell contains up to 512 nanopore channels for sequencing DNA or RNA in real-time. Select

Selected package

MinION Starter Pack StarterPack



Total: \$1,000.00

Chat icon

그런데...

재료가 좋아야 좋은 결과가 나옴.

→ 일반 DNA 추출과정으로는 조각난 DNA가 추출됨.

→ 그러므로 되도록이면 "긴" DNA를 추출하는 것이 성공적인

Nanopore sequencing의 1차적 관건임!

※ Long DNA = **High-Molecular Weight (HMW) DNA**

II. 식물체로부터의 HMW DNA 추출

최근 성신여자대학교 식물
분자계통학실에서 개발하
여 출판된 식물 HMW
DNA 추출 방법
(2023년 6월 공식 출판)

본 프로그램의 첫 part는
이 논문의 protocol에 준
하여 수행 됨.

Applied in Plant Sciences
의 DNA추출을 위한 특별
호에 게재됨.







<https://bsapubs.onlinelibrary.wiley.com/toc/21680450/2023/11/3>

Received: 26 September 2022 | Accepted: 4 May 2023

DOI: 10.1002/aps3.11528

PROTOCOL NOTE

High-molecular-weight DNA extraction for long-read sequencing of plant genomes: An optimization of standard methods

Myoungbo Kang¹  | Andre Chanderbali²  | Seungyeon Lee¹  |
Douglas E. Soltis^{2,3}  | Pamela S. Soltis²  | Sangtae Kim¹ 

¹Department of Biotechnology, Sungshin Women's University, Seoul 01133, Republic of Korea

²Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA

³Department of Biology, University of Florida, Gainesville, Florida 32611, USA

Correspondence

Sangtae Kim, Department of Biotechnology, Sungshin Women's University, Seoul 01133, Republic of Korea.
Email: amborella@sungshin.ac.kr

This article is part of the special issue "Emerging Methods in Botanical DNA/RNA Extraction."

Abstract

Premise: Developing an effective and easy-to-use high-molecular-weight (HMW) DNA extraction method is essential for genomic research, especially in the era of third-generation sequencing. To efficiently use technologies capable of generating long-read sequences, it is important to maximize both the length and purity of the extracted DNA; however, this is frequently difficult to achieve with plant samples.

Methods and Results: We present a HMW DNA extraction method that combines (1) a nuclei extraction method followed by (2) a traditional cetyltrimethylammonium bromide (CTAB) DNA extraction method for plants with optimized extraction conditions that influence HMW DNA recovery. Our protocol produced DNA fragments (percentage of fragments >20 kbp) that were, on average, ca. five times longer than those obtained using a commercial kit, and contaminants were removed more effectively.

Conclusions: This effective HMW DNA extraction protocol can be used as a standard protocol for a diverse array of taxa, which will enhance plant genomic research.

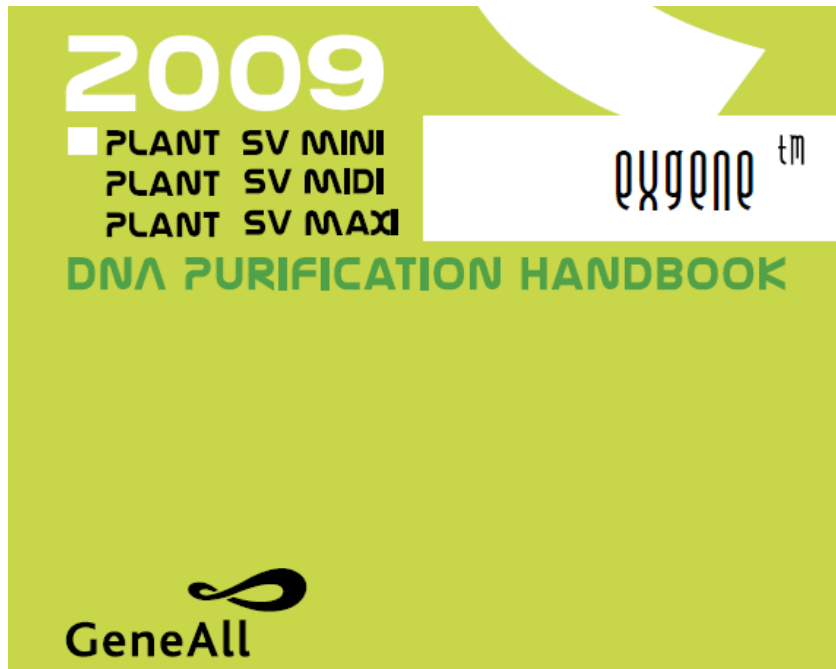
KEYWORDS

CTAB, DNA extraction, Femto Pulse system, high-molecular-weight DNA, nuclei extraction

기초지식: DNA의 추출을 위한 버퍼와 DNA의 성질(일반생물학실험 자료):

- 파쇄된 조직을 extraction buffer (CTAB buffer) 와 섞으면 조직으로부터 DNA가 분리된다 (버퍼속의 EDTA가 킬레이트(chelate) 작용을 함)
 - ※ 킬레이트: 한 개의 리간드가 금속 이온과 두 자리 이상에서 배위결합을 하여 생긴 착이온을 뜻한다.
- 조직이 파쇄되면 DNase가 세포내에서 빠져나와서 DNA를 파괴하게 된다. 그러므로 chloroform등의 단백질 비활성화 물질을 처리함으로써 모든 효소작용을 정지시킨다.
- DNA는 염(salt)의 존재 하에서 70%정도의 EtOH에서 엉기는 (pellet을 형성) 성질을 갖고 있다. 이 때 원심분리를 하면 엉긴 DNA는 가라앉고 다른 이물질은 용액속에 남아있게 된다. 이러한 성질을 이용하여 순수한 DNA를 추출할 수 있다.
- DNA는 TE (Tris-EDTA) buffer에서 매우 잘 녹는다.
- 전통적인 식물 DNA 추출방법: CTAB method (일반생물학실험에서 다름) link:

<http://amborella.net/2024-Nanopore-Lecture/03-General-Biology-DNA-extraction.pdf>



상업용 키트에 의한 일반적인 DNA 추출법:

Filter-binding method에 의한 방법

그러나 매우 단편화된 DNA가 추출됨.

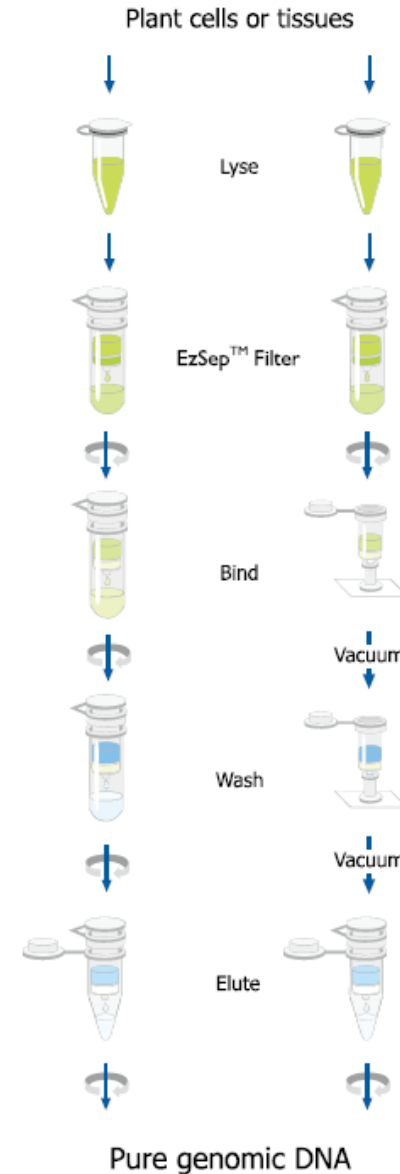
→ PCR 등 일반 실험을 위해서는 문제 없지만,

Nanopore로는 적합하지 않음.

Plant SV Kit Procedures

in microcentrifuges

on vacuum manifolds



내용 요약:

The successful application of third-generation sequencing technologies for sequencing nuclear genomes requires high-molecular-weight (HMW) DNA in sufficient quantity and quality for library preparation and sequencing (Healey et al., 2014). These DNA requirements are often challenging for non-model plant species and represent an important bottleneck for plant genome research; therefore, the development of an efficient HMW DNA extraction method is essential for the plant genomics community. Although several approaches have recently been provided for HMW DNA extraction from plants, they were only applied to a few taxa, required additional purification steps, or the essential factors influencing the process were not adequately discussed (Healey et al., 2014; Mayjonade et al., 2016; Li et al., 2020; Cai et al., 2021; Jones et al., 2021; Mavrodiev et al., 2021; Zerpa-Catanho et al., 2021). Therefore, there is a

need for an easy-to-use protocol that can produce HMW DNA from a wide range of plant taxa at a low cost.

In this study, we propose a HMW plant DNA extraction method that combines two classic protocols: (1) a nuclei extraction method (Green et al., 1987) and (2) a cetyltrimethylammonium bromide (CTAB) plant DNA extraction method (Doyle and Doyle, 1987), with modifications. The nuclear extraction step reduces the ratio of organelle genomes in the extracted DNA (Hanania et al., 2004). The CTAB method has been modified in our protocol to solve the problems associated with phenolics and polysaccharides: polyvinylpyrrolidone (PVP) was added to isolate genomic DNA, as suggested by Healey et al. (2014). To more efficiently meet the needs of genome sequencing, our combined protocol includes (1) improvements to optimize time and reagent requirements and (2) suggestions of favorable conditions for factors influencing the results

(number of pipetting steps, grinding time in liquid nitrogen, and centrifugation force in g). A combination of these two classic protocols has already been proposed for high-quality DNA extraction from *Vitis vinifera* L. (Hanania et al., 2004), but not with regard to HMW DNA and applicability in other taxa. Similarly, a method combining the nuclear isolation process and sodium dodecyl sulfate (SDS)-based DNA extraction protocol has recently been proposed for HMW DNA extraction (Zerpa-Catanho et al., 2021); however, its effectiveness has only been confirmed in a few plant taxa (six genera in three families), and it requires an extra purification step (QIAGEN Genomic Tip 20/G columns; QIAGEN, Hilden, Germany). By contrast, we have assessed the broad applicability of our protocol in species representing 18 orders of flowering plants from all major angiosperm lineages (Angiosperm Phylogeny Group, 2016), as well as a gymnosperm, *Pinus* L.

To confirm the effectiveness of our HMW DNA extraction method, we compared the results with those obtained using a commercial plant DNA extraction kit. The DNA length distributions and purity were evaluated as validation criteria for comparing the two methods. We also discuss factors influencing the results, such as the number of pipetting steps, grinding time in liquid nitrogen, and centrifugation force in g.

METHODS

HMW DNA extraction method

We sampled leaves of species from each of 18 major angiosperm orders and one gymnosperm to test the taxon-specific efficiency of our protocol. For details of all samples used in this study, see Appendix 1. Reagents, recipes, and a stepwise protocol are provided in Appendix 2. Our HMW DNA extraction protocol consists of three major steps: (1) grinding and nuclei isolation, (2) nuclear DNA extraction using CTAB buffer, and (3) RNase A and proteinase K treatment. We started with 2 g of tissue (preferably fresh, young leaves) and used a vacuum-aided cell strainer (40 μ m and 100 μ m; pluriSelect Life Science, Leipzig, Germany) to collect the nuclei suspension. We also conducted additional DNA extractions using the same samples from our HMW DNA extraction protocol. For this, we employed the Exgene Plant SV kit (GeneAll Biotechnology, Seoul, Republic of Korea), a commercial plant DNA extraction kit based on the DNA-binding filter method. Following the instructions in the manufacturer's manual, we used 0.1 g of leaf tissue, which is the recommended amount for fresh leaves.

Grinding and nuclei isolation

The protocol starts with 2 g of fresh, young leaves. We ground the leaves into a powder in liquid nitrogen (-80°C) and placed the powder in 20 mL of nuclei isolation buffer (IB). After 30 s of vortexing, we added Triton X-100 (20 μ L) and β -mercaptoethanol (1.5 mL). This step should be

conducted inside a fume hood as β -mercaptoethanol is toxic. The samples were placed on ice for 10 min, and then the mixture was filtered through a 100- μ m cell strainer (pluriStrainer 100 μ m; pluriSelect Life Science) seated in a 50-mL conical tube to collect the nuclear suspension. During filtration, gently scraping plant material accumulated on the filter with the side of a 1000- μ L pipette tip may facilitate a smoother filtration. The filtering step was repeated with a 40- μ m cell strainer (pluriStrainer 40 μ m; pluriSelect Life Science), and Triton X-100 (200 μ L) was added to the obtained nuclear suspension. This process lyses the cell and organellar membranes but not the nuclear membrane (Peterson et al., 1997). As a non-ionic detergent, Triton X-100 facilitates the release of nuclei from cells and prevents nuclei from clumping (Loureiro et al., 2007). To pellet the nuclei, the samples were centrifuged, and the supernatant was discarded. Centrifugation for 10 min at $3000 \times g$ (4°C) is recommended to prevent fragmenting long DNA molecules (see Results).

Nuclear DNA extraction using CTAB buffer

The nuclei pellet was resuspended in 5 mL of Carlson Lysis Buffer (Carlson et al., 1991). Adding β -mercaptoethanol (12.5 μ L) denatures globular proteins to make them insoluble in water (Jadhav et al., 2015). An incomplete resuspension can reduce yield; thus, we incubated the samples at 65°C for a minimum of 15 min for efficient resuspension. If the pellet still does not suspend, crushing the pellet with a pipette tip might be helpful. For easy handling, we transferred the suspended nuclei pellet to a 15-mL tube instead of proceeding with the 50-mL tube. We added 5 mL (equal volume) of chloroform:isoamyl alcohol (24:1 [v/v]) to remove impurities. During this step, chloroform (CHCl_3 ; a non-polar 3-hydrophobic solvent) dissolves non-polar proteins and lipids to promote the partitioning of lipids and cellular debris into the organic phase. Isoamyl alcohol ($\text{C}_5\text{H}_{12}\text{O}$) prevents the emulsification of the solution (Jadhav et al., 2015). After centrifugation ($3000 \times g$ for 10 min at 4°C), the aqueous upper phase containing DNA was collected and transferred into a new tube, while the organic phase containing lipids, proteins, and other impurities was discarded. The separation of a pure aqueous phase is critical for the purity of the end product, and we recommend collecting just four-fifths of the upper liquid volume to avoid including any cellular debris. Adding the proper ratio of sodium acetate (NaOAc) and isopropanol to the acquired supernatant is essential for precipitating the DNA: for every 10 mL of supernatant, a 1/10 volume of 3 M NaOAc (1 mL) and the same volume (including NaOAc) of room-temperature isopropanol (11 mL) are needed. It is essential to use room-temperature isopropanol for this step; otherwise, both polysaccharides and DNA will precipitate (Shepherd and McLay, 2011). The precipitated DNA was separated from other solvents through centrifugation ($3000 \times g$ for 10 min

at 4°C), and the resulting DNA pellet was washed with 70% cold ethanol, recentrifuged ($3000 \times g$ for 10 min at 4°C), and thoroughly dried. We recommend rapidly drying samples using room-temperature air blown by a hair dryer.

RNase A and proteinase K treatment

The DNA pellet was dissolved in 2 mL Tris-EDTA (TE) buffer. To remove RNA and protein efficiently, which account for most of the impurities in extracted DNA, we treated the samples with RNase A (10 mg/mL) and proteinase K (>600 units/mL), respectively. For each treatment, the proper incubation time and enzyme activation temperature are important: 5 min at 37°C for RNase A and 15 min at 50°C for proteinase K. The enzymes used in each step are removed by a treatment with 2 mL of chloroform: isoamyl alcohol (24:1 [v/v]). After treatment with RNase A and proteinase K, the same precipitation procedure as for the CTAB extraction is followed. The resulting pellet is dissolved using an appropriate amount of deionized water (50–500 μ L) according to the size of the pellet (recommended final concentration is ca. 200 ng/ μ L). If it is difficult to dissolve the pellet, we recommend incubating the tube at 50°C . If the pellet remains after incubation at 50°C , it is recommended to take only the dissolved aqueous layer after brief centrifugation.

Quality evaluation of extracted DNA

The quantity and purity (A_{260}/A_{280} and A_{260}/A_{230} ratios) of the extracted DNA were measured using a Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), respectively. The length distribution of the extracted DNA was evaluated using a Femto Pulse system (Agilent Technologies, Santa Clara, California, USA).

Optimization of conditions for HMW DNA recovery

We tested three factors influencing the results: (1) the number of pipetting steps, (2) the grinding time in liquid nitrogen, and (3) the centrifugation force in g. Three independent experiments were performed on different taxa in each case to evaluate each factor. First, we tested the impact of high g forces during centrifugation on DNA damage by comparing the setting in our protocol ($3000 \times g$; control group) and a higher setting ($5000 \times g$; experimental group). Second, the amount of grinding was compared. The control group was subjected to one minute of grinding (ensuring the sample was fully chilled before grinding began). The experimental group was subjected to an

additional two minutes of grinding after adding extra liquid nitrogen. Third, we assessed whether high-speed and multiple pipetting steps could potentially damage DNA. We conducted pumping at the maximum-achievable speed 200 times in a tube using a P200 tip (experimental group) and compared the resulting DNA size distribution with the original DNA (control group).

RESULTS

DNA quantity, size, and purity measurements

Usually, the quantity of the end DNA product per extraction is enough to generate 4–5 libraries (8–15 μ g) for long DNA sequencing with MinION or GridION (Oxford Nanopore Technologies, Oxford, United Kingdom), based on the library construction protocol (Ligation Sequencing Kit). The measurements obtained through the Femto Pulse system (peak height and percentages of fragments >20 kbp in the fragment-length distribution graph) confirm that our protocol successfully produced DNA fragments an average of five times longer than those generated using the commercial kit (Table 1, Figure 1), although the results of our standard HMW DNA extraction protocol showed different patterns depending on the taxon (Figure 2A, B). With our protocol, the taxon with the highest portion of >20-kbp fragments was *Chloranthus fortunei* Solms (Chloranthales; 83.6%), and the longest peak of DNA fragment distribution was obtained from *Alisma plantago-aquatica* subsp. *orientale* (Sam.) Sam. (Alismatales; 183.0 kbp) (Table 1). In the most efficient instance, our protocol yielded 35 times more DNA fragments over 20 kbp (77.1%) in *Lysimachia davurica* Ledeb. (Ericales) than the commercial kit, for which only 2.2% of fragments were greater than 20 kbp.

The quality of DNA extracted using the HMW method was superior to that obtained using the kit method in most samples. In the context of next-generation sequencing, high-quality DNA is characterized as predominantly HMW with an A_{260}/A_{280} ratio over 1.8 and without contaminating substances, such as polysaccharides or phenolics (Kasem et al., 2008; Desjardins and Conklin, 2010). With both methods, the A_{260}/A_{280} absorbance ratio, which measures protein contamination, showed similar results with low contamination (both averaged 1.83); however, our standard protocol more effectively removed carbohydrates and organic solvents (average A_{260}/A_{230} ratio = 1.88) than the commercial kit (average A_{260}/A_{230} ratio = 1.49) (Table 2; Figure 2C, D). Generally, A_{260}/A_{230} values between 1.8–2.2 indicate DNA is free of carbohydrates and organic solvents (Kasem et al., 2008; Desjardins and Conklin, 2010).

To address the statistical difference between the results from our protocol and a commercial kit, we performed paired *t*-tests on all pairs of DNA length and quality, with $P < 0.05$ considered significant. In the DNA length

TABLE 1 A comparison between our HMW DNA extraction method and a commercial kit. Fragment lengths were estimated using the Femto Pulse system.

Taxon	HMW method		Commercial kit		Ratio (a)/(b) × 100 (%)
	Peak (kbp)	% of >20 kbp (a)	Peak (kbp)	% of >20 kbp (b)	
<i>Platycladus orientalis</i>	21.57	58.8%	17.70	32.4%	181.5%
<i>Nymphaea tetragona</i> var. <i>minima</i>	22.10	59.7%	14.04	14.4%	414.6%
<i>Chloranthus fortunei</i>	38.21	83.6%	26.87	54.2%	154.2%
<i>Asarum sieboldii</i>	22.74	69.1%	28.81	66.5%	104.0%
<i>Alisma plantago-aquatica</i> subsp. <i>orientale</i>	183.00	56.8%	11.21	10.8%	525.9%
<i>Hemerocallis fulva</i>	24.80	66.2%	31.48	70.2%	94.3%
<i>Carex breviculmis</i>	107.36	83.2%	22.19	52.4%	159.8%
<i>Epimedium koreanum</i>	169.21	67.2%	10.60	10.2%	658.8%
<i>Euonymus alatus</i>	27.04	67.0%	22.96	58.1%	115.3%
<i>Viola collina</i>	142.52	75.1%	10.69	14.9%	504.0%
<i>Spiraea prunifolia</i> var. <i>simpliciflora</i>	165.50	70.1%	10.27	7.2%	973.6%
<i>Pelargonium inquinans</i>	22.45	65.2%	21.20	55.7%	117.1%
<i>Aesculus turbinata</i>	24.00	66.1%	15.38	26.1%	253.3%
<i>Lysimachia davurica</i>	154.32	77.1%	9.08	2.2%	3504.5%
<i>Isodon inflexus</i>	23.88	71.9%	21.05	49.1%	145.5%
<i>Ipomoea nil</i>	23.65	68.4%	25.78	42.1%	162.5%
<i>Adenophora erecta</i>	132.21	67.2%	13.36	19.1%	351.8%
<i>Cicuta virosa</i>	17.70	45.5%	21.05	54.4%	83.6%
<i>Sambucus williamsii</i>	157.43	64.8%	20.67	44.0%	147.2%
Average	77.88 ± 64.48	67.53% ± 0.09%	18.65 ± 6.67	36.00% ± 0.21%	455.34% ± 7.55%

criteria, the peak height and percentages of fragments >20 kbp show significant differences ($P = 0.002$ and $P = 1.66e-05$, respectively). Because the DNA extracts from our method and the commercial kit both showed excellent A_{260}/A_{280} ratios (both averaged 1.83), their quality was not significantly different ($P = 0.7769$); however, the A_{260}/A_{230} ratio was significantly different ($P = 0.001$), indicating that our protocol provided an advantage.

Factors influencing the results

Pipetting: avoid fast and frequent pipetting

Some long DNA extraction protocols suggest pipetting as little as possible or using a wide-bore tip to avoid shearing (Zerpa-Catanho et al., 2021). We confirmed that high-speed repeated pipetting damages DNA. For samples extracted based on our protocol (control group), 17.0% of the DNA fragments were >50 kbp in length, with the peak being 76.46 kbp. In contrast with the control, the sample subjected to repeated pipetting (200 times) at high

speed (experimental group) yielded just 13.4% of the fragments >50 kbp, with the peak being 49.93 kbp (Appendix 3A). High-speed over-pipetting, therefore, does affect HMW DNA extraction. The number of pipetting steps in our extraction protocol is fewer than 20, which is recommended to be performed gently with wide-bore tips to reduce the likelihood of DNA shearing.

Grinding: avoid excessive grinding

Generally, it is important to grind samples as long as possible (at least 25 min or more [Circulomics, 2021], although in practice the grinding time is much shorter) in DNA extraction to transform the plant tissue into a powder. Excessive grinding can provide a yield advantage, but it can also shear the DNA. The DNA sizes of the samples ground to different degrees were compared with the Femto Pulse system, and we concluded that additional grinding for 2 min (experimental group) has a negative effect on DNA fragment length (Appendix 3B). One minute of grinding is optimal, and additional liquid nitrogen is not needed.

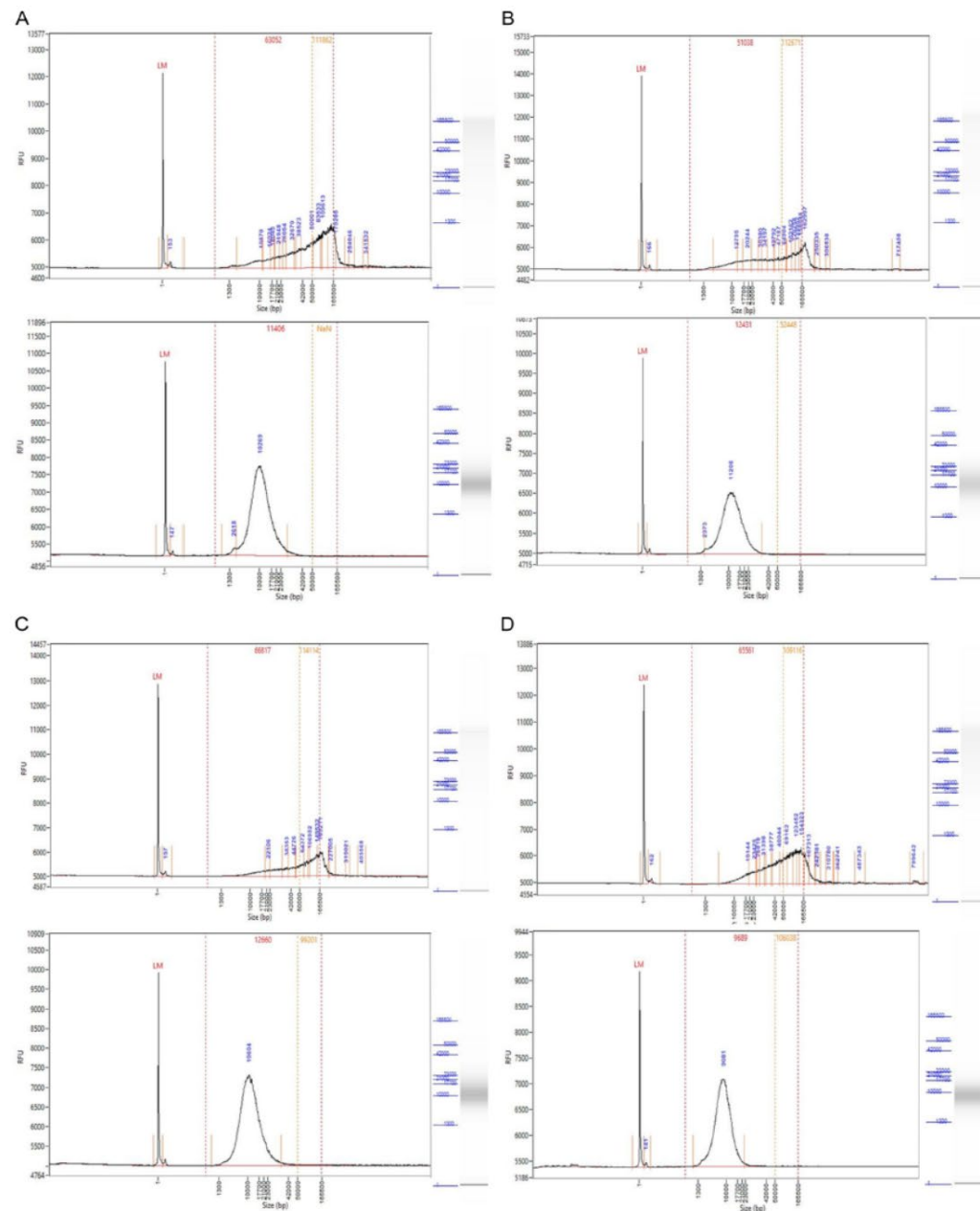
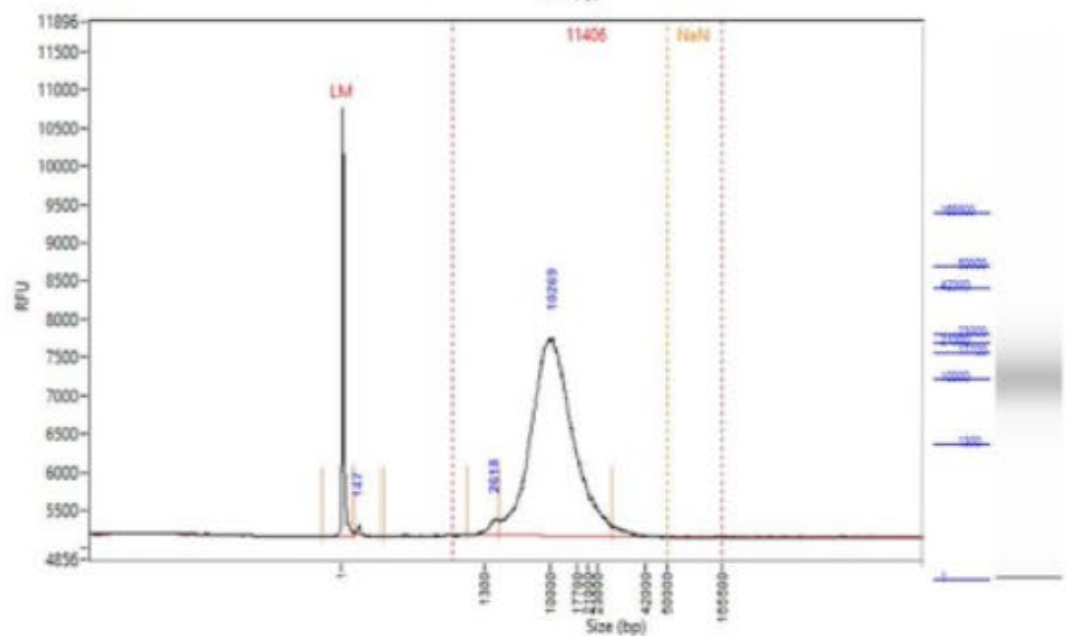
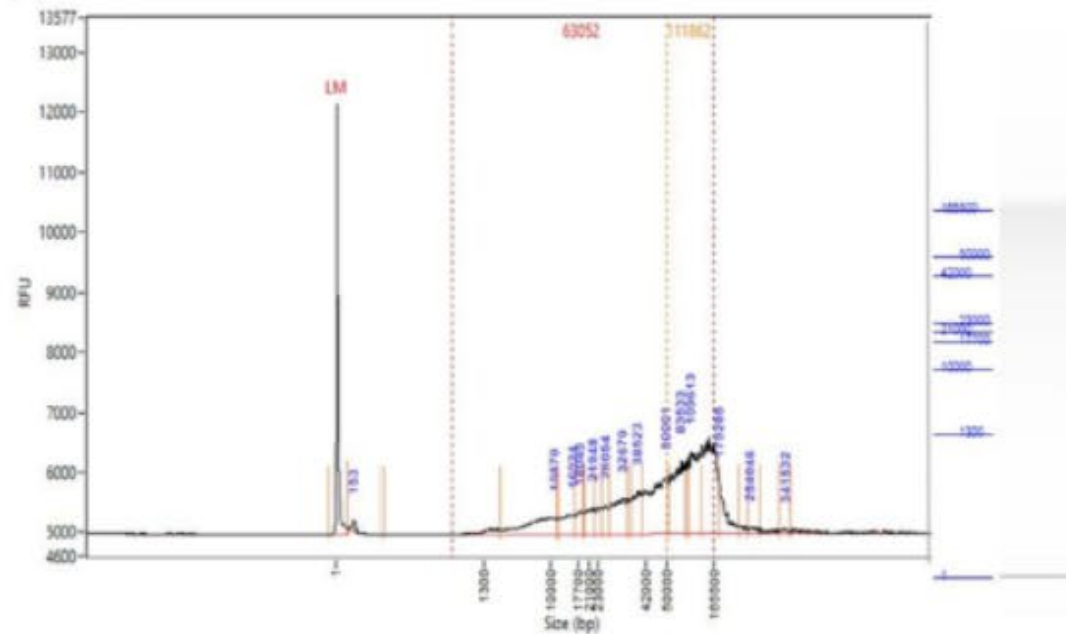
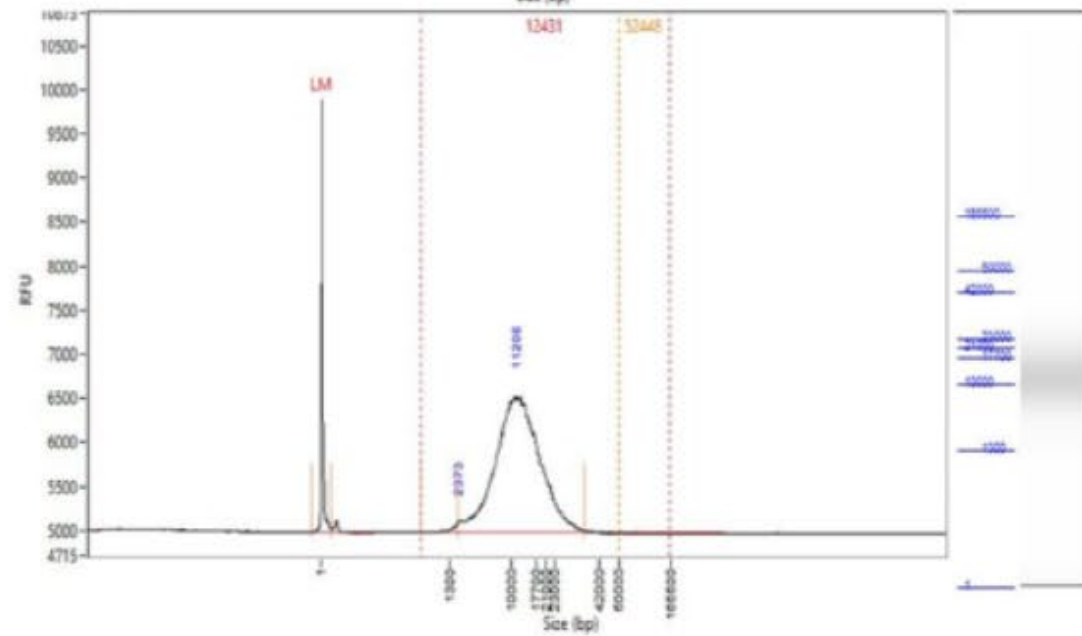
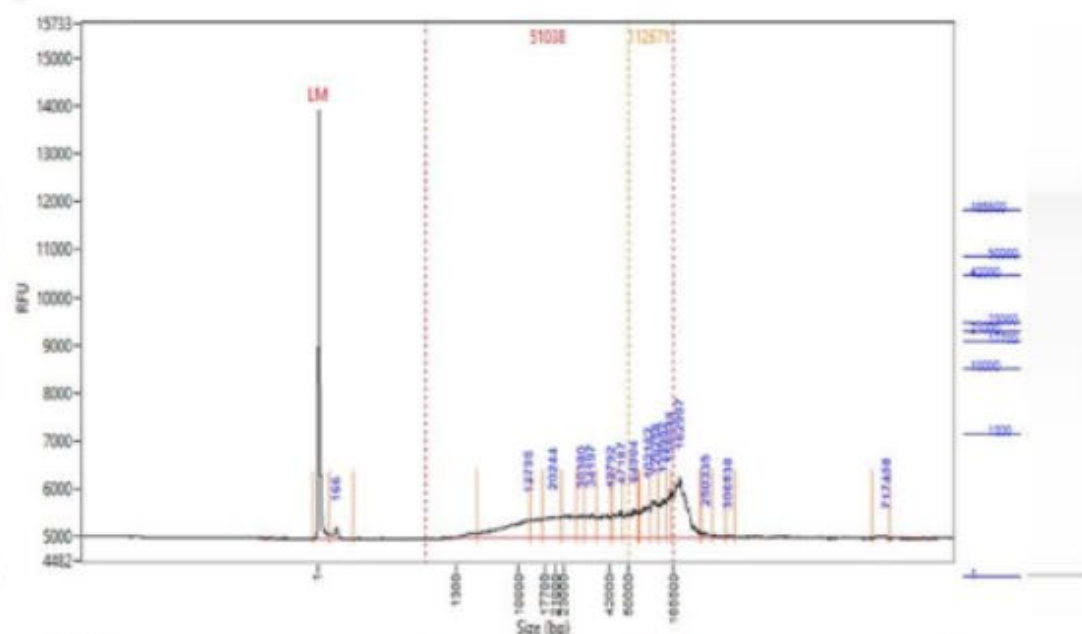


FIGURE 1 Comparison of the fragment-length distributions of the extracted DNA estimated using the Femto Pulse system (Agilent Technologies, Santa Clara, California, USA) for selected examples: (A) *Spiraea prunifolia* var. *simpliciflora*, (B) *Alisma plantago-aquatica* subsp. *orientale*, (C) *Epimedium koreanum*, and (D) *Lysimachia vulgaris* var. *davurica*. (A–D) The upper and lower graphs for each species represent the results of the HMW method and commercial kit, respectively. RFU, relative fluorescence units.

A**B**

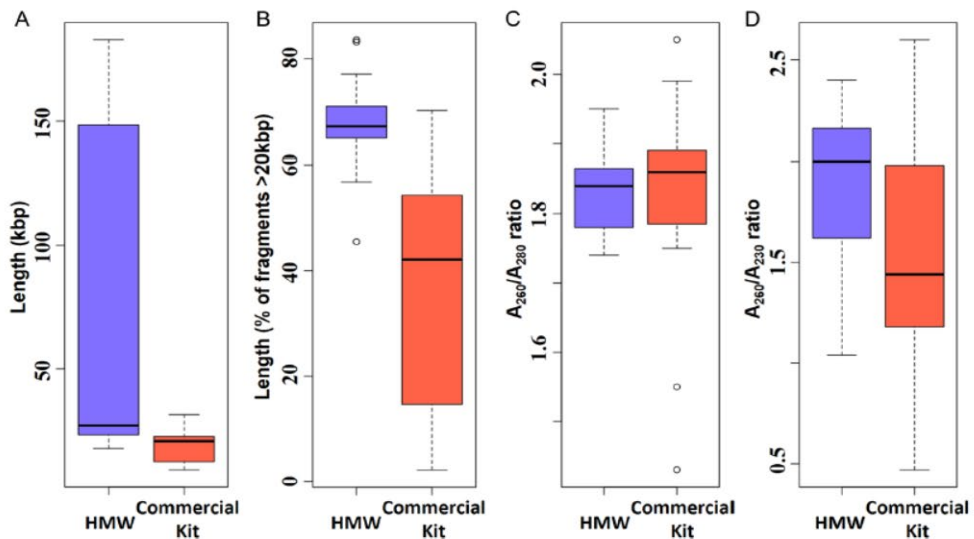


FIGURE 2 Comparison of the size and quality of DNA extracted using the two methods. (A, B) Size comparisons of (A) the highest peak and (B) the percentage of fragments >20 kbp. (C, D) Quality comparisons using (C) A_{260}/A_{280} ratio and (D) A_{260}/A_{230} ratio. The bold horizontal line in the middle of the box plot is the median value, and the lower and upper boundaries indicate the 25th and 75th percentiles, respectively.

Centrifugation: avoid high speeds

Centrifugation causes DNA molecules to collide, resulting in their molecular structure being subjected to high shearing forces (Peterson et al., 2012). Our results showed that the DNAs of the experimental group (centrifuged at $5000 \times g$) are more fragmented than the control group (centrifuged at $3000 \times g$), with measured DNA peaks (indicating HMW DNA) of 88.73 kbp and 146.94 kbp, respectively, yielding 30.0% and 39.7% of DNA fragments >50 kbp, respectively (Appendix 3C). Given this difference in producing very long fragments (e.g., >50 kbp or more), high-speed centrifugation over $3000 \times g$ is not recommended.

DISCUSSION

Here, we focused on optimizing and standardizing a HMW DNA extraction protocol for various plant taxa using economical techniques. We confirmed that our protocol successfully produced HMW DNA from various taxa in most cases; however, we expect that the experimental results will differ depending on the taxa investigated because each species has a different polysaccharide or phenolic content. Although not all species yielded good results using our protocol, we nevertheless confirmed that our protocol yielded DNA superior to the commercial kit in terms of length and purity, with statistically significant results.

To evaluate the results of various HMW DNA extraction methods, it is important to select an appropriate method and instrument with which the results can be compared. A common method to evaluate the length of the extracted DNA is a visualization of the position and brightness of DNA bands using electrophoresis through a low-concentration agarose medium (typically 0.7%) containing ethidium bromide. Alternatively, more efficient electrophoresis can be performed using a pulse-field power supply (e.g., Pippin Pulse system; Sage Science, Beverly, Massachusetts, USA). We tried pulse-field electrophoresis to check the quality of HMW DNA at the initial stage of our study; however, we confirmed that the result (the brightest position of a smeared DNA band) varied depending on the amount of loaded DNA (Appendix 4). As the quantity of DNA loaded in the agarose gel for electrophoresis is increased, the brightest position of the DNA band is shifted to a higher position (a position of higher molecular weight); that is, the quantity of DNA and the brightest position of the DNA band are positively correlated. Special attention is therefore needed to ensure that the same quantity of DNA is used for each sample when evaluating DNA length using pulse-field electrophoresis. Several automated electrophoresis techniques with fluorescence dye have been proposed for DNA length analysis to improve the unstable ethidium bromide visualization in normal electrophoresis (including pulse-field). Although the TapeStation

TABLE 2 A comparison between our HMW DNA extraction method and a commercial kit. DNA purity was evaluated using a NanoDrop.

Taxa	HMW method		Commercial kit	
	A_{260}/A_{280} ratio	A_{260}/A_{230} ratio	A_{260}/A_{280} ratio	A_{260}/A_{230} ratio
<i>Platycladus orientalis</i>	1.77	1.32	1.55	0.56
<i>Nymphaea tetragona</i> var. <i>minima</i>	1.88	2.00	1.84	1.72
<i>Chloranthus fortunei</i>	1.87	1.97	1.89	1.44
<i>Asarum sieboldii</i>	1.88	1.46	1.86	1.29
<i>Alisma plantago-aquatica</i> subsp. <i>orientale</i>	1.83	2.11	1.79	1.53
<i>Hemerocallis fulva</i>	1.79	2.33	1.79	2.60
<i>Carex breviculmis</i>	1.85	2.32	1.88	2.23
<i>Epimedium koreanum</i>	1.83	1.92	1.89	1.29
<i>Euonymus alatus</i>	1.86	2.19	1.75	0.92
<i>Viola collina</i>	1.94	2.40	1.86	2.17
<i>Spiraea prunifolia</i> var. <i>simpliciflora</i>	1.86	2.13	1.94	1.26
<i>Pelargonium inquinans</i>	1.84	2.09	1.99	1.16
<i>Aesculus turbinata</i>	1.76	1.62	1.43	0.47
<i>Lysimachia davurica</i>	1.84	1.14	1.85	1.20
<i>Isodon inflexus</i>	1.74	1.04	1.98	0.63
<i>Ipomoea nil</i>	1.76	1.62	2.05	1.80
<i>Adenophora erecta</i>	1.77	2.10	1.78	2.16
<i>Cicuta virosa</i>	1.82	1.77	1.78	1.52
<i>Sambucus williamsii</i>	1.95	2.20	1.87	2.29
Average	1.83 ± 0.06	1.88 ± 0.40	1.83 ± 0.14	1.49 ± 0.60

(Agilent Technologies) and the Fragment Analyzer (Agilent Technologies) are frequently used for size evaluations of extracted DNA fragments, they are not sensitive enough to separate HMW DNA (>60 kbp is not recommended in either instrument; Agilent Technologies, 2020a). Remarkably, the latter was used in a study of the development of a HMW DNA extraction protocol (Zerpa-Catanho et al., 2021). By contrast, the Femto Pulse system is the automated pulsed-field instrument designed for the purpose of analyzing HMW DNA. An automated pulsed-field power supply in the Femto Pulse system allows the separation of DNA up to 165 kbp (Agilent Technologies, 2020b).

CONCLUSIONS

The protocol introduced here can be used to efficiently extract HMW DNA using standard laboratory equipment (an average peak of 77.88 kbp and an average of 67.53% of fragments >20 kbp). Given its success with diverse flowering plant species and one gymnosperm, we hope our method will contribute to plant genome studies as a

broadly applicable protocol for poorly studied taxa. Additional investigations comparing DNA length, purity, and extraction cost between our protocol and commercial HMW DNA extraction kits will provide a more comprehensive understanding of the benefits of our approach.

AUTHOR CONTRIBUTIONS

S.K. and A.C. developed the experimental protocol. M.K. and S.K. performed all experiments and analyses. S.K. and M.K. wrote the preliminary manuscript draft, and all authors revised and approved the manuscript before submission. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Sungshin Women's University Research Grant of 2020 to S.K.

DATA ACCESSIBILITY STATEMENT

The taxonomic locations, vouchers, herbarium, and collection sites of all species used in this study are provided in Appendix 1.

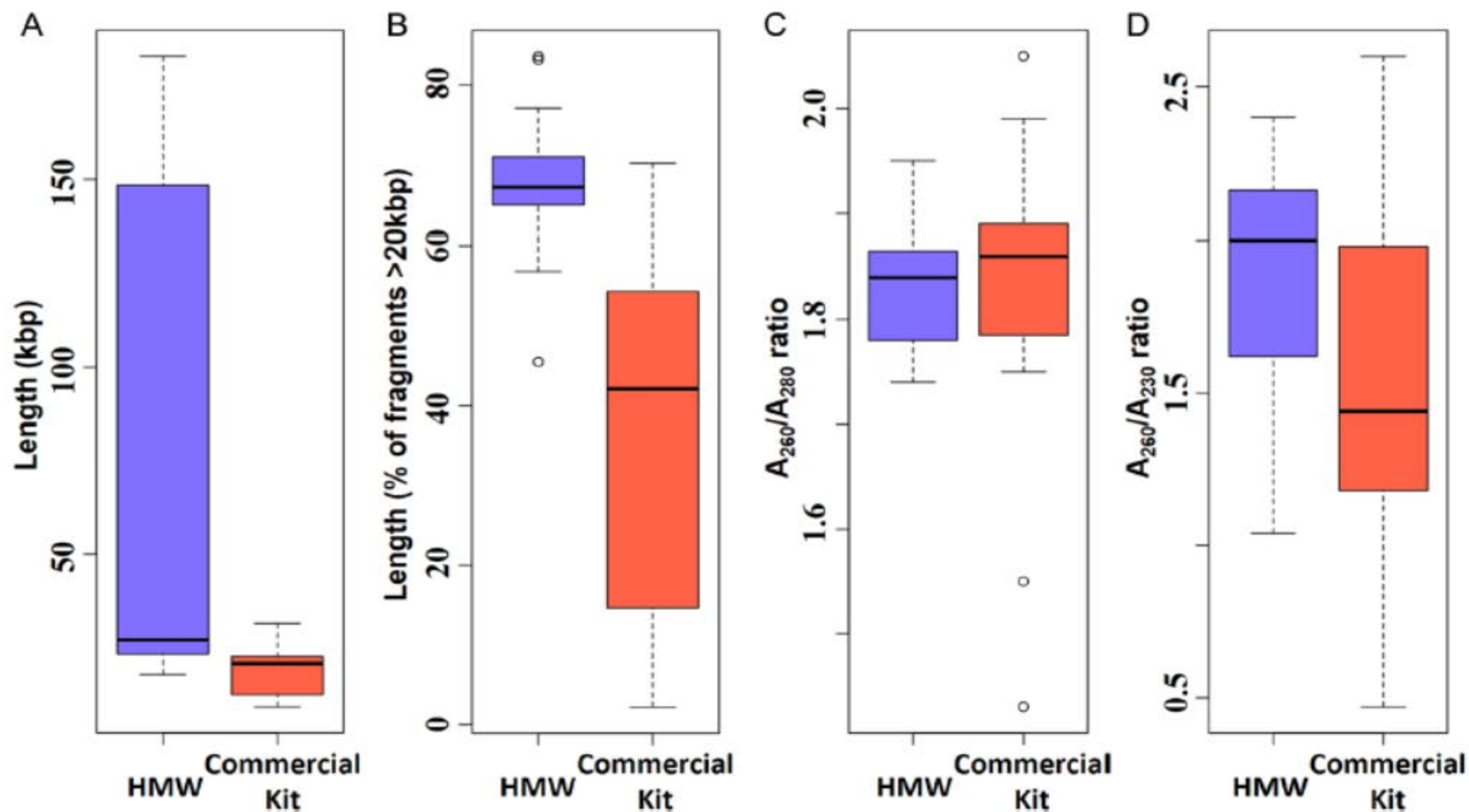


FIGURE 2 Comparison of the size and quality of DNA extracted using the two methods. (A, B) Size comparisons of (A) the highest peak and (B) the percentage of fragments >20 kbp. (C, D) Quality comparisons using (C) A_{260}/A_{280} ratio and (D) A_{260}/A_{230} ratio. The bold horizontal line in the middle of the box plot is the median value, and the lower and upper boundaries indicate the 25th and 75th percentiles, respectively.

Appendix 2: An optimized protocol for high-molecular-weight (HMW) DNA extraction in plant genomic studies.

Note: This protocol starts with 2 g of fresh, young leaves. Usually, the end product of one extraction process is sufficient to generate 4–5 libraries for sequencing with MinION or GridION (Oxford Nanopore Technologies, Oxford, United Kingdom).

I. Preparation of solutions

1. Preparation of nuclei isolation buffer (IB) (for 10 reactions)

- For 200 mL of nuclei IB, dissolve the following in ca. 100 mL of water:
 - 3 mL Tris-HCl (1 M stock, pH 9.5; final concentration: 15 mM)
 - 4 mL EDTA (0.5 M stock; final concentration: 10 mM)
 - 1.94 g KCl (final concentration: 130 mM)
 - 0.8 mL NaCl (5 M stock; final concentration: 20 mM)
- Gradually add 16 g of polyvinylpyrrolidone (PVP)-10 while rapidly stirring the solution with a magnetic stir bar.
- Use water to increase the volume to 200 mL.
- Add 0.05 g of spermine and 0.07 g of spermidine. Store IB at 4°C.
- Prepare 20 μ L of Triton X-100 and 1.5 mL of β -mercaptoethanol, to be added after mixing the IB with the ground tissue (final concentrations of 0.1% and 7.5%, respectively; this constitutes IBTB).

Note: Store at 4°C until use, or for a maximum of two weeks.

2. Preparation of Carlson Lysis Buffer (Carlson et al., 1991)

- Carlson Lysis Buffer = 2 \times cetyltrimethylammonium bromide (CTAB) buffer + 1% polyethylene glycol (PEG) 6000
- For 100 mL of Carlson Lysis Buffer:
 - 10 mL Tris-HCl (1 M stock, pH 9.5; final concentration: 100 mM)
 - 4 mL EDTA (0.5 M stock; final concentration: 20 mM)
 - 8.2 g NaCl (final concentration: 1.4 M)
 - 2 g CTAB (final concentration: 2%)
 - 1 g PEG (final concentration: 1%)

Note: Store at room temperature until use, or for up to two weeks.

3. Tris-EDTA buffer (TE) (1 \times)

- TE buffer = 10 mM Tris-HCl (pH 8.0) + 1 mM EDTA
- Note:** Store at 4°C until use.

II. Grinding and nuclei isolation (modified from Hanania et al., 2004)

- Chill mortar and pestle at -80°C before beginning the extraction procedure. Grind 2 g of fresh, young leaves in liquid nitrogen for 1 min.

Note: Ensuring the sample is fully chilled before grinding.

- Add 2 g of ground leaf powder to 20 mL of IB in a 50-mL conical tube and mix by inverting.

Note: Over-grinding negatively affects the extraction of HMW DNA. Grinding for 1 min is fine; additional grinding with extra liquid nitrogen is not needed.

Note: Increase the sample amount for succulent plants, and increase the volume of IB when the mixture becomes viscous.

- Immediately add 20 μ L of Triton X-100 and 1.5 mL of β -mercaptoethanol and mix by inverting.
- Keep on ice for 10 min.

Note: This step should be conducted inside a fume hood because the IBTB contains β -mercaptoethanol, which is toxic.

- Filter the mixture through a vacuum-aid cell strainer (pore size: 100 μm) seated on a 50-mL conical tube to collect the nuclei suspension (Figure A1).

Note: To aid filtration, gently scrape away plant tissue from the filter with the top of a 1000- μ L (blue) pipette tip. The filtrate should be light green.

- Repeat the filtering step with a 40- μm pore cell strainer.
- Add 200 μ L Triton X-100 to the nuclei suspension.

Note: This step lyses cell and organellar membranes, but not the nuclear membrane.

- To pellet the nuclei, centrifuge for 10 min at 3000 \times g at 4°C.

- Discard the supernatant.

III. Nuclear DNA extraction using CTAB buffer (modified from Doyle and Doyle, 1987)

- Add 5 mL of Carlson Lysis Buffer and 12.5 μ L β -mercaptoethanol to the tube and resuspend the nuclei pellet with brief tapping.

Note: Incomplete resuspension could reduce the yield as many nuclei will not have been lysed by CTAB. Briefly pipetting the pellet with an end-cut 1000- μ L pipette tip and gentle vortexing may aid resuspension.

- Incubate at 65°C for 15 min (maximum 2 h).

Note: If the pellet is not completely resuspended after incubation, a brief centrifugation (3000 \times g for 5 min) followed by only the use of the supernatant will help speed up processing.

- Transfer the suspended nuclei pellet to a 15-mL polypropylene tube and add an equal volume (5 mL) of chloroform:isoamyl alcohol (24:1 [v/v]) solution.

- Invert several times to mix.
- Centrifuge (3000 \times g) for 10 min at 4°C.

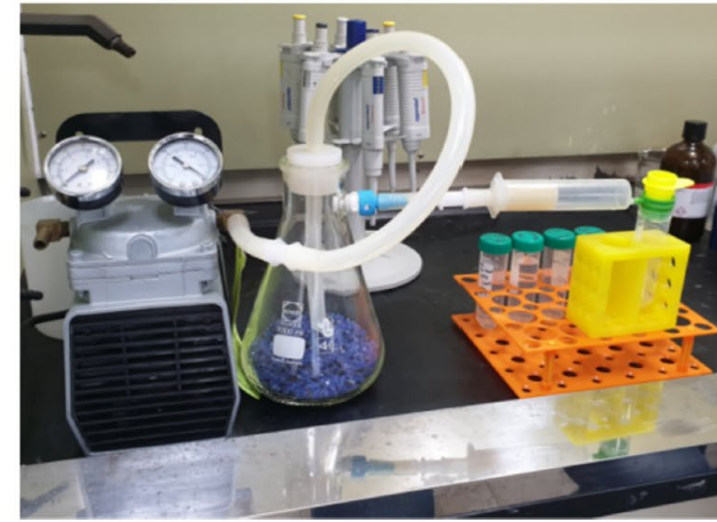


FIGURE A1 The setup of the vacuum-aid filtration including a liquid-overflow trap filled with silica gel. Using this setup can shorten extraction times.

- Transfer the aqueous upper phase to a new tube using a P1000 pipette.

Note: Take only 80% of the supernatant to avoid the inclusion of cellular debris. Take care removing the supernatant as this step is highly correlated with the quality of extracted DNA.

Note: If the supernatant is viscous, slow pipetting will help avoid sucking up the plant tissue.

- Repeat steps 3–6 (optional but highly recommended).
- Add a 1/10 volume of 3 M sodium acetate (NaOAc), mix gently, add the same volume of isopropanol (room temperature), and gently invert several times.

Note: For 4.5 mL of supernatant, add 0.45 mL of 3 M NaOAc and 4.95 mL of isopropanol.

- Precipitate at -20°C for more than 1 h.

Note: If precipitates are visible, moving to the next step is possible for faster processing. For highly viscous extracts, cold treatment makes the extract more viscous and more difficult to work with.

- Centrifuge (3000 \times g) for 10 min at 4°C.
- Discard supernatant.
- Wash pellets with 70% cold ethanol (ca. 20 mL per tube).
- Centrifuge (3000 \times g) for 10 min at 4°C.
- Discard supernatant.

Note: Keep the tube inverted for 1 min, and wipe out the tube wall with a Kimwipe.

- Dry the pellet completely.

Note: This step is very important for the quality of the DNA. The smell of alcohol is a good indicator of incomplete drying.

IV. RNase A and proteinase K treatment

- Dissolve the pellet with 2 mL of TE buffer.

Note: If the pellet is difficult to dissolve, incubate in a 50°C water bath for up to 10 min.

Note: Gently crushing the pellet with a pipette tip might be helpful for faster resuspension, but never vortex the sample. If the pellet is not completely resuspended after incubation, a brief centrifugation (3000 \times g for 5 min) followed by only the use of the supernatant will help speed up processing.

- Add 20 μ L (10 $\mu\text{L}/\text{mL}$) of RNase A (10 mg/mL conc.).
- Incubate at 37°C for 5 min.
- Add 20 μ L (10 $\mu\text{L}/\text{mL}$) of proteinase K (>600 units/mL conc.).

- Incubate at 50°C for 15 min.

- Add an equal volume (2 mL) of chloroform:isoamyl alcohol (24:1 [v/v]).

- Invert several times to mix.

- Centrifuge (3000 \times g) for 10 min at 4°C.

Note: Taking only 90% of the supernatant is best to avoid the inclusion of cellular debris. This is highly correlated with the quality of the extracted DNA.

- Repeat steps 6–9 (optional).

- Add a 1/10 volume of 3 M NaOAc, mix gently, add an equal volume of room-temperature isopropanol, and gently invert several times.

Note: For 3.5 mL of supernatant, add 0.35 mL of 3 M NaOAc and add 3.85 mL of isopropanol.

- Precipitate at -20°C for more than 1 h.

Note: If aggregates are visible, moving to the next step is possible for faster processing.

13. Centrifuge (3000 × g) for 10 min at 4°C.
14. Discard supernatant.
15. Wash pellets with 70% cold ethanol (ca. 5 mL per tube).
16. Centrifuge (3000 × g) for 10 min at 4°C.
17. Discard supernatant.

Note: Keep the tube inverted for 1 min, and wipe out the tube wall with a Kimwipe.

18. Dry the pellet completely.

Note: This step is very important for the quality of DNA. The smell of alcohol is a good indicator of incomplete drying.

19. Add 50–500 µL of deionized water to each tube to dissolve the pellet.

Note: If it is difficult to dissolve, incubate in a 50°C water bath for up to 10 min.

Note: Crushing the pellet with a pipette tip might be helpful for faster resuspension, but never vortex the sample. If the pellet is not completely resuspended after incubation, a brief centrifugation (3000 × g for 5 min) followed by only the use of the supernatant will help speed up processing.

V. DNA size and quality measurements

1. Check the quality (A_{260}/A_{280} and A_{260}/A_{230} ratios) and quantity of extracted DNA using a NanoDrop 2000

Appendix 3: Evaluation of factors influencing the DNA extraction process, using three taxa each as examples. Effects of (A) pipetting repeats (experimental group: additional 200 pipetting pumps with P200 tip), (B) degree of grinding (experimental group: additional 2 min of grinding with a second pour of liquid nitrogen), and (C) centrifugation force (control group: 3000 × g; experimental group: 5000 × g).

Example taxa for each factor	DNA fragment length		% of fragments >50 kbp	
	Control group (a)	Experimental group (b)	Control group (a)	Experimental group (b)
(A) Pipetting				
<i>Chloranthus fortunei</i>	25.99	22.00	12.5%	7.7%
<i>Alisma plantago-aquatica</i> subsp. <i>orientale</i>	165.50	91.6	26.1%	19.1%
<i>Scutellaria insignis</i>	37.89	36.18	12.3%	13.3%
Average ± standard deviation	76.46 ± 63.15	49.93 ± 30.03	17.0% ± 0.06%	13.4% ± 0.05%
Average (a) – average (b)	26.53 ± 41.04		3.6% ± 0.04%	
(B) Grinding				
<i>Chloranthus fortunei</i>	32.37	29.30	21.5%	0.0%
<i>Carex breviculmis</i>	107.36	88.95	34.7%	35.8%
<i>Viola collina</i>	142.52	147.49	46.8%	42.9%
Average ± standard deviation	94.08 ± 45.94	88.58 ± 48.25	34.3% ± 0.10%	26.2% ± 0.18%
Average (a) – average (b)	5.50 ± 11.88		8.10% ± 0.12%	
(C) Centrifugation				
<i>Chloranthus fortunei</i>	151.12	92.37	42.4%	31.1%

(Thermo Fisher Scientific, Waltham, Massachusetts, USA) and a Qubit 4 Fluorometer (Thermo Fisher Scientific).

2. Check the length distribution of the DNA fragments using the Femto Pulse system (Agilent Technologies, Santa Clara, California, USA).

VI. Special reagents and consumables

1. Reagents

- PVP-10: MilliporeSigma (Burlington, Massachusetts, USA) CAS 9003-39-8
- Spermine: MilliporeSigma S2876
- Spermidine: MilliporeSigma S2501
- Triton X-100: MilliporeSigma T8787
- PEG 6000: MilliporeSigma 81260
- RNase A: MilliporeSigma R6513
- Proteinase K: MilliporeSigma P2308

2. Consumables

- Vacuum-aid cell strainer (100 µm): pluriSelect Life Science (pluriSelect Life Science, Leipzig, Germany) 43-50100-51 yellow 100 µm
- Vacuum-aid cell strainer (40 µm): pluriSelect Life Science 43-50040-51 blue 40 µm
- Connector ring: pluriSelect Life Science 41-50000-03

Example taxa for each factor	DNA fragment length		% of fragments >50 kbp	
	Control group (a)	Experimental group (b)	Control group (a)	Experimental group (b)
<i>Scutellaria salviifolia</i>	132.27	46.21	41.4%	27.7%
<i>Sambucus williamsii</i>	157.42	127.62	35.3%	31.4%
Average ± standard deviation	146.94 ± 10.69	88.73 ± 33.33	39.7% ± 0.03%	30.0% ± 0.02%
Average (a) – average (b)	58.20 ± 22.97		9.63% ± 0.04%	

Appendix 4: Effect of DNA amount on the pulse-field gel electrophoresis image (0.7% agarose). Different quantities of the same extracted DNA (*Magnolia grandiflora*) were loaded in each lane. The length of the brightest position (peak; arrows) and the quantity of DNA show a positive relationship. M = MidRange I PFG marker (New England Biolabs, Ipswich, Massachusetts, USA); A = 125 ng; B = 250 ng; C = 500 ng; D = 1 µg; E = 2 µg.

