Next Generation Sequencing 기반 다중 DNA 바코드 개발법

Development of Multi-locus DNA Barcodes Using Next Generation Sequencing

나경주*

Gyoungju Nah*

서울대학교 NICEM 유전체분석센터

Genome analysis center at National Instrumentation Center for Environmental Management, Seoul National University

I. 서론

불량식품(Adulterated Food)이란 경제적 이득을 목적으로 원재료 대신 저렴한 유사재료를 사용하거나 섞어서 제조된 식품을 의미한다. 유사재료로 가공된 식품은 원재료에 비해 품질이 저하되고 심지어 독성 및 유해 물질을 포함할 수 있어, 외관상 구분이 어려운 유사재료를 과학적으로 판별하는 대책이 매우 중요하다. 유럽에서는 European Union Regulation(EUR)을 통하여 식품안전에 관한 대책으로 이력추적관리제를 추진하고 있고, 미국의 식품의약품청(FDA)과 영국의 식품기준청(FSA) 등에서도 증가하고 있는 불량식품에 대한 관리감독을 강화하여 불량식품으로부터 소비자를 보호하기 위한 대책을 마련하고 있다[1,2]. 국내에서도 식품 제조 과정 중 원재료 이외의 유사재료들을 비의도적으로 혼입하거나 대체 재료로 사용하는 것을 규제하기 위해 원료 물질을 모니터링 할 수 있는 효율적이고 정확한 분석 방법의 요구성이 시급한 실정이다.

Polymerase Chain Reaction(PCR) 기반 DNA 분석법은 비교적 시료가 열에 안전하고 특이성이 높아 간단한 방법으로 보다 정확한 분석결과를 낼 수 있다는 장점이 있어 많이 사용되어 왔다. PCR 기반 DNA 분석법은 종특이적 DNA 염기서열에 근거하는데 종 동정을 위해 DNA 바코드 마커(DNA barcode marker)를 활용하고 있다. DNA 바코드는 2003년 독일 Hebert의 연구팀에서 발표하였으며[3] genome 상의 보편화된 지역에서 종특이적 변이성이 존재하는 유전자 지역을 발굴하고 이 지역을 증폭할 수 있는 프라이머(primer)를 개발하여 PCR로 판별하는 기법이다. 식물에서는 핵유전체보다 비핵유전체인 chloroplast genome(cp genome)을 주로 사용하는데 그 이유는 모계유전을 따르는 엽록체 유전자들은 비교적 잘 보존되어 있으면서도 종특이적 마커를 발굴하기에 충분한 변이성이 존재하기 때문이다. 식물에서는 엽록체 유전체 상의 단일 유전자(single copy gene)인 Maturase K(*matK*)와 Rubisco large subunit(*rbcL*)가, 동물에서는 mitochondrial

* Correspondence to: Gyoungju Nah
Genome analysis center at National Instrumentation Center for Environmental Management,
Seoul National University, Seoul 08826, Korea
Tel: 02-880-4971 Fax: 02-888-4847 Email: GJNAH@snu.ac.kr

genome(mt genome)의 Cytochrome c oxidase subunit 1(*CO1*)이 범용성 핵심 바코드(core DNA barcode)로 사용되고 있다[4,5]. 이들은 각각의 장단점이 존재하는데 *matK*의 경우 상대적으로 높은 evolutionary rate로 인해 species 판별에 사용되고 있으나 근연관계에 따라 alignment가 잘 되지 않는 단점이 있다[6]. 이에 비해 *rbcL*은 family와 genus 수준에서 사용되는 바코드로서 대부분의 육지식물에 존재하므로 alignment가 잘 되고 amplify가 쉽다는 장점이 있으나 변이성이 낮다는 단점이 있다[7]. DNA 바코드의 중요성이 높아짐에 따라 국제 컨소시엄이 구성되었으며 2004년에 설립된 Consortium for the Barcode of Life(CBOL, http://barcodeoflife.org)가 대표적이다. CBOL은 2009년부터 *matK*+*rbcL*의 바코드를 공용으로 분석하는 것을 추천하고 있다[8]. 그러나 최근 *matK*+*rbcL*에 의한 판별력은 전체 육지식물의 약 75%로 보고되고 있어 범용성 핵심 바코드의 한계점이 드러나고 있다[9].

기존의 식품원료 종판별을 위한 DNA 바코드마커 증폭은 일반 프라이머(universal primer)를 사용하는 방법과 종특이적 프라이머(species-specific primer)를 사용하는 방법이 사용되고 있다[10]. 일반 프라이머 방법은 PCR 증폭 후 염기서열을 시행해야 하고 유전자은행에 등록되어 있는 비교대상종의 바코드마커 염기서열과 비교분석해야 하는 단점이 있다. 종특이적 프라이머 방법은 바코드마커의 증폭 유무로 결정되므로 이후에 염기서열 단계가 필요하지 않지만 technical error에 의해 증폭되지 않을 수 있거나 혼입시료인 경우 정량화가 어렵다는 단점이 있다. 또한 1~2개의 핵심 바코드마커의 사용은 판별 대상종이 근연종일 경우 마커의 다형성의 제한을 받을 수 있다. 이러한 제한성을 극복하려면 2가지 면에서 개선이 필요한데 공우성 마커(Co-dominant marker)와 다중 마커(Multi-locus marker)의 개발 및 활용이다. 공우성 마커는 기존의 일반 프라이머나 종특이적 프라이머 방법보다 다음과 같은 유리한 면이 있다:(1) PCR시 technical error에 의한 증폭 여부를 탐지할 수 있다.(2) 식품 원료의 경우 원재료와 유사재료가 혼입되는 경우가 종종 있는데 이때 공우성 마커일 경우 혼입에 사용된 재료종들의 정량적 측정이 가능하다.(3) Size로 판별되므로 일반 연구실에서 전기영동으로 쉽게 판별 가능하다.(4) Insertion-Deletion(InDel)은 종내보다 종간에서 주로 발견되므로 종간 구분에 용이한 마커의 특징을 지닌다. 또한 다중 바코드마커는 기존의 1~2개 마커 사용보다 몇가지 유리한 면이 있다:(1) 근연종과의 판별에서 정확도를 높일 수 있다.(2) 향후 판별대상이 될 유사재료 종과의 변이정도에 따라 마커를 선택적으로 사용할 수 있다. 공우성 및 다중 바코드마커의 발굴을 위해 엽록체 유전체 sequence를 대량으로 확보할 수 있다면 기존 방법의 한계점을 극복한 마커의 개발이 가능하다.

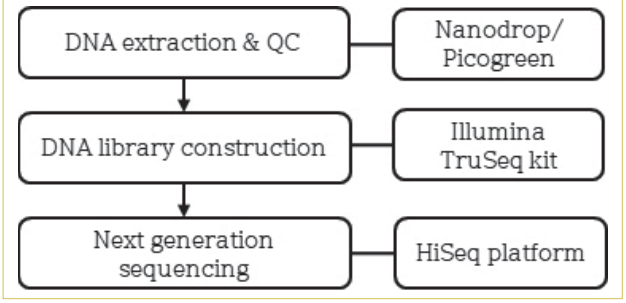
최근 Next Generation Sequencing(NGS) 기술의 발달과 비용 저하로 엽록체 유전체 염기서열을 대량으로 확보할 수 있게 되었다. NGS 기반 엽록체 유전체 염기서열법은 엽록체의 순수 분리 단계 없이 전체 genomic DNA를 사용하기 때문에 편리하게 sequencing 할 수 있다[11,12]. 엽록체 유전체는 핵유전체에 비해 크기가 매우 작아(약 120~180kb, 0.5Gb nuclear genome의 식물일 경우, 약 1/2,800~1/4,000의 크기) 상대적으로 적은 비용으로 NGS가 가능하고 multiplexing으로 진행할 수 있어 대량 염기서열이 가능하다. 또한 엽록체는 recombination이 일어나지 않는 haploid genome으로서 한 cell 당 수백 copy가 존재한다고 알려져 있어 전체 genomic DNA를 추출하였을 때 소량의 엽록체 DNA라도 핵 DNA보다 월등히 높은 copy number를 확보할 수 있어 assembly에 유리하다[12]. 보통 2~5Gb의 DNA 데이터양으로 assembly 하기에 충분한 엽록체 유전체 sequence를 확보할 수 있으므로 HiSeq 기반 NGS의 경우 기기 사양에 따라 1개 lane 당 50~100Gb의 데이터가 생성되는 것을 고려하면 수십개 이상의 sample을 동시에 sequencing 할 수 있게 된다. 생성된 raw read들은 샘플당 tag된 index sequence별로 분리되며, 분리된 raw read들을 assembly를 통하여 수십 kb 이상의 contig를 생성한다. 생성된 contig의 ordering과 연결이 완결되면

원래 엽록체 유전체 형태인 circular genome form을 만들 수 있다. Circular form으로 완결되지 않더라도 7개 이하의 contig의 합이 100kb 전후 크기를 포괄할 수 있으면 큰 어려움 없이 마커 발굴이 가능하다. 최종 contig가 결정되면 assembly된 contig들은 gene annotation을 거쳐 유전자 위치를 결정한다. 이후 2개 이상 재료종들의 유사지역을 나열하는 alignment 단계를 거쳐 변이성(InDel, SNP)이 발견되는 지역을 찾는다. 바코드마커 지역은 먼저 비교종과 alignment가 잘 되어야하고, 종 특이적 DNA variation이 존재해야 하면서, PCR product size가 약 200~500bp 범위가 되도록 primer를 design이 가능한 조건을 갖추어야 한다. 본문에서는 NGS와 정보분석에 의한 다중 바코드마커 발굴법과 근연종과의 계통분석을 통해 해상력 증대에 대한 다중 바코드마커의 역할을 제시하고자 한다.

II. 본론

식물 엽록체의 다중 DNA 바코드마커는 크게 염기서열과 정보분석의 2단계를 통해 개발된다. 본 연구에서는 식용 식물인 산마늘(*Allium victorialis*)의 엽록체 유전체 염기서열을 분석하고 그 유사재료로 비식용 식물인 박새(*Veratrum patulum*)와의 판별을 위한 다중 바코드마커 개발법[13] 을 소개한다.

1. Illumina 사의 HiSeq 기반 sequencing은 아래와 같이 3단계로 구분된다(Fig. 1).



〈Fig. 1〉 Workflow of HiSeq DNA library preparation

〈Fig. 2〉 Instruments for DNA QC

(1) DNA 추출 및 QC

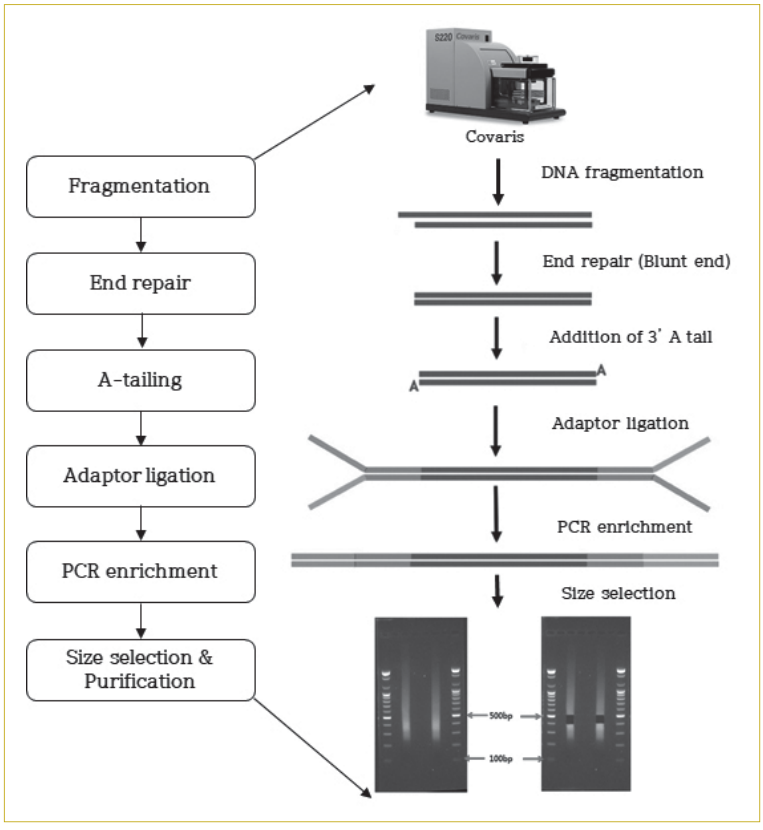
Total genomic DNA 추출은 Promega genomic DNA extraction wizard kit(Cat# A1120, USA)을 이용하여 user's manual에 따라 추출한다. DNA 추출 후 QC는 Fig. 2와 같은 장비를 사용하여 3단계로 진행한다. QC의 목적은 추출된 DNA 정량 외에 순도 등 정성적 점검을 통하여 DNA library 제작에 필요한 고품질 DNA를 확보하기 위함이다.

- QC1(Quantity 검사): Picogreen으로 농도를 측정하여 DNA library 제작에 필요한 양을 측정한다. 먼저 박테리오 파지 λ DNA 100ug/mL 짜리를 필요한 양만큼 1/50로 희석한 후 최종 200ul의 부피로 0~1ug/mL 농도 사이의 control을 5개 이상 준비하여 standard curve를 그릴 수 있는 control point를 5개 이상 잡는다. Standard curve의 비율과 동일하게 최종 volume 200ul에 DNA 2ul를 넣는다. 샘플에 picogreen을 넣고 2~5분 반응후 victor 기기로 480nm의 형광을 측정한다. DNA library 제작에 필요한 양은 2ug이 요구된다.
- QC2(Purity 검사): Nanodrop으로 260nm/280nm의 비율이 1.8~1.9가 되도록 한다. 만일 1.8 이하일 경우는 (1) DNA 농도가 매우 낮거나 (2) DNA 농도가 정상인데 낮을 경우 단백질의 오염이 있는 것으로 간주한다.
- QC3(Integrity 검사): Gel electrophoresis를 이용하여 picogreen으로 측정된 샘플 농도에 따라 100ng의 DNA를 1%의 gel 에 1kb ladder marker

와 100ng의 λ DNA와 같이 50V, 1시간 이상 전기영동하여 내린다. 전기영동 후 추출된 DNA의 intact한 정도와 크기, picogreen 농도 값이 정확한지 λ DNA와 비교 확인한다. DNA가 degradation 된 경우는 끌림 현상이 나타난다.

(2) HiSeq용 DNA library 제작

QC를 통과한 genomic DNA는 Covaris 등 물리적 초음파를 이용하여 무작위 절단한 후 Illumina 사의 TruSeq DNA library prep kit를 사용하여 sequencing용 library를 제작한다. 대용량 염기서열 분석을 위해 genomic DNA를 무작위적인 방식으로 절단하고 절단된 DNA의 말단은 Illumina 염기서열 분석기용 adaptor가 연결될 수 있도록 blunt화 한다. 무작위로 생성된 DNA단편에 adaptor를 연결하고 PCR 과정을 통해 양 말단에 adaptor가 연결되어진 DNA 단편을 증폭하여 library를 제작한다. Library span size는 300~500bp로 무작위 절단 장치인 covaris의 조건 intensity, duty cycles, burst의 값을 조정하여 평균 DNA의 크기를 조정할 수 있다. 상세한 library 제작 기술은 다음과 같다(Fig. 3).



〈Fig. 3〉 Overview of HiSeq DNA library construction

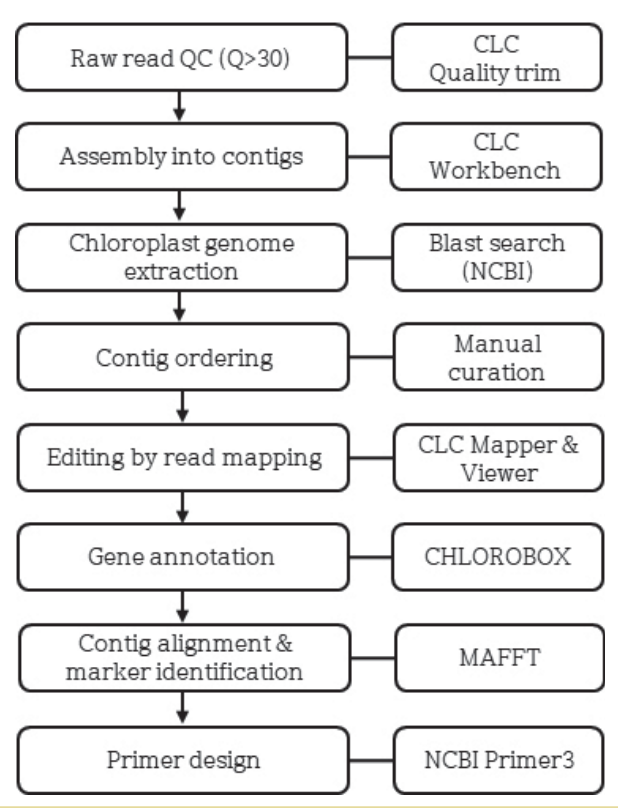
- Random shearing: Covaris는 물리적 초음파를 이용하여 일반 초음파보다 15배 이상 강한 초음파를 생성하며, 대용량 염기서열 분석 library 제작을 위한 무작위적인 DNA 절단에 사용되고 있다. Covaris의 intensity, duty cycles, burst값 조정과 예비 테스트를 통해 지정한 크기로 절단되는 조건을 찾는다.
- Fragment end repair: 절단된 DNA 단편은 polymerase와 Klenow 효소를 이용하여 손상된 DNA 말단의 nucleotide를 절단, 또는 채움 방식으로 blunting 한다.
- A-tailing: Blunting된 DNA 3'말단은 phosphorylation하고 adenosine을 ligation한다. 그 결과 DNA 단편 말단에 한 개의 A 뉴클레오타이드를 결합하여 A-overhang DNA 형태를 보여준다.
- Adaptor ligation: Illumina HiSeq 2500분석 위한 adaptor를 DNA 말단에 ligation한다. Adaptor는 P7, P5로 각각 구성되어 있으며 P5는 염기서열분석 프라이머 binding 위치, flowcell 결합서열로 구성되어 있다. P7 프라이머는 한쪽 방향으로 염기서열을 분석한 후 반대방향으로 염기서열 분석을 위한 서열로 구성되어져 있다. 그 결과 DNA 단편 말단에 adaptor 서열이 ligation된 library 형태를 보여준다.
- PCR enrichment: PCR은 프라이머 1, 프라이머 2, 프라이머 3인 세가지 프라이머를 혼합하여 PCR한다. 프라이머 3은 multiplex identifer로 여러 개의 시료를 혼합하

여 분석할 수 있다.

- Size selection: Library 증폭 후 분석하고자 하는 span 크기에 adaptor 크기(150bp)를 합한 분자량의 library가 주로 증폭되며, agarose gel 상에서 해당 분자량을 오려내어 DNA를 추출한다.

(3) HiSeq 기반 sequencing

- Real time PCR: 대용량 염기서열 획득을 위해 real time PCR을 이용한 library 분자량을 측정한다.
- Cluster generation: library DNA 단편을 flowcell 내에서 증폭하여 cluster를 생성한다.
- Sequencing 시행한다.



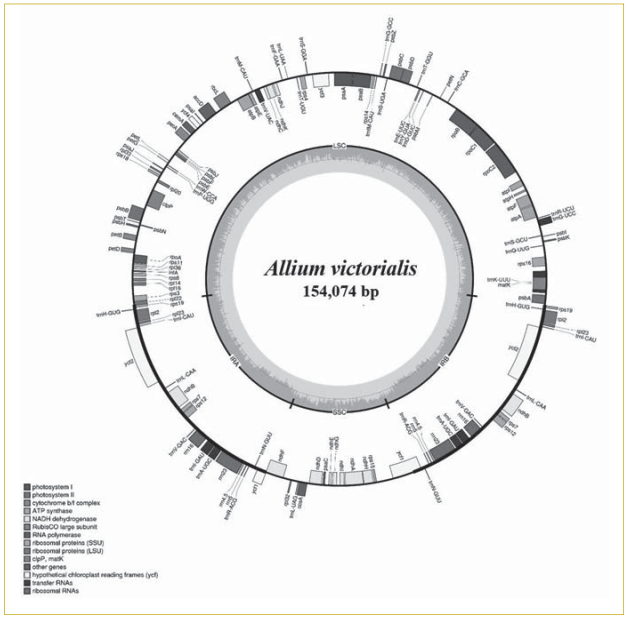
〈Fig. 4〉 Workflow of bioinformatics pipeline

2. 생물정보 분석

HiSeq sequencing으로 얻은 raw read들은 엽록체 유전체 assembly와 annotation을 거쳐 다형성을 보이는 지역에서 barcode 마커 개발에 사용된다. 단계별로 Fig. 4와 같이 CLC Workbench program(version4.0, CLC Inc. Arus, Denmark) 및 기타 software들이 사용된다.

- 데이터 QC: HiSeq 결과로 얻어진 raw read들은 Q-value 30이상으로 filtering 후 *de novo* assembly에 들어간다.
- Assembly: Filtering된 raw read들은 assembly 단계를 거쳐 contig들을 생성하게 된다.
- 엽록체 유전체 선발: Assembly된 contig들은 핵유전체와 비핵유전체가 혼재되어 있으므로 엽록체 genome의 validation을 위해 contig들을 기존의 Chloroplast genome database(NCBI)에 BLAST를 시행한다. 엽록체 유전체에 해당하는 contig들은 sequence similarity에 의해 가장 근연관계가 가까운 종엽록체 유전체 database에 mapping된다.
- Contig ordering: 선발된 엽록체 contig들은 contig가 1개 이상일 경우 기존 엽록체 유전체 정보를 참조하여 ordering을 진행한다. Contig 간의 gap은 N으로 처리한다. 만일 coverage가 충분하여 1개의 contig로 assembly 되면 circulation을 시도한다.
- Read mapping에 의한 수정: assembly에 사용되었던 raw read들을 다시 엽록체 contig에 mapping하여 duplication 등에 의해 mis-assembly된 지역을 확인 및 수정한다.
- 유전자 annotation: 엽록체 유전체 상의 유전자 위치를 동정한다.
- Contig alignment 및 다형성 지역 발굴: 비교 대상종(ex. 원재료종 vs. 유사재료종)과 alignment를 진행하고 비교종과 InDel 및 SNP를 보이는 지역을 발굴한다.
- Primer design: 마커 후보 대상 지역을 증폭할 수 있는 primer set를 정한다.

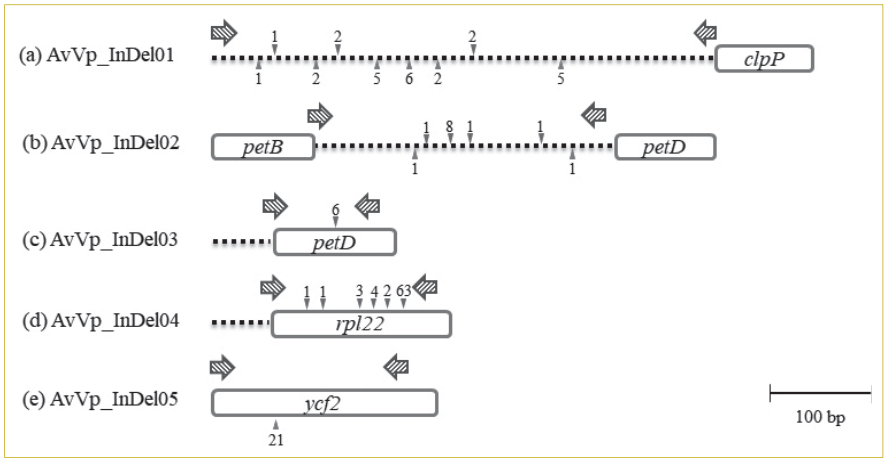
위 방법을 토대로 산마늘(*Allium victorialis*)과 박새(*Veratrum patulum*)를 구분하기 위해 산마늘의 엽록체 유전체의 염기서열을 분석하고 그 유사재료인 박새와 비교하여 엽록체 유전체 상의 5개 지역에서 InDel 부분을 증폭하는 primer를 개발하여 얻은 실제 데이터는 다음과 같다. 먼저 NGS로 얻어진 데이터로 assembly 후 circulation으로 map을 작성하고 annotation을 진행하였다(Fig. 5).



〈Fig. 5〉 Chloroplast genome map of *Allium victorialis*

이 정보로 NCBI에서 확보한 박새의 엽록체 유전체 sequence와 alignment 후 *clpP*, *petB*, *petD*, *rpl22*, *ycf2* 의 genic 및 intergenic region에서 5개의 바코드마커를 발굴하였다(Fig. 6).

Table 1은 산마늘과 박새의 엽록체 유전체 비교에서 발굴한 5개의 InDel 바코드마커와 primer sequence, 그리고 PCR 증폭시 예상되는 product size를 명시하였다. 여기서 얻은 5개 바코드마커로 근연종 포함 8개 식물종에서 단일 바코드마커를 사용하였을 경우와 5개의 다중 바코드마커를 사용하였을 경우를 phylogenetic tree를 구축하여(bootstrap value 1,000) 판별 해상도를 비교분석하였다(Fig. 7a~f). 그 결과 단일 바코드마커에서는 근연종들과 30~79%의 해상력을 보이는 clade가 발견되었는데 다중 바코드마커의 경우 모든 clade에서 86% 이상의 해상력을 보여 신뢰성 높은 근연관계 및 판별력을 나타내었다. 특히 *Allium* 속의 3개종의 분리는 단일 바코드마커에 따라 *Allium victorialis*와 34~98%의 차이를 보였으나(Fig. 7a~e) 다중 바코드마커의 경우 100%의 높은 해상력을 나타냈다(Fig. 7f).



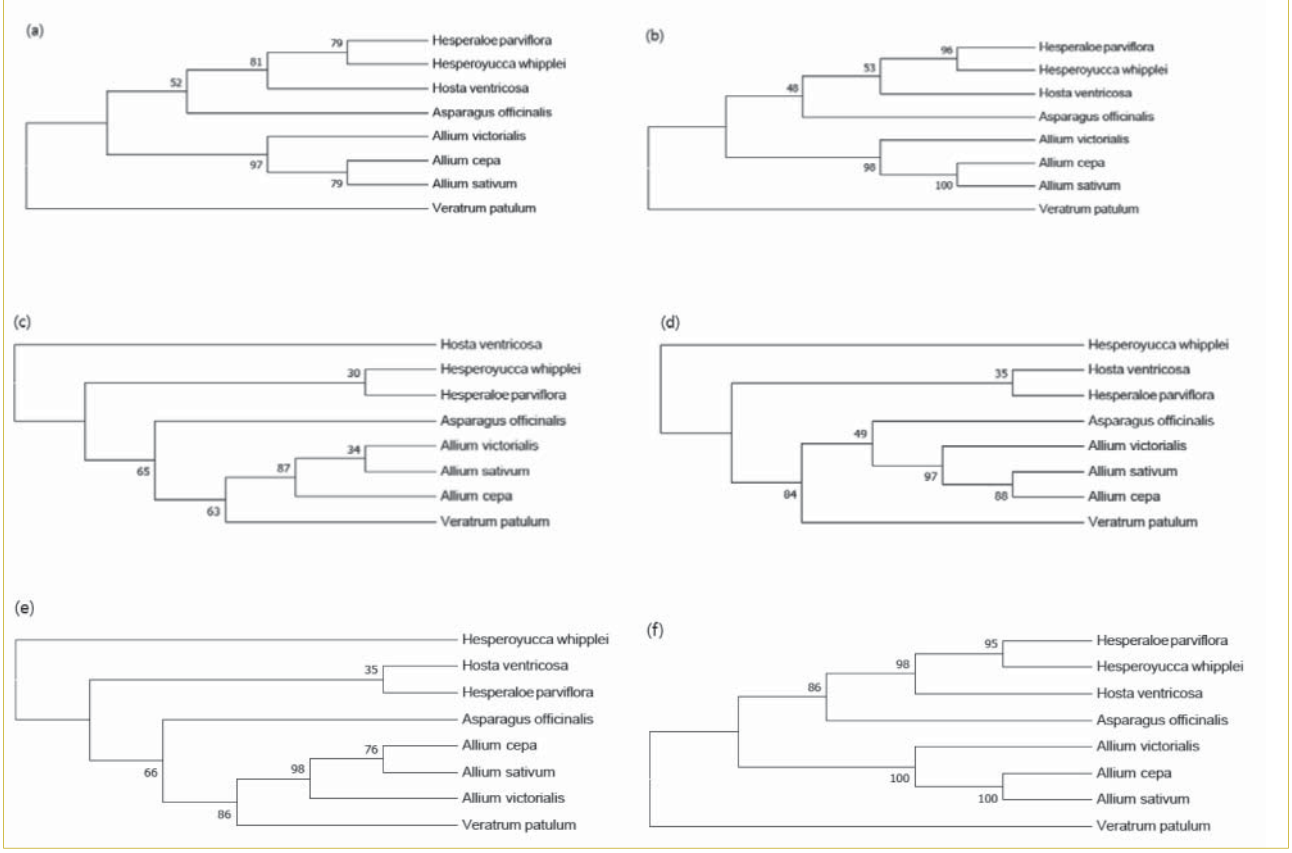
〈Fig. 6〉 Diagram of five co-dominant barcode markers of *A. victorialis* and *V. patulum*. The grey colored boxes indicate genes. Blue and red triangles indicate deletion of *A. victorialis* and *V. patulum*, respectively. The numbers above or below triangles represent the size of deleted nucleotides. Purple dashed arrows indicate designed primers.

III. 결론

본 연구는 NGS 기반 엽록체 유전체 염기서열 분석에 기반한 다중 공우성 바코드 마커 발굴 방법과 그 효용성에 대해 제시하였다. 본 연구에서는 기존의 *matK*나 *rbcL*외에 다양한 지역에서 바코드마커 발굴이 가능하며 단일 마커보다 다중 마커 발굴이 추후 근연종과의 판별에 도움이 될 것임을

Table 1. Estimated PCR product size and primer sequences for five candidate loci in *A. victorialis* and *V. patulum.*

Marker ID	Av (bp)	Vp(bp)	Primer	Description
AvVp_InDel01	503	487	F: AGGACAAATGATCTCAGTACCACT R: TGCCCATTGGTGTTCCAAAAG	*clpP* genic-intergenic region
AvVp_InDel02	292	301	F: ACACACTTTTGTATTGCCTCTTC R: TCTTCGGAGAATCCACTTCAACT	*petB* genic-intergenic region
AvVp_InDel03	110	116	F: GCTCGAGCCGGATGATGAAA R: AATACAGGATCATTCAAGTCAGGT	*petD* genic-intergenic region
AvVp_InDel04	173	247	F: GGGTTGTACCAAGTCTGAAACC R: CGATAAAAAGACCCACTTGTCAT	*rpl22* genic-intergenic region
AvVp_InDel05	210	189	F: CTAAGTCACTTCGTTTCTTTTTGTC R: TCAAATGAACGATTTGAACACCTAT	*ycf2* genic region



〈Fig. 7〉 Phylogenetic tree construction of eight related species of *A. victorialis* using five barcode markers (Maximum Likelihood, bootstrap value 1,000).
(a) AvVp_InDel01 (b) AvVp_InDel02 (c) AvVp_InDel03 (d) AvVp_InDel04 (e) AvVp_InDel05 (f) AvVp_InDel01~AvVp_InDel05

보여주었다. 특히 확보된 엽록체 유전체 염기서열 및 마커 정보들은 database 구축과 더불어 추후 다양한 판별 대상 종의 출현에 따라 마커를 선택 또는 추가 마커를 발굴할 수 있는 source의 역할을 하게 될 것이다. 또한 공우성 마커의 발굴은 혼입재료의 판정량적 판별에 도움이 될 것이다. 본 연구는 향후 식품 및 의약품 원료를 포함한 재료종의 진위를 판별하는 과학적 근거를 제시하여 불량식품을 예방하고 나아가 새로 발견되는 유사재료 종에 대한 대량 마커 발굴 기법으로 적극 활용될 것으로 기대된다.

참고 문헌

1. Barcaccia, G., Lucchin, M., and Cassandro, M.: DNA barcoding as a molecular tool to track down mislabeling and food piracy. *Diversity*. 8, 1-16(2016).

2. Kim, K.H., Kim, Y.S., Kim, M.R., Lee, H.Y., Jung, Y.K., Lee, J.H., Chang, H.S., Park, Y.C., Kim, S.Y., Choi, J.D., and Jang, Y.M.: Development of species-specific primer to determine the authenticity of vegetable raw materials in food. *Food Eng. Prog. Vol.* 18, 419~426(2014).

3. Herbert, PD., Cywinska, A., Ball, SL., and deWaard, JR.: Biological identifications through DNA barcodes. *Proc. Biol. Sci*. 270, 313-321(2003).

4. Li, X., Tang, Y., Henry, RJ., Rossetto, M., and Wang, Y. Plant DNA barcoding: from gene to genome. *Biol. Rev. Camb. Philos. Soc*. 90, 157-166(2015).

5. Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R.: DNA primers for amplification of mitochondrial cytochrome C oxidase subunit from diverse metazoan invertebrates. *Mol. Mar. Bio. Biotechnol*. 3, 294-299(1994).

6. Min, XJ., and Hickey, DA.: Assessing the effect of varying sequence length on DNA barcoding of fungi. *Mol. Ecol. Notes*. 7, 365-373(2007).

7. Hollingsworth, P.M., Graham, S.W., and Little, D.P.: Choosing and using a plant DNA barcode. *PLoS One*. 6, e19254(2011).

8. Group CPW.: A DNA barcode for land plants. *Proc. Natl. Acad. Sci*. USA. 106, 12794-12797(2009).

9. Wilkinson, M., Szabo, C., Ford, C.S., Yarom, Y., Croxford, A.E., Camp, A., and Gooding, P.: Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants. *Sci Rep*. 7, 46040(2017).

10. Fabric, T., Celia, M., and Catherine, H. Food and forensic molecular identification: update and challenges. *Trends Biotechnol*. 23, 359-366(2005).

11. Nock, C.J., Waters, D.A., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M., and Henry, R.J.: Chloroplast genome sequences from total DNA for plant identification. *Plan Biotech*. J. 9(3), 328-33(2011).

12. Garaycochea, S., Speranza, P., and Alveraz-Valin, F.: A strategy to recover a high-quality, complete plastid sequence from low-coverage whole-genome sequencing. Appl. *Plant Sci*. 12, 3(10)(2015).

13. Lee, J., Chon, J., Lim, J.S., Kim, E.K., and Nah, G.J.: Charaterization of complete chloroplast genome of *Allium victorialis* and its application for barcode markers. *Plant Breed. Biotech*. 5(3), 221-227(2017).