발간등록번호
11-1543000-001628-01

채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영

(Development and Establishment of Database System for Vegetable Breeding)

충 남 대 학 교

농림축산식품부 해양수산부 농촌진흥청 산림청

제 출 문

농림축산식품부장관 귀하

이 보고서를 “채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영”의 보고서로 제출합니다.

2017 년 2 월 14 일

프로젝트 연구기관명 : 충남대학교

프로젝트 책임자 : 임 용 표

연 구 원 : 최 수 련

연 구 원 : 이 소 영

연 구 원 : 김 윤 영

연 구 원 : 비그네쉬

연 구 원 : 우 효 나

연 구 원 : 이 소 남

연 구 원 : 주 광 생

연 구 원 : 방 문 성

연 구 원 : 파람스와리

연 구 원 : 고 낙 현

연 구 원 : 길 미 라
연 구 원 : 박 정 미
연 구 원 : 유 영 숙
연 구 원 : 윤 선 영
연 구 원 : 이 동 향
연 구 원 : 채 현 숙
연 구 원 : 최 종 만
연 구 원 : 한 희 순
연 구 원 : 시와난단
연 구 원 : 자 밀
연 구 원 : 강 동 현
연 구 원 : 김 다 솜
연 구 원 : 김 주 상
연 구 원 : 김 진 규
연 구 원 : 마 은 파
연 구 원 : 박 선 규
연 구 원 : 오 상 헌
연 구 원 : 이 성 호
연 구 원 : 임 수 빈
연 구 원 : 지 반
연 구 원 : 정 소 영
연 구 원 : 프라사스
연 구 원 : 홍 성 민

요 약 문

Ⅰ. 제 목

채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영

Ⅱ. 연구개발의 목적 및 필요성

1. 최종목표

○ 유전체 정보 통합 DB 구축으로 종자 개발에 있어 분자육종의 효율화를 지원.

○ 5대 채소작물(배추, 무, 수박, 고추, 파프리카)로부터 생산된 육종 특화 생물정보를 통합적으로 제공함으로써 각 채소작물 별 육종 활동에서의 시너지 효과를 유도.

○ 종자 개발에 있어 목표 형질 개체선발 및 계통선발, 목표 형질 접근 방법에 대한 효율적 체계의 구성.

2. 연차 과제별 목표

○ 한국 고유 식품인 김치의 주재료로서 이용되고 있는 배추는 우리나라 4대 채소작물 중의 하나로 국내에서 고추, 무, 양파에 이어 네 번째로 큰 종자시장을 이루고 있다. 십자화과 작물인 배추, 양배추, 무는 전체 종자 수출량의 약 46%를 차지하는 등 농업적, 경제적으로 매우 중요한 작물이다. 세계 채소종자산업의 현황을 보면, 국제종자시장에서 네덜란드, 미국, 프랑스 등의 종자회사에서 생산한 종자들이 높은 시장 점유율을 보이고 있으며 적극적인 M&A를 통해 대형화된 10대 다국적 종자기업으로부터 생산된 품종이 국제종자시장에서 67%의 점유율을 차지하고 있다. 뿐만 아니라 급속도로 발전하고 있는 여러 경제 작물의 유전체 연구 및 첨단 생명공학 기술이 재래적인 식물 육종 기술과 융합을 이루면서 세계 육종 환경에서의 GM작물 개발, 분자육종의 비중이 증가하여 왔다.

○ 2011년 배추과 작물의 경제적, 학문적 중요성으로 인해 국제 배추 유전체 염기서열 분석 프로젝트(Multinational brassica genome project)가 진행되어 배추의 inbred line 중 하나인 지부(chiifu)를 모델로 한 유전체 초안이 발표되었으며 이를 통한 배추 유전체 상의 유전자의 예측 및 이들에 대한 물리적 위치

가 특정되었다. 이러한 유전체 정보들은 대량의 분자 마커를 개발하기 위한 주요 기반이 되어 배추의 유전체 정보를 활용한 분자육종을 가속화하였다.

○ 국내 채소육종 기술과 육종가들이 가진 노하우는 세계적으로 매우 훌륭한 것으로 평가를 받아왔으나 IMF 이후, 다수의 국내 우량 종자회사들이 해외 거대 종자 회사에 합병되었다. 그 결과, IMF 이후 국내의 종자회사는 개인육종가 위주의 소규모의 회사 중심으로 영세화되었으며 육종에 있어서 분자마커, 종자생산 및 관리기술, 병리연구와 검정체계를 갖춘 기업은 극히 소수만 유지되었다. 근래에 들어 국가적 차원에서 종자개발을 위한 투자를 증진하고 국내에서 거대한 자본력을 갖춘 LG 그룹이 동부 팜한농을 인수하여 종자시장에 뛰어들고 노루에서 육종회사인 기반을 설립하고 연구설비의 확충 및 육종 전문인력의 신규채용을 시작하는 등 국내에서의 우리 종자 개발을 위한 관심과 기대가 점차 높아지고 있다.

○ 본 과제에서는 배추, 무, 수박, 고추, 그리고 파프리카를 국내외 채소 시장에서의 경제적 중요성이 큰 채소작물로 지정하여 이들의 분자육종을 활성화하고 우량 품종의 육성을 촉진을 목표로 한 채소육종 특화 데이터베이스에 대한 구축을 진행하였다. 이를 위해 지금까지 공개된 5대 채소작물의 표준 유전체 정보와 계통별 변이 데이터를 수집 및 재생산하여 신규 분자마커의 개발을 위한 기반을 마련하였다. 또한 채소작물의 다양한 표현형 및 형질 검정을 대상으로 한 분자마커 정보와 같은 다양한 육종관련 정보를 수집하고 이를 데이터베이스에 반영하여 국내의 육종가 및 기업과 같은 육종주체들로부터 예상되는 육종특화 정보접근에 대한 수요를 충족시킬 수 있는 환경을 제공하고자 하였다. 이를 통해 품종개발에 대한 효율성을 제고하고 분자육종을 통한 신품종 육성에 소요되는 육종연한을 단축을 유도하여 단기적으로 국내 시장에서 우리종자의 시장 점유율을 높여 해외종자 사용에 의한 로얄티를 점진적으로 감축해나가고 장기적으로는 세계 종자시장의 벽을 넘어 우리종자의 수출로 인한 국부의 증진에 기여하고자 한다.

가. 1차 년도

데이터베이스는 논리적으로 연관된 하나 이상의 자료의 모음으로 그 내용을 고도로 구조화하여 자료의 검색과 갱신의 효율화를 추구하는 시스템이다. 현재까지 다양한 종류의 정보가 구조화되어 전 세계의 관련 인력들에게 해당 정보를 제공하여 부가가치를 창출하고 있다. 국내의 육종에 종사하는 인력들이 채소작물의 육종을 목적으로 생산 및 수집한 정보를 구조화한 체계의 지원을 받는다면, 이는 국내의 채소 작물의 육종 활성화에 기여할 수 있는 하나의 수단으로 활용될 것으로 기대할 수 있다. 따라서 이를 실현하기 위한 5대 채소 작

물(배추, 무, 수박, 고추, 파프리카)의 육종 특화 데이터베이스의 구축을 목표로 하였다.

5대 채소작물들 중에서 배추를 모델로 하여 이의 육종에 이용될 것으로 기대되는 정보로서 배추의 유용 형질 조사 결과와 육종 소재의 정보, 유용 형질 관련 분자마커, 계통별 re-sequencing 결과로부터 산출된 SNP 정보를 선정하였다. 그리고 이를 데이터베이스라는 체계에서 표현할 방식을 구상하여 '채소 작물의 종자개발을 위한 육종 특화 통합 DB'의 프레임을 구성과 관련 정보의 수집을 목표로 계획을 수립하였다.

나. 2차 년도

2차 년도는 1차 년도에서 시작한 배추의 육종 관련 데이터의 생산과 수집을 지속하고 해당 결과를 데이터베이스에 update시키는 계획을 수립하였다. 따라서 육종에 필요한 형질별 분자마커의 정보와 과제 진행과정에서 생산한 전사체의 발현량 데이터와 계통별 변이정보를 생산 및 확충을 목표로 하였다. 이를 통해 데이터베이스내에서 활용할 수 있는 정보의 양을 확대하고 확보한 배추의 육종 관련 정보의 시각화를 통해 얻을 수 있는 자료를 구체화 및 다변화하여 데이터베이스에 반영하고자 하였다.

다. 3차 년도

1990년에 시작된 인간 유전체 프로젝트 이후, whole genome sequencing project는 애기장대, 초파리, 쥐와 같은 모델 생물체를 시작으로 진행되었고 이들과 근연관계인 경제적 가치를 지닌 생물들을 대상으로 확장되어 현재에도 관련 프로젝트가 활발히 진행되고 있다. 이를 계기로 염기 서열의 분석으로 유전자를 발굴하고, 형질과 유전자 사이의 연관성을 식별하는 연구(Genome Wide Assocation Study: GWAS)가 시작되었다. 3차 년도에서는 배추의 육종 관련 정보의 수집과 생산을 지속하면서 배추 수집단의 201개 계통을 대상으로 수행된 re-sequencing 데이터와 23개 표현형 정보를 활용하여 GWAS 수행을 위한 체계를 구축하고자 하였다.

라. 4차 년도

수박은 박과(Cucurbitaceae)에 속하는 작물로서 전세계적의 채소 재배 면적의 21.7%를 차지하고 있는 경제 작물중 하나이다. 수박의 세계 종자시장에서는 각 판매 지역 및 국가의 소비자의 기호에 따른 차이가 있지만 호피 단타원형, 호피 원형, 씨 없는 수박, 무지 원형 등 크게 4가지로 품종 개발이 이루어지고 있다. 수박에 대한 whole genome sequence 정보와 관련 정보는 오이와 멜론의 사례에 이어 2013년에 공개되었다. 국내에서는 공개된 수박의 유전체 정보를 활용하여 당도, 과육, 기능성 물질, 내병성, 종자 유무, 웅성불임성 등에 중

점을 두어 수박의 육종 활동이 진행되고 있다. 4차 년도에서는 구성한 배추의 육종 특화 데이터베이스의 구축 사례를 따라 국내의 수박 육종을 지원하고 활성화를 유도할 수 있는 수박을 대상으로 한 육종 특화 데이터베이스의 구축을 진행하고자 하였다.

Ⅲ. 연구개발 내용 및 범위

1. 유용 형질의 식별과 관련 정보의 수집 및 생산

국내 채소의 분자육종을 활성화하여 신품종 육성을 촉진을 목적으로, 국내 육종 환경에서 5대 채소작물의 육종 프로그램에 포함되는 형질 및 요소들에 관련된 정보를 하나로 통합하여 데이터베이스의 형태로 공개하고자 한다. 이를 위해 유용 표현형을 나타내는 특정 계통의 정보 및 소재 파악, 분자마커 관련 연구 사례, 그리고 특허에 대한 정보를 대상으로 정보 수집을 수행할 것이다.

또한 배추를 중심으로 수집한 분자마커의 다형성 검증을 본 연구기관이 보유한 배추 수집단의 유전자원을 통해 수행할 것이며 non-SNP 마커의 경우는 SNP 범용마커로 전환하여 논문 게재 및 특허 출원 및 등록과 신규 품종 육성 촉진과 같은 학술 및 경제적 부가가치 창출에 기여할 수 있도록 할 예정이다.

향후 형태적 표현형 관련 분자마커의 개발과 유전연구를 위한 기반 자료를 구성할 계획이다. 이를 위해 배추 수집단에 대한 반복적인 표현형 조사를 수행하고 이를 전산화하여 mapping 집단을 작성하기 위한 부모본의 선발 및 GWAS 분석을 위한 input 데이터의 생산 등에 활용할 것이다.

2. 채소의 종(species)내의 계통별 Re-sequencing과 GWAS 분석 체계의 도입

SNP(Single Nucleotide Polymorphisms)는 하나의 집단내의 계통들의 유전체정보로부터 인출된 동일한 유전적 영역상의 DNA 염기서열을 서로 비교했을 때 나타나는 특정 위치의 단일염기가 다형성을 나타내는 현상을 의미한다. 이는 SSR(Simple Sequence Repeats)과 InDel(Insertion and Deletion)에 비해 상대적으로 변이의 분포가 전체 유전체상에 고르게 나타나는 것으로 알려져 있으며 육종에서는 형질 연관 마커 및 고밀도 유전자 지도 작성에 유용한 것으로 알려져 있다.

이와 같은 SNP 데이터의 대량 발굴을 위해서는 채소의 계통별 re-sequencing을 진행하고 얻어진 데이터를 reference인 계통의 whole genome sequence와 비교하여 variant calling을 수행할 필요가 있다. 또한 1차적으로 얻어진 SNP 데이터들 중에서 마커로 안정적으로 활용할 수 없는 low-quality SNP를 제거하는 filtering 방식이 강구되어야만 한다.

또한 별도로 생산한 채소의 수치화한 표현형 데이터를 re-sequencing을 통해 얻은 변이데이터와 결합하여 GWAS를 수행하기 위한 작업체계를 확립하고자 하였다. 그리고 GWAS의 분석 결과를 computational method를 통해 형질에 관련된 변이를 식별하고 이를 구조 상에 포함하는 유전자를 확인하는 절차를 구성할 필요가 있다.

3. 유용 유전자 식별을 위한 RNA-seq과 DEG 분석

근래에는 NGS(Next Generation Sequencing) 기술의 보편화됨으로써 경제적 중요성을 지닌 생물(organism)들을 중심으로 한 전장유전체 해독 프로젝트가 시작되고 이들에 대한 유전체 지도 정보가 공개되기 시작했다. 이는 mRNA를 직접적으로 시퀀싱하여 그 발현량을 확인할 수 있는 RNA-seq 기술과 시너지를 이루어 유전체 지도 상에 위치한 모든 유전자에 대한 발현량을 확인할 수 있는 기반을 마련하였다. 또한 근연 종의 유전체 정보가 없는 생물의 경우에도 de novo assembly와 gene annotation을 통해 유전자의 식별 및 발현량 확인이 가능하다. 이를 통해 하나의 샘플에 대한 RNA를 추출하여 샘플이 처한 조건에서 발현하는 전사체의 전체적인 발현양상을 확인할 수 있게 되었다.

이를 통해 서로 다른 조건하의 샘플에 대한 RNA-seq을 수행하고 그 결과를 비교함으로써 조건에 대해 공통적 혹은 특이적으로 발현하는 유전자를 식별할 수 있게 되었다. 또한 이는 특정 형질 관련 candidate gene의 식별을 목표로 한 mapping 집단의 작성과 fine mapping을 통한 결과를 전사체적 연구의 측면에서 보완 및 지지를 할 수 있다.

본 과제에서는 다양한 채소작물의 표현형 관련 RNA-seq 데이터를 생산 및 수집을 위한 작업 절차를 구성하여 특정 조건에서 나타나는 DEG(Differentially Expressed Gene)를 식별하고자 한다. 이를 통해 특정 형질 및 표현형의 육종에 활용될 수 있는 유전자를 쉽게 선발하고 특정 형질의 선발을 위한 분자마커 개발에 활용할 수 있도록 할 것이다.

4. 육종 관련 수집 정보와 재생산된 생물정보를 활용한 데이터베이스 구축

웹데이터베이스의 구성체계로서 운영체제(Linux/Windows), Apache, Mysql, 그리고 php로 구성되는 LAMP/WAMP가 채택되었다. 과제 수행기간 동안 생산 및 수집된 채소의 육종 관련 정보들 간의 관계를 합리적으로 설정하고 정보의 검색과 인출을 효율적으로 수행할 수 있는 스키마(schema)를 구성하고자 하였다. 또한 스키마에 의해 저장된 채소의 육종 관련 데이터에 대하여 사용자의 질의(query)를 접수하고 이에 대응하는 Mysql에 저장된 정보를 인출하여 표 혹은 그림과 같이 시각화가 수행된 형태로 결과를 반환하는 html/php 스크립트를 개발을 수행하였다.

Ⅳ. 연구개발결과

1. 배추 육종 특화 데이터베이스 구축을 위한 육종 기반 정보의 생산 및 수집

가. 배추 유전체 정보의 수집과 annotation

데이터베이스에서 배추 모든 표준 유전자의 기본 정보로서의 활용 및 변이 데이터 생산 및 DEG 산출 과정에서 reference로 사용할 목적으로 현재 공개된 배추 유전체의 정보를 확보하였다. 지금까지 식별된 41,020개의 배추 표준 유전자에 TAIR ID, gene description, PFAM(Protein Family), PANTHER, KOG(euKaryotic Orthologous Groups), KEGG(Kyoto Encyclopedia of Genes and Genomes), 그리고 GO(Gene Ontology)와 같은 annotation을 부여하여 유전자의 검색과 기능 확인을 위한 환경을 조성하였다.

나. 배추 유용 형질에 대한 표현형 정보의 생산

향후 배추 육종 프로그램에서 유용한 목표 형질로서 기대되는 항목으로 Carotenoid, Flavonol, Vitamin C, Glucosinolate, Reducing Sugar, Mineral을 선정하였다. 자체적으로 보유한 계통들 내에서 해당 항목들에 대한 함량을 조사하여 유용 형질의 함량이 높은 계통을 각 항목 별로 선발하였다. 또한 배추 수집단이 보이는 다양한 morphological trait 23개 형질에 대한 3년간의 생육조사를 통해 해당 형질에 대한 표현형 데이터를 생산하였다.

다. 배추 유용 형질 관련 분자마커 정보의 수집 및 생산

육종가들이 각자의 육종 프로그램에 반영할 것으로 기대되는 다양한 형질에 관련된 유전자 정보와 분자마커 정보를 수집하였다. 그 결과, 내병성 형질에 대해서는 뿌리혹병, TuMV, 노균병, 무름병에 대한 정보를 수집하였으며 그 외의 형질로는 웅성불임, 자가불화합성, 개화기, 초장, 종피색, 잎털, 환경 저항성 그리고 Glucosinolate에 대한 정보가 수집되었다.

유용 형질 마커의 수집 이후 일부 비 SNP 마커와 연관된 배추 유전체상의 영역에 대한 배추 수집단에서의 SNP 변이를 조사하여 해당 마커의 SNP 마커로의 전환을 도모하였다. 그 결과, 잎털, 결각, GMO 검정에 대하여 각각 2개, 3개, 2개의 SNP 프라이머를 제작하였다.

2. 배추 육종 특화 데이터베이스 구축을 위한 생물정보의 생산과 재가공

가. 배추의 계통별 변이 정보의 생산

1차 년도부터 시작한 배추 수집단에서의 high-quality 변이 정보의 생산은 현재 총 201개 계통(*B. rapa* spp. pekinensis: 148 계통, *B. rapa* RIL 집단: 26계통, 기타: 27계통)에 대하여 수행되었다. 그 결과 배추의 모델 계통인 지부

(chiifu)에 대해 137X, 2개의 엘리트 계통에서 30X, 다른 7개의 엘리트 계통에서 10X 그 외 계통에서 3~5X genome coverage를 갖는 re-sequencing 데이터를 생산하였다. 계통별로 생산한 re-sequencing 데이터를 GATK 파이프라인을 통해 배추 표준 유전체에 개별적으로 대조하여 계통별 변이 정보를 얻고 이를 통합하여 배추 수집단에서 총 1,888,669개의 SNP position에 대한 정보를 얻을 수 있었다. 그리고 이를 조사한 배추의 23개 표현형 정보와 연결시켜 배추의 유용 형질에 대한 GWAS 분석을 실현할 기반을 조성하였다.

나. 배추의 조직별 전사체 정보의 생산

배추의 5개 조직(어린 잎, 화기, 뿌리, 결구 내엽, 결구 외엽)에서 RNA의 추출 후 cDNA 라이브러리의 합성과정을 거쳐 배추의 조직별 RNA-seq을 수행하였다. 또한 RNA-seq으로부터 생산한 배추의 조직별 transcriptome read 데이터로부터 발현량을 산출하기 위한 read의 quality check, quality trimming, read alignment, expression value calculation, 그리고 normalization 순으로 수행되는 파이프라인을 확립하였다. 조직별로 산출한 발현량을 어린 잎을 control로 하여 다른 4개 조직에 대한 DEG를 추출하였을 때 화기, 뿌리, 결구 내엽, 결구 외엽에서 각각 2,442, 2,646, 268, 866개의 조직 특이적으로 발현하는 유전자를 식별할 수 있었다. 또한 배추 산물의 노화에 관여하는 유용 유전자의 식별을 위해 수확 후 노화 양상에 큰 차이를 보이는 두 계통을 식별하고 수확 전과 수확 4주 후로 시기를 나누어 샘플을 채취하고 이에 대한 RNA-seq을 수행하였다.

3. 생산 및 수집한 생물정보에 기반한 배추의 육종 특화 데이터베이스 구축과 운영

과제수행 과정에서 생산 및 수집한 배추 육종 관련 정보의 활용 및 시각화를 위하여 1차 년도에는 배추의 육종 특화 데이터베이스의 기본 프레임을 구성하였다. 해당 데이터베이스는 3차 년도까지 생산한 배추의 육종 관련 정보 및 생물정보를 반영하였다. 배추 유전체 및 육종 관련 정보에 대한 접근을 위한 기능으로서 배추 육종 특화 데이터베이스는 지놈 브라우저, web-BLAST 시스템 그리고 키워드 중심의 search 옵션을 보유하고 있다.

지놈 브라우저를 통하여 육종가가 관심을 가지고 있는 특정 배추의 표준유전자에 대하여 sequence 정보, 변이 및 발현량 등 입력된 모든 정보를 열람할 수 있다. 또한 수집 및 생산한 MAS(marker assisted selection)를 위한 분자마커 정보를 배추의 10개 chormosome 상에 시각화하고 분자마커의 sequence를 화면 하단에 정리하여 육종가들이 쉽게 표현형 관련 마커를 선발할 수 있도록 하였다. 그리고 지부(chiifu)와 권심(kenshin)을 부모본으로 작성한 RIL 집단의 SNP 정보를 활용하여 육종가들이 MAB(Marker assisted back-crossing)에 필요한 마커 정보와 BIN map을 제공한다.

4. 배추의 유용 유전자 발현량 데이터베이스 (BrTED)의 구축과 운영

전세계의 연구진들에 의해 농업적으로 유용하게 이용 가능한 배추의 다양한 형질들에 대한 전사체 데이터는 현재 NCBI 등 거대 생물정보 데이터베이스에서 열람 및 이용이 가능하다. 그러나 일반적인 연구자들이 이러한 전사체 데이터에서 각자가 필요한 데이터의 가공과 해석에는 전문적인 지식과 상당한 수준의 computational resource의 투입이 필요하다. 이러한 일반적인 연구자가 겪을 수 있는 기술적 장벽의 해소를 목표로 배추의 전사체 데이터베이스를 구축하였다.

총 10개의 공개된 실험들로부터 다양한 조건을 반영한 92개의 전사체 데이터를 수집하여 조건별 유전자 발현량을 산출하고 데이터베이스에서 조건별로 추출할 수 있는 DEG에 배추의 41,020개의 유전자의 annotation을 연결시켰다. 또한 산출된 DEG의 KEGG와 GO에 대하여 enrichment 분석을 하는 체계를 데이터베이스 플랫폼 내에 구축하여 이용자가 활용할 수 있도록 하였다.

5. 수박의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영

수박의 whole genome sequencing project의 모델로 이용된 97103의 유전체 정보를 최신 정보를 이용하여 annotation을 수행하였다. 그리고 이를 바탕으로 수박의 20개 계통의 SNP에 대한 변이데이터와 과피, 과색, 웅성불임성과 같은 형질에 대한 발현량 데이터를 생산하였다. 생산된 모든 데이터는 배추를 기준으로 구성한 데이터베이스 플랫폼에 입력되어 현재 지놈 브라우저와 web-BLAST를 통해 유전자의 검색과 관련 정보의 시각화가 가능하다.

Ⅴ. 연구성과 및 성과활용 계획

1. 연구성과

가. 연차별 연구성과 목표 및 달성

(단위: 건 수)

구분		논문		특허		인력 양성	데이터베이스 구축 사례	분자 마커	기타
		SCI	비SCI	출원	등록				
1차 년도	목표								
	달성			1			1		
2차 년도	목표	1							
	달성	1		2	2			1	GMO 마커검정
3차 년도	목표		1						
	달성	7		4	1	2		6	GMO 마커검정
4차 년도	목표	1					1		
	달성	3			3	3	1		
계	목표	2	1						
	달성	11	0	7	6	5			

나. 논문게재 성과

○ 논문게재 SCI 11건

순번	발간연도	논문명	주저자	학술지명	Vol (No)	구분
1	2014	Genetic Detection of Clubroot Resistance Loci in a New Population of *Brassica rapa*	Wenxing Pang, Shan Liang, Zhongyun Piao	Horticulture, Environment, and Biotechnology	55(6)	SCI
2	2015	Quantitative trait loci mapping of partial resistance to Diamondback moth in cabbage (*Brassica oleracea* L)	Nirala Ramchiary, Yong Pyo Lim	Theoretical and Applied Genetics	128(6)	SCI
3	2015	Mapping QTLs of resistance to head splitting in cabbage(*Brassica oleracea* L. var. *capitata* L.)	Wenxing Pang, Xiaonan Li, Yong Pyo Lim	Molecular Breeding	35(5)	SCI
4	2015	The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases.	Arne Schwelm, Yong Pyo Lim, Jutta Ludwig-Müller, Christina Dixelius	Scientific Reports	10 (1038)	SCI
5	2015	Construction of chromosome segment substitution lines enables QTL mapping for flowering and morphological traits in *Brassica rapa*.	Xiaonan Li, Yong Pyo Lim, Zhongyun Piao	frontiers in plant science	6	SCI
6	2015	The 2015 KSM-ICWG-GSP Joint Clubroot Symposium Meeting Report.	Vignesh Dhandapani, Yong Pyo Lim	Journal of Plant Growth Regulation	34(2)	SCI
7	2015	Anatomic Characteristics Associated with Head Splitting in Cabbage (*Brassica oleracea* var. *capitata* L.).	Wenxing Pang, Yoon-Young Kim, Yong Pyo Lim	PLOS ONE	10(11)	SCI
8	2015	Genomic and Post-Translational Modification Analysis of Leucine-Rich-Repeat Receptor-Like Kinases in *Brassica rapa*.	Jana Jeevan Rameneni, Yeon Lee, Man-Ho Oh, Yong Pyo Lim	PLOS ONE	10(11)	SCI
9	2016	Quantitative Trait Loci for Morphological Traits and their Association with Functional Genes in *Raphanus sativus*	Xiaona Yu, Su Ryun Choi, Yong Pyo Lim	frontiers in plant science	7	SCI
10	2016	Genome-Wide Analysis and Characterization of Aux/IAA Family Genes in *Brassica rapa*	Parameswari Paul, Vignesh Dhandapani, Yong Pyo Lim	PLOS ONE	11(4)	SCI
11	2016	Genome wide identification and functional prediction of long non-coding RNAs in *Brassica rapa*	Parameswari Paul, Yong Pyo Lim	Genes & Genomics	38(6)	SCI

다. 특허 성과

○ 특허실적 13건 (특허출원 7건, 특허등록 6건)

순번	구분	출원여부	년도	특허명	출원인	발명인	출원(등록)번호
1	특허	출원	1	칼슘 함량이 증가된 배추 품종 및 이의 육종방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-2013-0153649
2	특허	출원	2	배추좀나방 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-2014-0062420
3	특허	출원	2	열구 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-2014-0062417
4	특허	등록	2	배추좀나방 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-1474910
5	특허	등록	2	열구 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-1474914
6	특허	출원	3	열근 무 품종을 구분하기 위한 특이 마커 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-2015-0015252
7	특허	등록	3	칼슘 함량이 증가된 배추 품종 및 이의 육종방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1557420
8	특허	출원	3	배추 글루코시놀레이트 함량 조절인자 기반의 SNP 마커 및 이의 용도	충남대학교	임용표, 이소남, 비그니쉬 단다파니, 최수련, 강동현, 정소영, 박선규	10-2015-0190309
9	특허	출원	3	초장이 짧은 야생무로부터 초장이 긴 개량무 품종을 특이적으로 구분하기 위한 프라이머 세트 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-2015-0188806
10	특허	출원	3	해조류 추출물을 이용한 위타니아 솜니페라 유래의 스테로이드계 락톤의 생산 방법	충남대학교	임용표, 시와난단 가내산	10-2015-0142520
11	특허	등록	4	글루코브라씨신 함량이 증가된 배추 품종 및 이의 육종 방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1602022
12	특허	등록	4	글루코나스튜틴 함량이 증가된 배추 품종 및 이의 육종 방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1602536
13	특허	등록	4	열근 무 품종을 구분하기 위한 특이 마커 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-1627197

라. 인력활용/양성 성과

순번	양성일자	연구기관	학위	성별	국적	비고
1	2015.1.12	충남대학교	석사	남	대한민국	
2	2015.12.10	충남대학교	석사	남	대한민국	
3	2016.2.25	충남대학교	박사	여	인도	
4	2016.2.25	충남대학교	석사	남	대한민국	
5	2016.8.25	충남대학교	석사	남	대한민국	

마. 데이터베이스 구축 성과

순번	구축년도	연구기관	데이터베이스명	URL	비고
1	2013	충남대학교	채소작물의 육종 활성화를 위한 데이터베이스	www.vegetable.or.kr	2017년 기준, 생산 및 수집된 배추와 수박의 육종관련 정보가 반영
2	2016	충남대학교	BrTED(*Brassica rapa* Transcriptome Expression Database)	brted.cnu.ac.kr	2017년 DATABASE지에 submit 예정

바. 분자마커 개발

특성 분류	순번	기보고된 후보 유전자명	분자마커 형태	SNP 마커 전환 및 범용성 평가
GMO	1	CaMV 35S promoter	gene base	SNP 마커 개발
종피색	2	TTG1	SCAR	SNP 마커 개발 완료(잎털과 동일)
잎털	3	TTG1	SSR	SNP 마커 개발 완료(종피색과 동일)
결각	4	GASA4a	SNP	SNP 마커 개발 및 범용생 평가 완료
	5	GASA4b	SNP	SNP 마커 개발 및 범용생 평가 완료
	6	AGL12	SNP	SNP 마커 개발 및 범용생 평가 완료

사. 분자마커 서비스

순번	실시년도	연구기관	분석 내용	비고
1	2014	충남대학교	시중 판매 중인 양배추 F1에 대한 GMO 검정	이상 없음
2	2015	충남대학교	농우 바이오 시판 품종 4 작목에 대한 GMO 검정	이상 없음

2. 성과활용계획

가. 배추 육종 특화 데이터베이스 구축을 위한 육종 기반 정보의 생산 및 수집

채소작물의 육종에 활용할 수 있는 성과로서 5편의 SCI급 논문의 게재와 12건의 특허실적을 달성하였다. 해당 성과들은 과제 수행을 통해 구축한 데이터베이스를 통해 연구자 및 육종가들의 접근이 가능하도록 서비스하여 국내의 채소작물의 분자육종의 활성화를 실현할 수 있도록 할 것이다. 또한 과제 수행 중의 연구에서 개발한 분자마커 및 관련정보들을 분자육종 기술이 부재한 종자회사에게 마커 검정 서비스를 제공할 수 있을 것으로 기대된다.

(1) 배추 수집단 표현형 정보 조사 결과

본 실험실에서는 3년에 걸쳐 배추 수집단의 생육조사를 수행하였다. 해당 결과는 23개의 표현형에 대한 정량적 및 정성적 결과를 가지며 정성적 결과의 경우 index에 따른 값을 부여 받아 정량화 되었다. 이 생육조사 결과는 전산화되어 23개 표현형에 대한 유전연구의 부모본 선발과 mapping 집단의 전개 및 GWAS의 표현형 정보로서 활용되어 표현형 관련 candidate gene의 식별에 활용될 수 있을 것으로 기대된다.

(2) 배추 유용형질의 분자마커 수집 및 재생산

기보고된 유용 표현형관련 분자마커가 지정하는 유전자의 위치는 현재 공개된 배추의 표준 유전체 정보를 통해 쉽게 접근이 가능하다. 수집한 분자마커의 세부사항을 확인하고, 본 실험실에서 작성한 배추 수집단의 표현형 및 유전적 변이 정보와 수집단의 계통별로 추출한 DNA를 활용하여 기존의 분자마커를 SNP 범용마커로 전환할 수 있다. 개발 및 검증이 종료된 범용마커는 육종공동체에 보급되어 신속한 목표 형질의 선발 및 신품종 육성에 기여할 수 있다.

나. 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스 구축을 위한 생물정보의 생산과 재가공

(1) 채소작물 표준 유전체 정보를 활용한 gene annotation

현재까지 알려진 배추 표준 유전자 41,020개와 수박 표준 유전자 23,440개에 대한 gene annotation을 수행하였다. gene annotation에는 KEGG, KOG, GO, PANTHER, uniprot annotation 등 배추와 수박 genome 데이터베이스에서 다루지 않는 추가적인 annotation을 부여하여 사용자가 추가적인 검색 없이 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스를 허브로 삼아 관련 정보를 쉽게 열람할 수 있다. 향후 본 데이터베이스에서 다루는 5대 채소작물의 표준 유전체 정보가 갱신되면 이를 기반으로 재생산된 데이터 또한 분석

을 신규 유전체 정보를 반영하여 재처리하여 갱신된 분석정보를 사용자들에게 제공할 계획이다.

(2) 배추 수집단의 re-sequencing을 통한 GWAS 분석 기반의 구축

현재 배추 201계통 및 수박 20계통의 변이 정보가 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스에 반영되어 있다. 이를 통한 변이 정보가 재생산되어 배추와 수박의 수집 집단 내의 변이 정보를 현재 이용가능한 상황이다. 배추의 경우 3년간의 생육조사를 통한 23개의 표현형 정보가 전산화되어 있다. 배추 201 계통의 SNP 정보와 표현형 정보를 GAPIT을 통해 분석하면 표현형 별 GWAS가 가능하다. GAPIT의 분석 결과인 맨하탄 플롯과 SNP의 표현형 연관 분석 정보를 통해 표현형에 강하게 연관된 SNP를 보유한 유전자를 식별하고 이를 분자마커로 전환할 수 있다. 이를 통해 GWAS 기반의 형질 관련 SNP 마커의 대량 발굴 체계를 구축할 것이다. 또한 생산된 SNP 마커의 검정 체계를 통해 신규 SNP의 검정력이 평가를 받는 기준을 제시하여 이를 충족하는 분자마커의 특허출원을 실시할 것이다. 특허 출원 및 등록이 완료된 신규 분자마커의 정보는 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스의 분자마커 정보 검색 체계에 편입되어 육종가들에 의해 활용될 수 있도록 할 계획이다.

(3) 채소작물의 유용 유전자 식별을 위한 RNA-seq 기반 DEG 연구

채소작물의 조직 및 조건에 따른 전사체 발현양상의 비교분석을 통해 비교조합에 따른 DEG의 식별 체계가 채소종자의 육종 특화 데이터베이스와 배추 전사체 데이터베이스에 갖추어져 있다. 이를 통해 특정 형질에 관련된 기보고된 연구결과와 데이터베이스에서의 DEG 산출을 통한 candidate gene의 리스트를 쉽게 얻을 수 있다. 추후 DEG를 기반으로 한 네트워크 분석을 도입하여 표면적으로 보이는 표현형에 관련된 candidate gene을 조절하는 패스웨이 상의 upstream에 존재하는 유전자를 규명할 것이다. 그리고 GWAS 구축 과정에서 얻은 표현형별 변이 데이터와 RNA-seq 분석 결과를 통합하여 서로 다른 표현형을 보이는 계통간에서 특정 유전자상의 염기서열이 나타내는 변이가 유전자의 발현 양상을 지배할 수 있다는 것을 규명하고자 한다. 그리고 이러한 유전자 상의 변이를 범용 마커로 전환할 수 있는 체계를 구성하여 단순 표현형 연관 마커를 넘어선 유전자 발현 원리에 기반을 둔 근원적인 분자마커 생산을 위한 체계를 구성하고자 한다.

다. 생산 및 수집한 생물정보에 기반한 채소작물의 육종 특화 데이터베이스 구축과 운영

(1) 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스

현재, 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스는 5대 작물에 대한 기본 플랫폼이 완성되었고 1차에서 3차 년도까지 배추에 대한 데이터 구성 및 입력이 완료된 상태이다. 그리고 4차 년도에는 배추의 사례를 기반으로 수박으로부터 표준 유전체의 기본 정보와 annotation, 계통별 변이 정보 및 전사체 발현 정보를 입력한 상태이다. 입력된 정보는 지놈 브라우저를 중심으로 구성한 스키마에 따라 출력되며 BLAST 및 키워드를 통한 검색이 가능하다. 그러나 아직 수박에 대한 데이터 입력의 절대량과 분석 및 검색 기능이 배추의 사례에 비해 미진한 부분이 있다. 따라서 차후의 작업으로 배추와 수박의 데이터베이스 간의 완성도에 대한 차이를 좁혀나갈 것이며 이후 무, 고추, 그리고 파프리카에 대한 정보 생산과 수집을 실시하여 5대 채소 작물에 대한 데이터베이스 구성을 종결하고 향후 새로운 자료가 확보하는 즉시 이를 데이터베이스의 data pool에 반영해나갈 것이다. 이로써 국내의 채소 분자육종 공동체에 각 작물에 대한 육종 관련 정보를 통합적으로 제공하여 신규 마커의 개발과 신품종 육성의 촉진을 기대할 수 있다.

(2) 유용 형질의 DEG 기반 전사체 데이터베이스

배추(*Brassica rapa*)를 대상으로 한, 총 10개의 공개된 실험의 92개 전사체 데이터들이 수집되었다. 이를 통해 다양한 조건들에 대한 유전자의 발현량이 산출되었으며 이를 활용한 DEG 분석으로 KEGG와 GO에 대한 enrichment 분석이 데이터베이스내에서 실용화되었다. 이로서, 연구자 및 분자육종가들이 각자가 목적으로 하는 형질 관련 후보 유전자의 선정에 DEG 분석 결과를 사용할 수 있게 되었다. 앞으로도 BrTED는 새롭게 생산된 발현량 데이터를 지속적으로 생산 및 수집하여 더 많은 data pool을 연구 및 육종 공동체에 제공할 수 있도록 할 것이다. 또한 DEG 분석방식을 정밀화하여 더욱 정확한 분석 결과를 제공할 수 있도록 할 것이며 이를 기반으로 한 네트워크 분석을 도입할 예정이다.

SUMMARY

Ⅰ. Title

Development and Establishment of Database System for Vegetable Breeding

Ⅱ. Objectives and Necessity

1. Final objectives

○ Supporting the effectiveness of molecular breeding for new seed development by constructing informative database integrating genomic information of 5 main economic crops in korea.

○ Triggering the synergy effect for breeding activity for by releasing information which is related to breeding from the data created by 5 main crops(chinese cabbage, radish, watermelon, pepper, and paprika).

○ Constructing optimal system regarding line selection and finding candidate genes for specific traits to facilitate seed development effectively.

2. Objectives of annual projects

○ Chinese cabbage is one of the four major vegetable crops in Korea and forms the fourth largest seed market in Korea following pepper, radish and onion. Chinese cabbage, cabbage, and radish which are members of Cruciferous vegetables are the most important agricultural and economically important crops, accounting for about 46% of total national seed exports of korea. When we take a look the global vegetable seed industry, seeds produced by seed companies of the Netherlands, USA, and France show a high market share. Also, varieties which are produced from 10 large multinational seed companies that have been enlarged through aggressive M & A account for 67% of the seed market. Genome research and high-tech biotechnology of various economic crops, which are rapidly developing, have

been combined with conventional plant breeding techniques, and the proportion of GM crop development and molecular breeding in the world breeding environment has increased.

○ Due to the economic and academic significance of the Chinese cabbage crop in 2011, a multinational brassica genome project has been carried out. As a result, a genome draft based on the chiifu, which is one of the inbred lines of Chinese cabbage has been released to world-wide brasscia research and breeding community. The gene prediction on the genome and the physical location for each genes have been specified. These genomic information has become the main basis for the development of molecular markers, accelerating the molecular breeding using the genomic information of chinese cabbage.

○ Korean vegetable breeding techniques and breeders' know-how have been evaluated as excellent around the world, but since the IMF, a number of korean seed companies have been merged into oversea's large seed companies such as monsanto and syngenta. As a result, since the IMF, korean seed companies have been maintained by small breeding companies or individual breeders, and very few breeding companies have the system for molecular markers, seed production and management technology, pathology research and testing. In recent years, korean government has promoted investment for seed development at the unprecedented level. Also LG Group, which has huge capital power in Korea, merged Dongbu Palm Hanong and tried to enter the seed market. In case of Noroo, they established a breeding company names as 'Kiban', investing the budget for constructing R&A center and recruiting workers having experience for breeding. Like above case, there are growing interest and expectation for the development of seeds in Korea.

○ This project designated chinese cabbage, radish, watermelon, red pepper, and paprika as vegetable crops with high economic importance in the domestic and overseas vegetable market and aimed to construct database specialized to molecular breeding for facilitating the breeding of good varieties. To realize this concept, we collected standard genomic information from five major vegetable crops. Subsequently, we reproduced sequential variant calling data from raw data and established the foundation for the

development of new molecular markers. Also we tried to support the web-based environment which allows breeders to access information related to molecular breeding derived from our collected or re-organized data such as molecular marker information for phenotype data of various traits by managing the database focused on breeding Through this, it is possible to improve the efficiency of breed development and to shorten the breeding period required for breeding new varieties through molecular breeding, thereby gradually increasing the market share of our seeds in the domestic market and gradually reducing royalties by using overseas seeds. And in the long terms, It will contribute to the promotion of national wealth through the export of our seeds.

2-1. First year

A database is a collection of logically related data, organizing its contents with a highly structured way, and seeking to efficiently retrieve and update data. Until now, various types of information have been structured, creating added value by providing relevant information to related personnel around the world with specific purposes. If the work-forces involved in domestic breeding are supported by a structured system of information which is produced and collected with purpose of breeding vegetable crops, this will be used as a meaningful tool for the efficient promotion of breeding of domestic vegetable crops. Therefore, we aimed to establish a database of breeding of five vegetables (Chinese cabbage, radish, watermelon, red pepper, paprika).

In the first year, we selected chinese cabbage as a model case of constructing Database System for Vegetable Breeding for 5 economicr crops in korea. Subsequently, we decided creating category of necessary information which are able to be utilized in breeding program with phenotyping results for interesting lines and their availability in breeding programs, molecular marker information for specific purpose, and SNP data derived from re-sequencing data of each lines in crop's core collection. Finally, we designed a way to display data utilized for breeding in the web-database system

.

2-2. Second year

Second year's project will continue production and collcetion for data

related to breeding of chinese cabbage and update them to database. Therefore, we planned to enlarge transcriptome expression data and variants calling data from core collection with molecular marker information which is necessary to breeding program. Based on this strategy, we tried to enlarge the volume of information which is able to be displayed in database, and support diverse ways to visualize stored data in web-database structure in the second year project.

2-3. Third year

Since the human genome project, which began in 1990, the whole genome sequencing project has been started with model organisms such as Arabidopsis, Drosophila and Rats, and has been extended to and economically valuable organisms. Under this circumstance, The GWAS(Genome Wide Assocation Study) was started to identify the key genes for specific traits by analyzing the nucleotide sequences and to identify the relationship between the traits and the genes. In the third years, we tried to construct system for GWAS by re-sequencing data and phenotype data from 201 lines in core collection and 23 traits respectively, continuing works for collection and reproduction of information of chinese cabbage breeding.

2-4. Fourth year

Watermelon is a crop belonging to the genus Cucurbitaceae and is one of the economic crops accounting for 21.7% of the world's vegetable growing area. Although there are differences according to consumers' preferences in each sales region and country, world-wide watermelon's breeding program is nowadays focused on largely a series of kinds for traits such as stripes on surface, fruit shape, and existence of seeds. Whole genome sequence information and related information on watermelons were released in 2013, following the case of cucumber and melon. In the korean domestic breeding program for watermelon, it is emphasized on sugar content, fruit flesh, functional secondary metabolites, disease resistance, seed existence and male sterility, utilizing released standard genome information of watermelon lines 97103.

In the fourth year project, we tried to construct watermelon breeding database which is expected to vitalize domestic watermelon's molecular breeding industry, following the case of constructing and management of breeding database for chinese cabbage.

Ⅲ. Methods of studies

1. Identification of useful traits and the collection and production of relevant information

For efficient breeding new cultivar by molecular vegetable breeding in korea, we tried to release integrated results regarding phenotypes or elements which are utilized in breeding program of 5 domestically cultivated economic crops. To achieve this goal, first, we evaluated availability of useful lines in breeding program, and collected case study of molecular marker and patent focusing on breeding. Also, polymorphisms test for collected molecular marker was confirmed against lines belonging our own core collection. In case of released non-SNP markers, they were converted to SNP marker, anticipating additional values with two direction; academic and economic, by publishing the papers, patents and contributing breeding program of new cultivar.

We constructed basic data for development of molecular markers involved in MAS(marker assisted selelection) about interested phenotypes by repetitive phenotype evaluation from '*Brassica rapa* core colletion lines'. We largely used result of phenotype evaluation with two direction such as selecting parental lines for developing mapping population and creating input data for GWAS research.

2. Introduction of re-sequencing and GWAS analysis system for lines in collected vegetable populations

SNP refers to a phenomenon in which a single nucleotide at a specific position is polymorphic when DNA sequences on the same genetic region of the each lines in the population are compared with each other. It is known that the distribution of the variants is evenly distributed over the entire genome compared to SSR (Simple Sequence Repeas) and InDel(Insertion and Deletion), and it is useful to create high-density genetic map with this information.

In order to massively discover such SNP data, it is necessary to conduct re-sequencing of each plant and to perform variant calling by comparing the obtained data with the whole genome sequence of the reference. Additionally, a filtering method for eliminating low-quality SNP that are not able to stably be utilized as a marker must be considered.

We also tried to establish a pipeline for GWAS by combining the phenotypic data produced independently with the SNP matrix obtained through re-sequencing analysis. Also, it is necessary to construct a process to identify the sequence polymorphisms related to the trait and to evaluate effect

of the gene containing it in the structure through the computational method for analyzing results of GWAS.

3. RNA-seq and DEG analysis for finding candidate genes for useful traits

Recently, the NGS (Next Generation Sequencing) technology has become universalized, and a project to decode the whole genome of organisms having economic importance has started and comprehensive information of genetic map about them has begun to be released. This provides a basis for confirming the expression level of all the genes located on the genome map by synergy with the RNA-seq technology for directly sequencing mRNA and confirming the expression level. In addition, even in the case of organisms that do not have genomic information of their relatives, it is possible to identify genes and their expression levels through de novo assembly and gene annotation. By this techonology, identification of overall expression profiling of transcriptome is available through RNA extraction from each samples.

By this system, we are able to identify genes expressed commonly or specifically from the compared results of RNA-seq expression profilings which generated from samples under different conditions. Also, this system can support and complement the result of candidate genes derived from the QTL results from the mapping population and fine mapping.

We tried to identify the DEG(Differentially Expressed Gene) appearing under specific conditions by constructing a work process for the RNA-seq data representing various vegetable crop's phenotype which we produced or collected.

4. Constructing database utilizing collected information and reproduced bioinformatics results for breeding

LAMP/WAMP consisting of operating system(Linux/Windows), Apache, Mysql, and php was adopted as the structure of web-database. We established a reasonable relationship within the all available information, which were produced or collected during the project, and constructed a schema that is able to efficiently search and retrieve information in the database system. Also, we developed html/php script which visualize returned results with table or figure, interacting with user's query for deposited crop's breeding data in database system.

Ⅳ. Results and conclusion of the research

1. Producing and collecting basic information for the database focused on chinese cabbage breeding

1-1 Collecting basic information of *Brassica rapa* genome and its annotation

We obtained standard information for the *B. rapa* genome which is currently available to display its information as a basic data for all genes in the database and utilize it for variant calling and expression profiling as a reference. We conferred the annotation such as TAIR ID, gene description, PFAM(Protein Family), PANTHER, KOG(euKaryotic Orthologous Groups), KEGG(Kyoto Encyclopedia of Genes and Genomes), and GO(Gene Ontology) for identified 41,020 genes in *B. rapa* genome to develop the search options for all available *B. rapa* genes.

1-2 Producing phenotype data of useful traits in chinese cabbage breeding

We selected Carotenoid, Flavonol, Vitamin C, Glucosinolate, Reducing Sugar, and Mineral as a list for potential traits of breeding programs in the future. After investigating contents of each subject, we selected lines showing high contents for each subjects in our own germplasm. Additionally, we produced phenotype data for 23 morphological traits within our core collection of the chinese cabbage by 3 years phenotyping works.

1-3 Collecting molecular marker information regarding useful traits of *B. rapa* breeding program and converting to SNP marker format

We collected diverse molecular marker informations and their target genes which are expected to be respected to breeding programs by breeders or seed company. As a result, we collected marker information for pathogen resistance(club root disease, downy mildew, TuMV, and soft rot), other traits(male sterility, self incompatibility, flowering time, plant height, seed coat color, trichome, environment resistance, and glucosinolate).

After collecting marker information for useful traits in the vegetable breeding program, we investigated SNPs on the target genomic region of collected marker withing our re-sequencing data pool of chinese cabbage core collecion, and tried to convert released markers into SNP marker. As a resutl we developed novel SNP markers for a series of traits; trichome, seed coat color, lobation and GMO evaluation.

2. Bioinformatics researches for constructing database focusing on breeding of chinese cabbage

2-1. Variants calling from chinese cabbage core collection

The production of high-quality variants information on the chinese cabbage genome was started from the first year project. Now, this process was applied to 201 lines in our core collection(*B. rapa* spp. *Pekinensis*: 148 lines, *B. rapa* RIL population: 26 lines, others: 27 lines). As a result, we produced re-sequencing data with genome coverage of 137X for chiifu, 30X for two elite lines, and 10X for other elite lines and 3-5X for others of chinese cabbage. The re-sequencing data generated by each lines were individually compared with the standard genomes of chinese cabbage through the GATK pipeline, and the variant calling results of each lines was integrated. Finally, a total of 1,888,669 SNP positions in the cabbage group were calculated by our own re-sequencing data from chinese cabbage core collection. Also we initiated to create the basic system for performing GWAS research by combining analysis result of re-sequencing data and phenotype data from the 23 traits.

2-2. Producing transcriptome data of chinese cabbage

After extracting RNA from five tissues of Chinese cabbage(young leaves, flowers, roots, inner leaves in the head, outer leaves on the head surface), cDNA library was synthesized and RNA-seq of each cabbage was performed. Subsequently, the pipeline was constructed in the order of read quality check, quality trimming, read alignment, expression value calculation, and normalization for calculated expression level from the transcriptome read data of the chinese cabbage produced from RNA-seq. When we set the young leaf as the standard and calculated DEG for other 4 tissues, we found 2,442, 2,646, 268, 866 DEGs from flowers, roots, inner leaf, and outer leaf respectively. Additionally, we conducted RNA-seq for two lines showing the strong difference for senescence aspect to identify key genes involved in product senescence of chinese cabbage. For the sampling, we extracted RNA from leaves before harvest and repeated identical work with 4 weeks old leaves after harvest.

3. Database construction and management based on results of bioinformatics researches for chinese cabbage.

In order to utilize and visualize the information related to the production and

collection of information for chinese cabbage breeding, a basic frame of the specialized database for chinese cabbage breeding was constructed in the first year. This database reflects the breeding information and bioinformatics results of chinese cabbage breeding which have been produced until the third year project. Chinese cabbage breeding database contains genome browser, web-BLAST and search options based on keyword system to allow users to access into information regarding standard genome and breeding for chinese cabbage.

Through the genome browser, all the deposited information in the database such as sequence information, variant calling data, and gene expression data can be browsed with respect to a interested gene in the standard chinese cabbage genome. In addition, molecular marker information for marker assisted selection (MAS) was collected and visualized on 10 chormosomes of chinese cabbage. Sequences of molecular markers are arranged at the bottom of the screen so that breeders can easily select phenotype related markers and utilize them to their work. Also, We utilized SNP information derived from RIL population which was developed by two parental lines; chiifu and kenshin to support information regarding MAB(Marker assisted back-crossing) and BIN map.

4. Construction and management for BrTED.

Transcriptome data for various varieties of chinese cabbage that are available with agricultural purpose from world wide researcher's work are currently available for viewing and utilizing from the macrobiological information databases such as NCBI. However, general researchers don't have knowledge and a considerable amount of computational resources to process analysis and interpretation for the deposited transcriptome data. A database of transcriptome of chinese cabbage was constructed with the aim of solving the technical barriers that general researchers might face to.

Total 92 transcript data reflecting various conditions were collected and used for calculating DEGs within diverse combination of samples from the total of 10 published experiments. The annotation of 41,020 genes of chinese cabbage was connected to DEGs which are extracted from the specific conditions in the database. In addition, a system for enrichment test of KEGG and GO from calculated DEGs by the web-page was constructed in the database platform so that it could be utilized by users.

5. Database construction and management based on results of bioinformatics researches for watermelon.

Annotation of 97103 genome information, which was used as a model of whole genome sequencing project of watermelon, was performed using the latest information. Based on 97103 genome information, we produced variants calling data for SNPs within 20 lines of watermelon and expression data for usefult traits such as rind, flesh color, male sterility. All the reproduced data is opened in the database platform following the methods used for case of breeding database of chinese cabbage. This system is possible to visualize data involved in watermelon breeding and search gene's information through the current genome browser and web-BLAST.

목 차

CONTENTS

제 1 장 연구개발과제의 개요

제1절 연구개발의 목적

1. 최종 목적

● 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스(DB)의 구축 및 운영
- 유전체 정보 통합 DB 구축으로 종자 개발에 있어 분자육종의 효율화를 지원함.
- 5대 채소작물(배추, 무, 수박, 고추, 파프리카)로부터 생산된 육종 특화 생물정보를 통합적으로 제공함으로써 각 채소작물 별 육종 활동에서의 시너지 효과를 유도.
- 종자 개발에 있어 목표 형질 개체선발 및 계통선발, 목표 형질 접근 방법에 대한 효율적 체계의 구성.

2. 연차별 연구 목적

가. **1차 년도의 연구 목적**
- 2011년 공개된 배추 표준 유전체, 유전자, 분자마커를 수집하여 DB를 구축할 수 있도록 재가공.
- 기존의 연구를 통해 확보한 비교 유전체 정보 수집하여 DB 구축을 위한 재가공.
- 배추의 유용 형질을 연구할 수 있는 육종 소재 및 형질 조사 내용 수집.
- 유용 형질 관련 조사를 통한 관련 유전자 및 분자마커 정보 수집.
- 수집된 형질 관련 유전자 및 분자마커 정보의 DB화.
- 교배친 및 우수자원의 re-sequencing 분석을 통해 배추 유전체상에서 대량의 Single Nucleotide Polymorphism(SNP) 발굴.
- 배추 분자육종 활성화를 위해 특화된 DB를 구축하기 위한 구성요소, 제공내역, 사용자 편의성을 고려한 배추 육종 특화 생물정보 기반의 DB 디자인.
- 배추 표준 유전체의 지놈 브라우저(Genome Browser) 구현.
- 유전체 정보와 유용 형질 관련 유전자 및 분자마커 정보 연동.
- 배추 유용 형질 관련 유전자 정보 검색 및 유전자 서열 추출 시스템 구현.
- 유전자 서열 유사도 검사를 위한 BLAST(Basic Local Alignment Search) 시스템을 DB에 도입.
- DB 및 시스템 구축을 위한 제반 환경 구현.

나. **2차 년도의 연구 목적**

- 배추의 유용 형질을 연구할 수 있는 육종 소재 및 형질 조사 내용 수집.
- 유용 형질 관련 조사를 통한 관련 유전자 및 분자마커 정보 수집.
- 배추의 조직별 DEG(Differentially Expressed Gene) 확인을 위한 전사체 (transcriptome) 데이터 생산과 데이터 분석.
- Genome-wide SNP를 이용한 여교잡 선발용 분자마커 (Marker Assisted Back-crossing) DB의 구축.
- DB 및 시스템 구축을 위한 제반환경 구현.
- 배추 Recombinant Inbred Line(RIL)의 re-sequencing 데이터 생산과 대량 SNP 발굴.
- 배추 유용 유전자원의 표현형 관련 유전자에 대한 발현량 정보 생산.
- 배추 유전체의 유용 형질 관련 유전자 구조 정보의 도식화.
- 배추의 조직 및 처리에 대한 전사체 발현 정보를 이용한 형질 관련 유전자 발굴.
- 유전체 정보를 이용한 형질 관련 배추 전사체의 발현량 정보 도식화.
- 배추 RIL의 genetic map DB 구축.
- 배추 RIL 집단을 이용한 배추 유전체 상의 변이 위치 정보를 DB화.
- 목표형질 선발용(Marker Assisted Selection) 분자마커 DB 구축.
- 집단내의 연관 불균형(Linkage Disequilibrium: LD)을 이용한 배추 RIL의 유전체 재조합 정보 분석 및 Haplotype 정보 도식화.
- 형질 관련 유전자의 대사/신호전달회로 정보 DB 연동.

다. **3차 년도의 연구 목적**

- 배추의 유용 형질을 연구할 수 있는 육종 소재 및 형질 조사 내용 수집.
- 유용 형질 관련 조사를 통한 관련 유전자 및 분자마커 정보 수집.
- 배추의 조직별 DEG(Differentially Expressed Gene) 확인을 위한 전사체 (transcriptome) 데이터 생산과 데이터 분석.
- Genome-wide SNP를 이용한 여교잡 선발용 분자마커 (Marker Assisted Back-crossing) DB의 구축.
- DB 및 시스템 구축을 위한 제반 환경 구현.
- 수집된 형질 관련 유전자 및 분자마커 정보를 활용한 DB 구축.
- 대량의 SNP 정보와 표현형 데이터를 이용하여 배추의 표현형 관련 유전적 영역 탐색을 위한 GWAS(Genome Wide Association Study)의 기반 시스템 구축.
- LD(Linkage Disequilibrium)을 이용한 배추 RIL의 유전체 재조합 정보 분석 및 Haplotype 정보 도식화 DB 구축.

라. **4차 년도의 연구 목적**

_ 수박 표준 유전체, 비교 유전체 정보 수집 및 DB 구축을 위한 정보 재가공.

- 수박의 유용 형질 정보 조사.
- 문헌상의 유용 형질 관련 유전자 및 분자마커 정보 수집 및 DB화를 위한 정보 재가공.
- 수박 표준 유전체를 기반으로 수박 유용 계통의 re-sequencing 데이터 생산 혹은 수집.
- 수박의 유용 형질 관련 전사체 데이터를 통해 수박 전사체의 발현량 정보 생산.
- 수박 집단내의 계통별 re-sequencing 정보를 이용한 대량의 SNP 발굴.
- 수박 유전체의 유용 형질 관련 유전자 구조 정보를 도식화.
- 배추를 기준으로 디자인된 DB 플랫폼에 재생산한 수박의 생물정보를 입력하여 수박의 육종 특화 DB 구축.
- 수박 지놈 브라우징 시스템 구축.
- 수박 유용 형질 관련 유전자 정보 검색 및 유전자 서열 추출 시스템 구현.
- DB 및 시스템 구축을 위한 제반환경 구현.

제2절 연구개발의 배경 및 필요성

1. 연구개발의 종합적 배경 및 필요성

가. 국내 농업의 활성화를 위한 5대 채소작물 종자 개발의 필요성

- 배추, 무, 고추, 파프리카 그리고 수박은 벼와 함께 우리나라뿐 만이 아니라 전 세계적으로 거대한 시장을 형성하고 있는 경제 작물.
- 우량 종자의 개발과 해외 수출 활성화를 통한 이윤의 창출은 국내 종자 산업에 대한 투자를 증진하여 농업의 선진화 및 농가 소득 증대와 우리 농업의 영속화를 도모할 수 있는 핵심 방안.
- 국내 종자 생산량은 급속한 감소 추세를 보이는 반면 수입 종자의 국내 종자 시장의 점유율은 갈수록 증가하고 있음.

나. 식물 게놈 프로젝트에 의한 경제 작물의 유전체 정보 접근의 실현.

- 현재 세계 각국의 연구를 통해 배추, 무, 고추, 수박등에 대한 유전체 염기서열 정보가 공개되어 연구 및 육종 목적으로 이용이 가능한 상황(표 1).
- 작물의 표준 유전체를 활용하여 해당 작물의 유전자원을 대상으로 계통별 re-sequencing 데이터의 생산 가능.
- Re-sequencing 데이터를 가공하여 얻은 계통별 변이 정보를 생산하여 이를 신규 분자마커 개발 및 이를 활용한 신품종 육성에 활용할 수 있음.

표 1. 5대 작물 유전체 공개 현황

작물	종명	유전체 크기	공개일	수행기관
배추	*Brassica rapa*	529 Mb	Nature Genetics; 2011.09.	MBGP
무	*Raphanus sativus*	573 Mb	DNA research; 2014.05.	Kazusa DNA Research Institute
수박	*Citrullus lanatus*	425 Mb (draft)	Nature Genetics; 2012.12.	IWGI
고추/파프리카	*Capsicum annuum*	2700 Mb	Nature Genetics; 2014.01.	서울대

다. 유용 형질 관련 분자마커의 개발을 통한 세계 시장에서의 생존 활로 모색.

- 현재까지 형질관련 분자마커 개발은 단일 유전자에 의해 조절되는 형질들을 중심으로 이에 대한 문헌정보를 활용하여 수행됨.
- 여교잡 선발(Marker-assisted backcrossing: MAB)을 목적으로 교배조합별

로 이용 가능한 genome-wide SNP 마커의 개발은 육종연한의 단축으로 시장의 수요변화에 대처할 수 있음.

- 유용 유전자의 계통 내 집적을 위한 형질 연관 마커의 개발 및 교차율을 이용한 후손세대 예측과 같은 육종에 실질적인 정보를 생산하고 이를 통합적으로 제공할 필요가 있음.

라. 분자육종 지원을 위한 채소작물의 육종 특화 유전체 정보 통합 DB의 구축 필요성.

- 국내 육종 환경에서는 육종포장에서 수집되는 작물의 정보는 육종회사 및 개인 육종가들이 개별적으로 작성하고 각자 폐쇄적으로 활용하여 육종 정보의 상호교환을 통한 시너지 효과를 도출하고 있지 못하는 실정.
- 유전체 분석기술(염기서열 해독 및 분석)은 비약적으로 발전하고 있으나 작물의 육종목적으로 확보한 유전자원으로부터의 자료 생산과 활용은 미비하여 개선이 필요.
- 배추의 경우, 표준 유전체 염기서열이 해독되어 물리지도로 활용되고 있으며 반수체(Doulbe haploid: DH), 여교잡(Back-Crossing: BC), RIL의 mapping 집단에서 연관지도의 작성과 같은 작물의 유전 연구에 대한 방대한 자료들이 대학과 육종회사를 중심으로 육성되었으나, 아직 이를 활용한 연구단계로의 진입이 미비함.

제 2 장 국내 외 기술개발 현황

제1절 국내외의 기술 연구 현황

1. 배추의 육종 지원을 위한 배추의 표준 유전체 정보, 육종 관련 정보의 수집 및 생산

가. 배추 표준 유전체, 유전자, 분자마커의 데이터베이스 구축을 위한 재가공

(1) 배추 유전체 정보의 수집

1. 본 연구관련 국내외 기술수준 비교

개발기술명	관련기술 최고보유국	현재 우리나라 기술수준	기술개발 목표수준	비고
형질 정보 DB	미국	40	70	http://www.phenome-networks.com/
식물 비교 유전체 DB	미국	70	90	http://www.plantgdb.org/
유전체 재조합 분석소프트웨어	미국	20	70	http://en.wikipedia.org/wiki/Haploview
육종지원을 위한 소프트웨어	미국	20	70	http://www.teamcssi.com/

1) 개발기술명은 본 연구과제 최종 연구개발 목표기술을 의미
2) 현재 기술수준은 선진국 100% 대비 우리나라 및 신청한 연구팀의 기술수준 표시
3) 기술개발 목표수준은 당해과제 완료 후 선진국 100% 대비 목표수준 제시
4) 부가설명이 필요한 경우 비고란에 작성

2. 관련 특허분석

가. 특허분석의 범위

채소 종자의 육종 활성화를 위한 데이터베이스의 구축에 응용 가능한 특허 정보를 검색하여 이를 국내외의 출처로 분류 후 상호 비교하였다. 특허 정보의 검색을 위하여 특허정보원의 데이터베이스(www.kipris.or.kr)에서 최근 5년간의 정보를 특허의 제목 및 초록을 중심으로 검색하였다.

나. 특허분석에 따른 본 연구과제와의 관련성

개발기술명		육종을 위한 분자 마커 시스템	육종을 위한 배추 형질 정보
Keyword		breeding molecular marker system	breeding brassica trait
검색건수		31,471	9,209
핵심특허 및 관련성	특허명	기능정보 융합과 MAS 활용을 위한 통합 유전마커시스템 SSR-FMM과 이를 이용한 육종방법(Functional data integrated genetic marker system, SSR-FMM, and improved genetic research and Marker Assisted Selection with SSR-FMM)	후향 후대 맵핑 (Reverse Progeny Mapping)
	보유국	대한민국	네덜란드
	등록년도	2008	2011
	관련성(%)	50	30
	유사점	서열 정보를 이용한 분자마커의 발굴.	육종을 위한 적용 기술
	차이점	선발되는 마커의 종류가 SSR에만 해당하는 것을 SNP, SSR, In/Del등 모든 polymorphism에 적용 가능함. EST를 이용한 RNA정보만을 이용하는 것에서 DNA, RNA 정보를 모두 활용 가능함. 마커 발굴에만 그치는 것이 아니라 발굴된 마커를 이용한 실질적인 육종 지원에 적용하고자 함.	신품종 개발을 위하여 후향 후대 맵핑 방법론을 사용하나 본 연구에서 제안하는 방법은 채소 작물의 어린 잎에서 바로 유전체적 정보로 형질을 특징을 파악할 수 있도록 구성하는 것에 주안점을 두고 있음.

1) 개발기술명은 본 연구과제 최종 연구개발 목표기술을 의미
2) keyword는 검색어를 의미하며, 검색건수는 keyword에 의한 총 검색건수를, 유효특허건수는 검색한 특허 중 핵심(세부)개발기술과 관련성이 있는 특허를 의미
3) 핵심특허는 개발기술과의 관련성이 높고 인용도가 높은 특허를 기준으로 분석

3. 관련 논문분석

가. 논문분석의 범위

채소 종자의 육종 활성화를 위한 데이터베이스의 구축에 응용 가능한 논문 정보를 검색하여 이를 국내외의 출처로 분류 후 상호 비교하였다. 특허 정보의 검색을 위하여 NCBI의 pubmed 데이터베이스(https://www.ncbi.nlm.nih.gov/pubmed)에서 최근 5년간의 정보를 특허의 제목 및 초록을 중심으로 검색하였다.

나. 논문분석에 따른 본 연구과제와의 관련성

개발기술명		육종을 위한 분자 마커 시스템	표현형-유전형 관계성 규명 시스템
Keyword		genome assisted breeding	phenotype genotype database
검색건수		628	979
핵심논문 및 관련성	논문명	Development of genomics-based genotyping platforms and their applications in rice breeding	SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments
	학술지명	Curr. Opin. Plant Biol.	NAR
	저자	Chen et al.	Jessen *et al.*
	게재년도	2013	2013
	관련성(%)	60	40
	유사점	- NGS 정보를 이용한 마커 선발 시스템 구성. - 선발된 마커를 이용하여 MAS, MAB의 가능성 제시.	- 표현형과 유전형의 관계성 규명
	차이점	- Rice에 한하여 정보를 제공함. - 형질과 마커의 연관성을 제공하지 않음.	- 단백질 정보를 이용하여 관계성을 규명하나 본 연구에서는 유전체와 유전자를 이용하여 관계성을 규명함.

1) 개발기술명은 본 연구과제 최종 연구개발 목표기술을 의미
2) keyword는 검색어를 의미하며, 검색건수는 keyword에 의한 총검색건수를, 유효논문 건수는 검색한 논문 중 핵심(세부)개발기술과 관련성이 있는 논문을 의미
3) 핵심논문은 개발기술과의 관련성이 높고 인용도가 높은 논문을 기준으로 분석

4. 제품 및 시장 분석

가. 생산 및 시장현황 (배추의 사례)

(1) 배추 종자의 해외 채종 비율과 종자 자급률 (2010년 기준)

배추는 국내에서 재배되고 있는 다른 작물에 비해 월등하게 낮은 해외 채종률을 보이고 있는 채소이다. 국내 채종에 대한 환경조건은 해외의 경우와 견주어 보았을 때 불리한 점이 없는 것이 가장 큰 이유이다. 배추의 경우 다른 작물에 비해 채종시기가 장마철 이전이므로 채종과정에서 습해에 의한 수발아등 채종 종자의 수량을 저하시키는 요인이 적은 것이 특징이다. 그 결과 배추의 경우 국내 시장에서의 종자 자급률이 거의 100%에 달한다.

(2) 국내 종자산업 현황

국내 1위의 종자회사는 농우바이오로 해외 4개 현지법인과 연구소를 보유하고 있으며, 2011년 매출액 560억 원으로 국내 종자시장의 25%를 점유하고 있다. 또한 농우바이오는 중국, 인도네시아, 미국, 인도에 현지법인을 설립하여 해외시장에 진출, 총 매출액 중 수출 비중이 30%를 차지하고 있으다. 국내 시장의 정체에도 불구하고 농우바이오는 고기능성 신 종자 개발 및 시장 확대를 모색하고 있고 내수시장에서의 영향력을 유지하고 있다.

그 외의 국내 종자회사의 경우, 동부팜한농(주)는 Monsanto Korea의 채소 종자 부분을 인수하고 Monsanto의 연구인력, 소재를 활용하여 채소종자시장에서 두각을 나타내고 있다. 아시아종묘(주)는 생명공학육종연구소, 남부육종연구소 등의 국내 연구소와 인도 뱅갈로 육종연구소를 중심으로 내수 및 수출용 품종을 개발하고 있으며 매년 매출액 대비 20%이상을 종자 개발 연구에 투자하고 있다.

현재 국내의 종자산업은 약 5000억원 규모 수준의 시장으로 성장하였다. 동부, LG, 노루 그리고 농협과 같은 국내 거대 기업이 종자산업에 대한 투자를 점진적으로 늘리고 있고 국내의 채소 종자기업의 수는 2011년에 185개로 증가하였으나 여전히 중소 및 개인 육종가와 같은 영세한 규모의 사업체가 대다수이다.

(3) 국외 종자생산 및 시장 현황

전 세계 채소종자 시장규모는 58억 달러(약 6조 3800억) 수준이며 유럽 및 아시아에서 생산된 종자가 전체 시장의 70% 점유하고 있다. 채소 종자산업은 종자산업 중에서 가장 복잡하고 세분화된 분야로, 전반적인 시장규모가 증가하는 추세이다. 또한 인구의 증가와 함께 개인당 소득의 증가 및 건강에 대한 관심과 수요가 증대됨에 따라 채소 소비도 지속적으로 증가 추세이며, 이에 따라 종자 시장규모도 증가하는 것으로 나타났다.

대륙별 종자 시장 현황

	아시아	유럽	북미	중동 /아프리카	중남미	합계
면적	17.1	3.9	1.9	2.8	0.9	26.6
시장규모 (채소종자)	21.1	18.6	9.6	5.9	2.8	58.0

(단위 : 백만ha, 억 불)

전 세계의 상업용 채소종자 시장은 2011년 기준으로 약 5조 5천억 원으로 추정되며, 2020년에는 9조 6천억 원까지 성장할 전망이다(연평균 성장률 7%). 이러한 상업용 채소종자 시장 성장의 주요 원인은 종자가격의 상승효과(54%)이며, 이밖에도 재배면적 증가효과(20%), 교배종으로의 전환효과(16%)에 의해서 시장이 성장하고 있는 것으로 나타났다.

아시아 지역의 경우, 2011년 총 21.1억 달러 규모의 신흥시장으로 평가되며, 향후 가장 큰 시장으로 성장할 것으로 예상된다. 특히 중국과 인도 시장은 기존 재래종에서 교배종으로 급격하게 전환되는 추세로서 시장규모도 급증할 것으로 예측된다.

반면, 유럽과 북미의 경우, 안정기에 접어든 성숙기 시장으로, 각각 18.6억 불, 9.6억 불의 시장을 형성하고 있다. 또한 유럽과 북미에서는 대부분의 상용되고 있는 채소종자에 대하여 교배종에 대한 개발이 완료되어 종자 가격이 상당히 높은 고단가 시장이며, 글로벌 업체의 점유율이 높다. 향후, 유럽과 북미 지역에서의 채소에 대한 수요가 개발도상국으로부터의 수입으로 충족되는 비중이 증가할 것으로 전망되어 해당지역에서의 지역 원산의 종자 소요량은 줄어들 것으로 보인다.

중동과 아프리카 지역의 전체 시장규모는 5.9억 달러로 추정되는 신흥시장이며, 중남미 지역도 2.8억 달러의 시장 규모를 갖고 있는 것으로 평가되고 있다. 중동시장의 경우 점차 고부가 가치의 종자를 생산할 수 있는 전환기를 맞고 있으며 아프리카의 경우, 현재는 저가 시장이나 향후 성장 가능성이 높은 상황이다. 그 외의 중남미 시장의 경우, 유럽과 미국 회사의 종자가 지역 시장을 강하게 지배하고 있는 것으로 나타났다.

나. 개발기술의 산업화 방향 및 기대효과

(1) 산업화 방향(제품의 특성, 대상 등)

세계의 각 주요시장별 및 세부기술별로 종자 생산에 관계된 특허 출원의 증가율을 분석한 결과, 작물 육종을 목표로 한 분자마커의 출원 증가율이 두드러지게 나타났다. 분자마커의 전체 출원 추이는 한국에서 출원한 추이와 동일한 양상을 나타내는 것으로 보아 국제적으로 분자 마커 분야에서 한국이 기술을 주도하는 것으로 나타났다.

작물의 형질 관련 유용 유전자에 대한 분야의 경우, 2000년대 중반부터 최근까지 한국의 특허 등록에 대한 추이가 관련 특허 시장에 큰 영향을 미치는 것으로 분석되었다.

전통육종 분야는 중국의 특허 출원이 강세인 것으로 나타났으며 분자 육종의 경우 한국, 미국 및 중국을 중심으로 해당 기술의 개발이 이루어져 세계 시장을 선도하는 것으로 나타났다.

종자 처리 분야의 경우 분석 초기부터 2000년대 중반까지는 미국이 기술을 주도하는 것으로 나타났으나, 이후 최근까지 중국의 출원 급증에 따른 여파가 전체 기술에 적극적으로 반영되어 개발이 이루어지고 있다.

5. 3P(Patent, Paper, Product)분석을 통한 연구추진계획

가. 분석결과 향후 연구계획

(1) 특허분석 측면

기존 특허는 EST 데이터를 이용한 전사체상의 SSR 마커 발굴 분야에 치중되어 있으므로, 본 연구과제에서는 NGS 데이터를 이용하여 전사체뿐만 아니라 유전체 전체에서 SNP, SSR, In/Del과 같은 사용가능한 모든 마커를 이용하여 육종에 지원할 수 있도록 연구를 추진하여 분자마커 선발 시스템 특허 등을 국내 및 국외에 출원할 계획이다.

(2) 논문분석 측면

기존 논문은 식물체를 이용한 마커의 발굴 및 mapping 분야 자체에 치중되어 있으므로, 본 연구과제에서는 5대 채소작물의 육종 프로그램에 효율적으로 활용될 수 있는 마커를 선별하기 위한 시스템 고안 및 마커의 확인 방향으로 연구를 추진하여 효율적인 분자마커 선발에 대한 논문 등을 학술지에 게재할 계획을 수립하였다.

(3) 제품 및 시장분석 측면

국내 및 국외시장에 대한 동향을 분석한 결과, 육종지원시스템과 같은 시스템의 제품화는 아직 존재하지 않으나, 이러한 시스템을 이용하여 도출하게 될 유용 신품종 채소에 대한 산업화가 이루어 질 수 있을 것으로 기대된다. 본 연구과제에서는 신품종 개발을 가속화하는 체계화된 육종 지원 시스템에 대한 통합적인 모델을 제시하여 최종적으로는 이를 효율적인 채소 육종을 위한 정보화 시스템으로서 판매할 계획이다.

제 3 장 연구개발수행 내용 및 결과

제1절 배추의 육종 특화 데이터베이스(DB) 구축을 위한 육종 기반 정보의 생산 및 수집

1. 배추의 육종 지원을 위한 배추의 표준 유전체 정보, 육종 관련 정보의 수집 및 생산

가. 배추 표준 유전체, 유전자, 분자마커의 데이터베이스 구축을 위한 재가공

(1) 배추 유전체 정보의 수집

배추는 김치의 주재료로서 국내에서 1조원 규모의 높은 시장성을 갖는 작물이다. 또한 세계적으로 김치의 수요가 늘어나면서 배추는 산업 전략작물로 자리를 잡은 상태이다. 우리 배추가 식품 및 세계 종자시장에서 지금보다 더 높은 국제경쟁력을 갖기 위해서는 다양하고 우수한 형질을 가진 품종들이 지속적으로 개발되어야 한다. 이를 위해서는 국내 종자산업에서 관행적으로 이루어져 왔던 전통육종방식에 분자마커를 이용한 분자육종 기술체계 확립과 인프라 구축이 반드시 요구되고 있는 실정이다. 더욱이 지난 2011년에는 한국이 하나의 주축이 되어 수행한 국제 배추 유전체 염기서열 분석 프로젝트를 통해 배추의 표준 유전체가 공개되었기 때문에(MPGSP, 2011) 배추 분자육종 체계 및 인프라 구축이 국내에서 형성되기 매우 좋은 여건을 가진 상황이다. 배추의 분자육종 활성화를 위한 정보 수집이 가능한 출처는 표 1과 같으며 분자마커로서 활용 가능한 배추 유전체 상의 SNP 분포는 표 2와 같이 알려져 있다.

표 1. 배추 표준 유전체 및 유전자 정보의 열람 가능 데이터베이스

데이터베이스 명	데이터 종류	데이터베이스 주소
Brassica database	Genome data, Predicted gene, Raw reads, etc.	http://brassicadb.org/brad/
The Multinational *Brassica* Genome Project	Data links, etc.	http://www.brassica.info/
Brassica EST Database	SSR database, SNP database, Unigenes, etc.	http://brassest.cnu.ac.kr/
Brassica Genome Gateway	Genome sequence Unigene set, etc.	http://brassica.nbi.ac.uk/
BrTED	EST data, Unigene sequence, Tissue specific EST set, etc.	http://brted.rna.kr/

표 2. 수집된 유전체 염색체 내의 SNPs 분포

Chromosome	Homotype SNPs	Heterotype SNPs	Total SNPs
A01	4,526	3,231	7,757
A02	735	446	1,181
A03	6,240	3,634	9,874
A04	2,238	1,348	3,586
A05	3,011	2,292	5,303
A06	4,570	3,663	8,233
A07	4,110	2,633	6,743
A08	2,071	1,253	3,324
A09	4,201	2,929	7,130
A10	3,683	2,531	6,214
a00	11,243	18,895	30,138
a01	3,360	3,045	6,405
a02	5,602	3,879	9,481
a03	263	96	359
a04	2,088	1,809	3,897
a05	2,680	1,964	4,644
a06	1,687	1,078	2,765
a07	1,051	776	1,827
a08	3,228	2,734	5,962
a09	4,287	3,305	7,592
a10	213	157	370
Total	71,087	61,698	132,785

나. 배추 유용 형질의 선정과 관련 정보의 생산 및 수집

(1). 배추의 유용 형질에 대한 표현형 정보의 생산

1차 년도에는 배추의 육종 프로그램에 포함될 것으로 기대되는 유용형질로서 Carotenoid, Flavonol, Vitamin C, Glucosinolate, Reducing Sugar, Mineral을 선정하여 보유한 계통을 사용하여 이에 대한 표현형을 평가하고 계통간의 성분 함량을 비교 분석하여 집단에서의 유용형질의 양에 대한 유전적 관계를 찾고자 하였다. 조사 결과 유용형질 함량이 높은 우수한 자원을 선발하여 육종 소재로의 가능성을 분석하였다.

(가) Carotenoid

표 3. 수집된 육종 소재의 Carotenoid 분석 결과

(μg/g D.W.)

		Access No.	Average
β-carotene	Chinese cabbage	11639	*33.34*
		11640	*19.92*
		11707	*41.88*
Lutein	Chinese cabbage	11639	*25.53*
		11640	*14.56*
		11707	*34.75*

(나) Flavonol

표 4. 수집된 육종 소재의 Flavonol 분석 결과

(mg/g D.W.)

		Access No.	Average
Quercetin	Chinese cabbage	11639	*0.19*
		11640	*0.18*
		11643	*0.15*
Kaempferol	Chinese cabbage	11639	*0.07*
		11640	*0.07*
		11643	*0.06*
Total	Chinese cabbage	11639	*0.26*
		11640	*0.25*
		11643	*0.20*

(다) Vitamin C

표 5. 수집된 육종 소재의 Vitamin C 분석 결과

(mg/100g D.W.)

		Access No.	Average
Vitamin C	Chinese cabbage	11636	*13491.86*
		11637	*12558.48*
		11644	*61251.05*

(라) Reducing sugar

표 6. 수집된 육종 소재의 Reducing sugar 분석 결과

(mg/g D.W.)

		Access No.	Average
Reducing sugar	Chinese cabbage	11695	*242.1*
		11696	*336.6*
		11697	*281.7*

(마) Glucosinolate

표 7. 수집된 육종 소재의 Glucosinolate 분석 결과

(μmol/g D.W.)

		Access No.	Average
Methoxy glucobrassicin	Chinese cabbage	11495	*0.67*
		11496	*1.32*
		11497	*0.76*
Neoglucobrassicin	Chinese cabbage	11491	*1.63*
		11492	*1.17*
		11493	*0.44*
Progoitrin	Chinese cabbage	11488	*2.14*
		11490	*1.15*
		11491	*1.63*
Glucoraphanin	Chinese cabbage	11495	*0.67*
		11496	*1.32*
		11497	*0.76*
Gluconasturtiin	Chinese cabbage	11499	*2.13*
		11500	*1.73*
		11501	*1.69*
Gluconapoleiferin	Chinese cabbage	11503	*2.73*
		11504	*0.45*
		11505	*1.25*

		Access No.	Average
Gluconapin	Chinese cabbage	11710	*1.41*
		11711	*0.61*
		11712	*1.17*
Glucocohlearin	Chinese cabbage	11685	*0.11*
		11686	*0.45*
		11687	*0.06*
Glucobrassicin	Chinese cabbage	11712	*1.17*
		11713	*0.79*
		11714	*0.45*
Glucobrassicanapin	Chinese cabbage	11488	*2.14*
		11490	*1.15*
		11491	*1.63*
Glucoalyssin	Chinese cabbage	11483	*1.21*
		11484	*0.82*
		11485	*0.85*

(바) Mineral

표 8. 수집된 육종 소재의 Mineral 분석 결과

(ppm D.W.)

		Access No.	Average
Ca	Chinese cabbage	11602	*14660*
		11603	*18370*
		11604	*13170*
Mg	Chinese cabbage	11605	*29680*
		11606	*15490*
		11608	*25410*
Mn	Chinese cabbage	11609	*20.4*
		11610	*8.6*
		11635	*14.4*
Zn	Chinese cabbage	11672	*56.45*
		11673	*55.15*
		11677	*123.4*

		Access No.	Average
Fe	Chinese cabbage	11673	*73.15*
		11677	*67.1*
		11678	*70.2*

(2) 배추 유용 형질 관련 분자마커 정보의 수집

배추의 유용 형질에 대한 문헌조사를 통해 해당 형질에 대한 유전자 정보와 분자마커 정보를 수집하였다. 1차 년도에는 내병성 형질에 대해서는 뿌리혹병, TuMV, 노균병, 무름병에 대한 정보를 수집하였으며 (Hatakeyama et al, 2013; Jin et al, 2014; Farinho et al, 2004) 그 외의 형질로는 웅성불임, 자가불화합성(Chase et al, 2007; Hiscock and Mclnnis 2003) 그리고 Glucosinolate에 대한 정보를 수집하였다.

2차 년도에는 1차 년도에 수집된 형질들과 개화기, 초장, 종피색, 잎털, 환경저항성에 대해 심화된 문헌조사를 통해 관련 논문에서 형질에 관련된 candidate gene의 정보와 표현형 검정에 사용된 분자마커의 정보를 얻었다(Zhao et al, 2010; Li et al, 2013; Rahman et al, 2007; Zhang et al, 2009). 그로부터, 배추의 표현형 관련 등록된 특허 및 정보 검색의 대상에 포함시켜 상용되고 있는 분자마커의 정보도 얻을 수 있었다. 그 특성과 관련 정보는 표 9와 같다.

표 9. 수집된 저항성 유전자 및 분자마커 정보

특성 분류	특성	대표적 분자마커 또는 후보 유전자명	분자 표지 형태
내병성	뿌리혹병	*CRa*	RFLP, RAPD, SCAR, STS, RAPD, SCAR
		GC2360/GC1680	STS
		Crr1	SSR, SNP, InDel, CAPS
		Crr2	SSR, SNP
		Crr4	SSR
		Crr3	RAPD
		CRk	RFLP
		CRc	RAPD/STS
		CRb	AFLP/SCAR
	TuMV	*TuMV-R* *TuRB07*	SSR
	노균병	*BraDM*	RAPD/Isozyme
		BrRHP1	SCAR/InDel
	무름병	*PinII/aii*	-
개화	유전자적 웅성불임	GMS	AFLP, SCAR, SSR
	웅성불임	*BrRF1*	-

특성 분류	특성	대표적 분자마커 또는 후보 유전자명	분자 표지 형태
	자가불화합성	*SLG,SRK,SCR/SP11*	STS
	개화기	*FLC1*	SNP
형태적 형질	종피색	*TTG 1*	SCAR
	초장	*CKX5* *GRF3*	SNP
	털	*GL1* *EGL3* *TTG1* *TRY*	SSR
환경 저항성	저온저항성	*BrCSR* *OsAOX1a*	-
	염 저항성	*BrSSR*	SSR

(가) 순무모자이크 바이러스 (TuMV)

TuMV는 배추과 식물에 수확량 손실 및 품질저하 등의 많은 피해를 주는 대표적인 바이러스 병으로 알려져 있다. TuMV는 진딧물에 의해 비영속적으로 전염되므로 재배기간 동안의 완벽한 방제가 어렵다. 따라서 저항성 품종을 육성하는 것이 가장 효과적인 방제법으로 평가되고 있다. 보고된 특허(대한민국특허. 등록번호 10-1311249)와 논문(Jin et al. 2014)에서 공개된 TuMV 내병성 판별용 마커 정보를 수집하였다. 수집된 SSR 프라이머를 이용하여 배추 순무모자이크 바이러스(TuMV-C4) 내병성 인자와 연관된 SSR 마커에 의한 밴드를 검출함으로써, 바이러스의 접종 없이 내병성 개체와 이병성 개체를 유전자 수준에서 간편하고 정확하게 구분할 수 있으며, 육종 세대 단축을 통해서 품종 개발의 효율을 증진할 수 있다.

표 10. 수집된 배추 TuMV 내병성 판별용 SSR 프라이머 정보

프라이머명	염기서열 (5'-3')	증폭산물 (bp)
KS10960(F)	TCTTCACGCAATGGCTTT	266
KS10960(R)	TCCCCATTAATGACACGC	
H132A24-s1(F)	CTCAAATAGCAACGACGCAT	279
H132A24-s1(R)	CAGCAGTGGGATATCGGG	

본 프라이머 세트는 공우성 마커로서 내병성 집단의 경우 저항성 증폭 산물(R 밴드)과 이병성 증폭산물(r 밴드)을 전부 나타낸다. 반면에, 이병성 집단의 경우 이병성 증폭산물(r 밴드)만을 나타내는 것을 통해 주어진 시료의 저항성 유무를 판별할 수 있다.

표 11. TuMV 내병성 판별 마커 평가에 사용된 배추 소재 정보

순번	계통명	소재 정보
1	VC-1	TuMV 내병성 DH 계통
2	SR-5	TuMV 이병성 DH 계통
3	VCS-3	F1 계통(VC-1 x SR-5)

(나) 노균병

배추 속 식물에서 노균병 저항성의 유전 양상에 대한 광범위한 연구가 진행되어 오고 있다. 특히, 유채(*Brassica napus*)와 양배추(*Brassica oleracea*)에서 다양한 출처의 저항성이 보고되었다. 애기장대에서는 노균병 저항성을 나타내는 20개의 유전자좌 *RPP(Recognition Peronospora parasitica)*가 확인되었고, 이 중 6개의 *RPP* 유전자 또는 유전자 클러스터가 클로닝되었다. 배추의 노균병 저항성과 관련되어 유묘 단계에서 노균병 저항성을 조절하는 주요 양적 형질 유전자좌가 확인되었으며, *BraDM*이 A8 연관 그룹(Linkage group) 상에 위치하는 것으로 밝혀졌다. 성체식물 단계에서 노균병 저항성을 조절하는 신규의 유전자좌인 *BrRHP1*이 밝혀져 있으며 이와 관련된 SCAR 표지가 보고되었다.

표 12. 수집된 배추 노균병 분자표지의 프라이머 서열

프라이머명	염기서열 (5'-3')
OPA08	GTGACGTAGG
A-F1	CTGGTTTCTTCCTTGCATTGCCCGATA
B-F1	AGTTCATCGGTTTGAACCGGCTTGTTG
Co-R1	GACGCCGGCCTGTTGGTAAATCACAT
BN-F1	TCTGAGCTCCCGTCTAAGTTG
BN-R1	TGTCCAACATTCAGCAAAGC
G17-F	GCGGGTTGACCCCTAGTAAT

G17-R	TGCAAGTTGTGTCGGACAAT
M22-F	ATACCAAAAGCAACGGCAAC
M22-R	TGGGGAAGAAGGTTTGTTTG
N18-F	GAGGCAAGAACCTTCTCCAG
N18-R	TTGCTCAACATCATCGGTCT
M05-F	ACAACATTAGCAACGCACCA
M05-R	CTTTTCTATCGCGCCTGAAC
J11-F	TGTGGGAGAGATAGGGTTGG
J11-R	TTTGTCGGAGGGATCAAAAA
A03-F	AGGTTCGACCACCATGACTC
A03-R	TGGGGTGTTTACACAAAGCTC

※대한민국특허(배추 노균병에 대한 성체식물 저항성을 부여하는 유전자좌, 및 이와 연관된 분자 표지. 10-2012-0136772)

(다) 무름병

무름병은 배추, 무, 당근 및 감자 등 주요 채소류에 경제적 손실을 초래한다. 특히, 배추의 무름병은 배추의 피해를 크게 주는 3대 병 가운데 하나이다. 무름병은 세균성의 토양전염성 병으로 감염 시 치료가 불가능할 뿐 아니라 적합한 예방법도 없는 실정이다. 또한 현재 무름병에 대한 내병성 육종소재가 없기 때문에 병이 많이 발생하는 여름철에 상습적인 발병지에 재배하면서 병이 적게 발생하는 계통을 선발하는 육종법이 진행되고 있다. 무름병 저항성은 양적형질로 미동유전자에 의해 지배되기 때문에 유전자의 집적이 어려워 저항성 품종이 전무한 실정이다. 현재까지 보고된 무름병 저항성에 관련된 정보는 외래유전자 도입 또는 특정 유전자를 과발현시킨 형질전환 식물체를 생산한 것이 대부분이다. 또한 *BrWRKY12* 유전자가 식물체의 무름병에 대해 저항성을 증진시킬 수 있음이 확인되었고 이 유전자를 이용한 형질전환 식물체 및 이의 제조방법이 보고되었다. 배추에서 유래한 무름병 저항성 관련 유전자 외에 감자 유래 PinⅡ(protease inhibitor Ⅱ)의 신호 펩티드 코딩 유전자 및 바실러스 sp. GH02 유래 aii(autoinducer inactivation) 유전자 재조합 식물 발현 벡터를 이용하여 식물의 병충해 내성을 증진시키는 것이 보고되었다.

그림 1. 시스테인 프로테아제 유전자를 이용한 형질전환체 벡터의 모식도 (대한민국특허. 10-2011-0069351)

그림 2. 배추 BrWRKY12 과발현 벡터의 모식도 (대한민국특허. 10-2014-0094167)

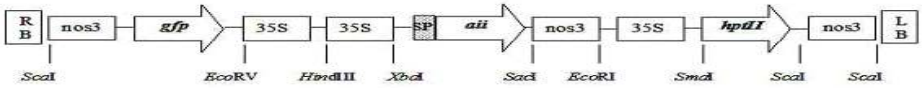


그림 3. PinII+Aii 재조합 T-DNA 영역의 모식도 (대한민국특허. 10-2011-0001312)

(라) 웅성불임

특허 정보 조사를 통해 세포질적 웅성불임(CMS)에 대하여 공개된 마커 정보를 수집하였다. 본 실험실 보유 유전자원 및 타 기관 분양소재를 활용하여 마커 검정을 통해 수집된 분자마커의 특성을 평가하였다.

표 13. 수집된 배추 CMS Type 웅성불임성 판별용 마커 정보

프라이머명	증폭산물의 크기 (bp)	구분	Reference
Brams A	1,302(CMS)/1,000(MF)	CMS와 MF 구분	대한민국특허 (공개번호10-2012-0041192)
Brams B	1,020(CMS)/963(MF)	CMS와 MF 구분	
Brams C	303	Ogura CMS 구분	
Brams D	1218	Ogura 및 SNU3 CMS 구분	
Brams E	660	Polima CMS 구분	
Brams F	666	Polima CMS 구분	
Brams H	513	Polima CMS 구분	
Brams G	825	Polima CMS 구분	
Brams K	891	SNU3 CMS 구분	

표 14. 수집된 배추 GMS Type 웅성불임성 판별용 마커 정보

프라이머명	증폭산물의 크기 (bp)	구분	Reference
syau_scr01	378(Ms)	GMS 구분	대한민국특허 (공개번호10-2010-0117170)
syau_scr04	204(Ms)/208(ms)	GMS 구분	
syau m-13	296(Ms)/300(ms)	GMS 구분	
CNU-m273	271(Ms)/273(ms)	GMS 구분	

실험에 사용된 육종 소재는 총 34개이며, 이의 웅성불임성 여부가 구분된다. 그림 1은 수집된 배추 CMS 타입 웅성불임성 판별용 마커를 평가한 결과이며 Brams D, Brams K 마커는 Ogura 및 SNU3 CMS를 구분할 수 있다. 일부 마커의 경우 CMS/GMS type을 나눌 수는 없었지만, 웅성불임성이라는 형질 자체에 대해서는 판별이 가능한 마커로 평가할 수

있었다.

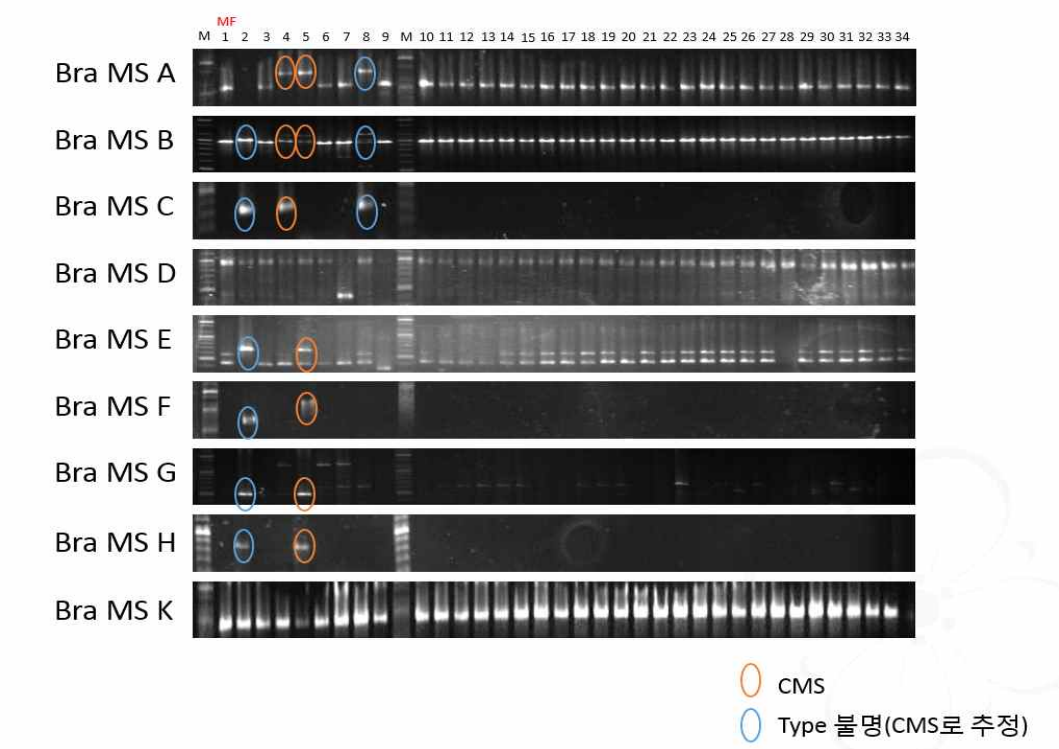


그림 4. 수집된 배추 CMS Type 웅성불임성 판별용 마커 평가

표 15. 마커 평가에 사용된 배추 소재 정보

순번	소재 등록번호	소재 정보
1	CNU_143006	Chiifu(MF)
2	CNU_11374	Akimeki(*Crr1* and *Crr2* homo, *CRb* hetero) F1
3	CNU_29008	MS AB line
4	25045	1-30, Rcb (Mst1), Ogura cytoplasm source, AB1aa
5	25046	1-31, Rcb (Mst2), Anand cytoplasm sourec, AB1aa
6	25047	1-35, Ant(7) (mst2/mst2,mst2/MST2), purple, Aaa
7	25048	1-107, anl/anl, Pan(6):Hir(1) (mst2/1:1), nonpurple
8	MS-2	농협종묘
9	13042(MS maintainer)	-
10 - 34	SYC_14XXX	실험실 수집 자원

(마) 자가 불화합성

NCBI로부터 기 보고된 S locus 관련 염기서열과 특허 정보 조사를 통해 공개된 마커 정보를 수집하였다. 배추에서 자가 불화합성과 관련이 있는 유전자로 SLG, SRK, SP11 등이 있으며, 총 72종의 관련 유전자 서열이 보고되었다. 수집된 정보는 배추 유전체 정보와 비교를 통해 육종 소재의 SI type 분석에 활용 가능하도록 가공되었다.

표 16. 수집된 배추 자가 불화합성 판별용 멀티플렉스 프라이머 정보

프라이머명	염기서열 (5'-3')	Reference
BrSRK22	GCATACCAGGGGATCAATAT	대한민국특허 (등록번호10-1413116)
BrSRK25	GTGCGATACGTACAGAGCG	
BrSRK46	CTCAAATAGCAACGACGCAT	
BrSRK54	CAGCAGTGGGATATCGGG	
BrSRK55	ATAATCTTGTCCTCCTTGAT	
BrSRK AS-1	GCAGCCAATCTGACATAAAG	

(바) 개화기

특허 정보 조사를 통해 배추 속 식물의 개화억제 유전자인 *FLC1* 관련 마커 정보를 수집하였다. 이 마커 세트는 *FLC1* 유전자의 여섯 번째 엑손과 인트론이 스플라이싱 되는 영역에 존재하는 SNP 확인이 가능하다.

표 17. *FLC1* 유전자 관련 분자마커 정보

프라이머 조합	증폭산물의 크기 (bp)	구분
1, 8	169	여섯 번째 Exon과 여섯 번째 Intron의 Splicing 지역의 SNP 판별
2, 8	169	
3, 8	169	
4, 8	169	
5, 8	169	
6, 8	169	
7, 8	169	
9, 10	206	다섯 번째 인트론 내에 있는 SNP 및 여섯 번째 Exon과 여섯 번째 Intron의 Splicing 지역의 SNP 동시 판별
11, 12	206	

표 18. *FLC1* 유전자 판별을 위한 SNP 지역 증폭용 프라이머 정보

프라이머명	염기서열 (5'-3')	Reference
1	ATGTTTTNNNTAGCCAGA	대한민국특허 배추속 작물의 개화시기 판별용 조성물 (출원번호10-2012-0061704)
2	ATGNNTNNNTAGCCAGA	
3	ATGTTTTNNNNNGCCAGA	
4	ATGTNNTNNNTAGCCAGG	
5	ATGTNTTNNNTNGCCAGG	
6	ATGTTNTNNNTAGCCNGG	
7	ATGTTTTNNNNNGCCAGG	
8	AGGCTCCCCAGATAATATGT	
9	AGGTCGCAAGCCTATCTCT	
10	CCGTAACAAAAAAAACTTTAGTTAT	
11	AGGTCGCAAGCCTATCTCC	
12	CCGTAACAAAAAAAACTTTAGTTAC	

마커의 검정능력 평가를 위해 실험실에서 수집한 유전자원들의 개화 관련 표현형을 조사한 결과에서 만기개화를 나타낸 5개의 소재와 조기개화를 나타낸 5개의 소재를 선정하여 마커 검정에 이용하였다.

표 19. 마커 평가에 사용된 배추 소재 정보

순번	소재 등록번호	소재 정보	특성
1	120012	NLDCGN_CGN06817	만기 개화
2	CNU_11387	OHCR	
3	CNU_11380	CR-GJ	
4	25084	常州烏塌茶・상주오탑차・sang ju o top cha	
5	120031	NLDCGN_CGN06790	
6	CNU_28064	caixin	조기 개화
7	CNU_28067	Z062282 / 내혼계	
8	27077	Pusa kalayani	
9	28732	Brown sarson Tora Type	
10	CNU_11418	HKC-005	

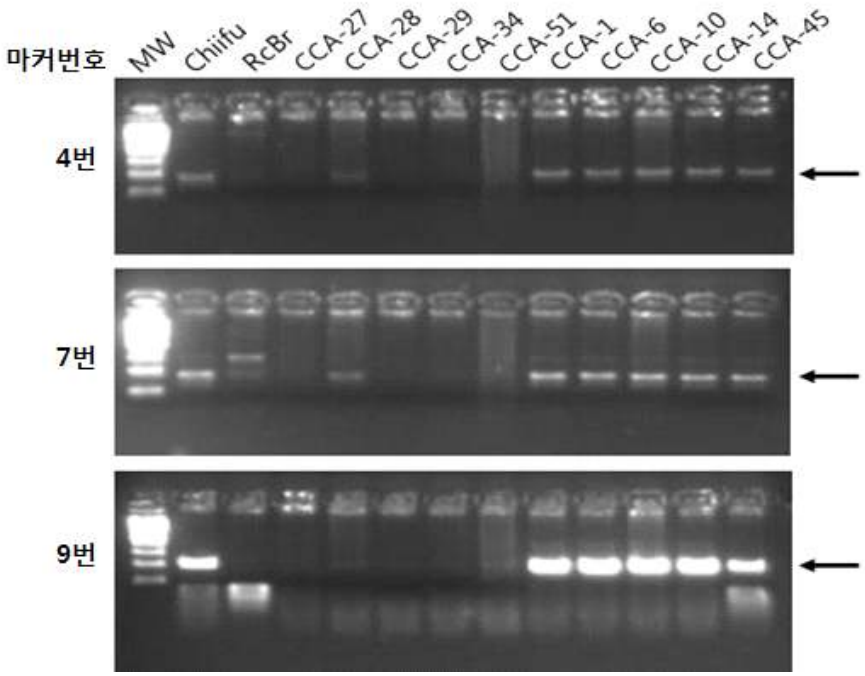


그림 5. FLC1 유전자 판별을 위한 SNP 마커 평가 결과

(사) 종피색

종피색은 짙은 갈색 종자와 노란색 종자를 맺는 품종이 존재하며, 식물 표피의 털과 종실유 함량과 연관되어 유전되는 양상을 보인다. 검은색/갈색 종자보다 노란색의 종피를 가진 종자의 oil 및 단백질 함량이 높게 나타났다. 육종 소재로 'Y177-12(A)' - 검은 종피와 'Y195-93(B)' - 노란 종피를 이용한 연구가 보고되었다.

표 20. 종피색과 연관된 SCAR 마커 정보

프라이머명	프라이머 염기서열 (5'-3')	Reference
JF39	CCGCATGTTTCACCAACC	
JF40	TGGCCTTACATAGTGGAAG	
JF5G3	ATAGAAAAGTAAAGGTACTCTCTT	
JF5G4	GGTACTCTTTTTAGTGCGA	Zhang et al. 2009. Plant
JF5G5	ACCAGTTCCTTGTTCGTTC	Molecular Biology.
JF87a	GTCCCAACCTGCGTTCTA	
JF106	CAGAGCATAAATCTCCTGC	
JF106b	CCAGAGCATAAATCTCTTATG	

전기영동을 통해 분리된 증폭밴드의 염기서열 결정과정으로 얻은 염기서열과 애기장대 데이터베이스의 비교를 통해 AT5G26680, AT3G62850, AT5G63330, AT2G19110의 총 4개 유전자 서열을 수집하였다.

(아) 초장

배추의 초장은 양적형질의 양상을 나타내고 있다. 문헌 조사와 특허 검색을 통해 이와 관련된 유전자좌로 총 7개의 QTL 지역과 SNP 정보를 수집하였다.

표 21. 초장과 관련된 총 7개의 QTL 지역

QTL	LGs	Confidence interval (cM)	Marker interval	Block	Reference
qPH1	A2	0.5-15.2	cnu_m616a-At5g12290	R	Li et al. 2012. DNA Research.
qPH2	A3	7.3-9.7	ACMP00100-ACMP00396	R	
qPH3	A3	40.8-44.4	cnu_m288a-ACMP00087	J	
qPH4	A5	67.3-86.5	ACMP00868-cnu_m029a	J	
qPH5	A7	92.6-104.0	cnu_m056a-cnu_m308a	E	
qPH6	A9	138.2-139.6	cnu_m356a-At2g20490	H	
qPH7	A10	11.9-26.0	ACMP00858-nia_m015a	A	

표 22. QTL 지역에 위치하는 plant height와 관련된 putative candidate genes과 SNP

Gene family	Gene name	Block (LGs)	At gene ID	*B. rapa* gene ID	SNP location (number)	Reference
Geibberellin biosythesis	GA20OX3	R(A2)	AT2G34555	Bra028706	-	Li et al. 2013. DNA Research
Cytokinin oxidase/deh-ydrogenases	CKX5	E(A7)	AT1G75450	Bra015842	Intron & exon (2)	
	CKX1	J(A3)	AT2G41510	Bra000229	-	
Gast1 protein homolog	GASA1	E(A7)	AT1G7575	Bra015820	Exon (1)	
	GASA1	E(A7)	AT1G7575	Bra003743	Intron (2)	
Growth-requlating factor	GRF3	J(A5)	AT2G36400	Bra005268	Intron (1)	
	GRF3	J(A3)	AT2G36400	Bra023066	Intron & exon (3)	

표 23. 논문에서 현재 디자인된 SNP 마커

Primer name	Maker name	Primer sequence(5′-3′) Forward	Reverse	SNP Type	
SNP-Bra023066	SNP-GRF3a	CATTCATCATGAGAATATGCTTC	CATATCACAGATGAGATCAC	Chiifu	A
				Rcbr	G
SNP-Bra015842	SNP-CKX5	TGGCCCTATCCTTATCTACCC	AGAGAAAGATGAGCCGACGA	Chiifu	T
				Rcbr	G

※ Li et al. 2013. Quantitative Trait Loci Mapping in *Brassica rapa* revealed the structural and functional conservation of genetic loci governing morphological and yield component traits in the A, B, and C subgenomes of *Brassica* species. DNA Research

표 24. 마커 개발에 사용된 배추 소재 정보

순번	소재 정보
1	Chiifu 401 - 42
2	Rapid Cycling *B. rapa* (RCBr)

(자) 잎 털

유채 유전 정보 사이트 TAIR로부터 기 보고된 털(trichome) 관련 유전 정보를 배추에 서로 같은 (homolog한) 부위를 배추 유전체 데이터베이스인 BRAD에서 확인하여 그 정보를 수집하였다.

표 25. 마커 평가에 사용된 배추 소재 정보

순번	소재 등록번호	소재 정보
1	CNU_143006	Chiifu(MF)
2	-	RcBr
3	-	권심

표 26. 잎털 형성의 positive regulators

Positive regulators	애기 장대	배추
R2R3 MYB transcription factor GL1	AT3G27920	Bra025311
WEREWOLF (WER)	AT5G14750	Bra008740, Bra023486
MYB23	AT5G40330	Bra025589
bHLH factor GL3	AT5G41315	Bra025508
EGL3	AT1G63650	Bra027796, Bra027653
WD40-repeat factor TTG1	AT5G24520	Bra009770, Bra029411
GLABRA2 (GL2)	AT1G79840	Bra003535
TRANSPARENT TESTA GLABRA2 (TTG2)	AT2G37260	Bra005210, Bra023112

표 27. 잎털 형성의 negative regulators

Negative regulators	애기 장대	배추
CAPRICE (CPC)	AT2G46410	Bra004539, Bra039283
TRY	AT5G53200	Bra029089, Bra022637
ENHANCER OF TRY AND CPC1 (ETC1)	AT1G01380	Bra032635
ETC2	AT2G30420	
ETC3	AT4G01060	Bra037388, Bra000941, Bra008539
TRICHOMELESS1	AT2G30432	

표 28. Trichome 형성 주요 유전자 프라이머 정보

프라이머명	증폭산물의 크기 (bp)	구분	Reference
GL1	259(ms)	Myb-like protein, helps in induction of trichome development	Nayidu et al. 2014. PLOSone.
GL2	289(ms)	positive regulators	
EGL3-1	235(ms)	positive regulators	
EGL3-2	266(ms)	positive regulators	
TTG1-1	210(ms)	WD-40 protein involved in trichome development	
TTG1-2	237(ms)	WD-40 protein involved in trichome development	
TRY-1	212(ms)	negative regulators	
TRY-2	298(ms)	negative regulators	

(차) 저온 저항성 관련 유전자 및 관련 마커

저온 저항성 관련 유전자 *BrCSR*이 2014년, Yu et al. 에 의해 발표되었다. 또한 앞서 벼에서 연구된 AOX 관련 유전자에 관한 특성에 따르면 OsAOXla가 저온 저항성에 관련되며 QTL에 매우 조밀하게 존재해 서로 연관되어 있는 것을 밝혀내었고, 이를 토대로 배추에서 AOX 합성 관련 EST를 대상으로 다형성을 보이는 4개의 분자 마커가 보고되었다.

표 29. AOX 관련 유전자 분석을 통해 개발된 4개의 분자 마커 정보

AOX Marker	Primer name	Nucleotide sequence (5′ → 3′)
AO8 (DN963778.1)	AO8_F	ACCTGCTCCGGCTATC
	AO8_R	CGACCTTGGTAGTGAATGT
AO9 (EX098853.1)	AO9_F	AGCGTGAACACTTGGATCT
	AO9_R	GTAAATCATCAAGAGTAACGATTG
AO10 (ES930525.1)	AO10_F	GACAACACAAACGTTAACGG
	AO10_R	TCAAAGCGCCTGAGAGAT
AO11 (EX037292.1)	AO11_F	GCTGGAGCTTCCTTTAG
	AO11_R	CCATTTTGCATCTGTAAGTG

(카) 염 저항성 및 건조 내성 관련 유전자 및 관련 마커

배추 유전자 정보를 이용한 24K Oligo microarray chip을 이용한 발현량 평가 시험을 통해 저온, 염 저항성, 건조 내성 관련 유전자가 평가되었다. 이들 정보에서 염 및 건조 내성과 관련한 유전자 연관 서열을 확보하였다. 또한 염 저항성 관련 유전자 *BrSSR*이 2013년에 보고되었다.

표 30. 배추 유래 염 저항성 관련 유전자 *BrSSR*로부터 개발된 분자 마커

Primer name	Sequence (5'-3')
BrSSR-F	ATGGCTTCGTATTACTCTGCGT
BrSSR-R	CTACGCGACCACGAGTGTT

(3) 배추의 유용 형질 관련 분자마커 정보의 수집

3차 년도(2015년) 과제 수행 시 2차 년도 과제 수행 결과를 현재 구축한 Web DB에 업로드 하였으며, 조사한 분자마커들을 이용하여 보유하고 있는 자원들을 이용해 마커테스트를 진행 하였으며, 일부 유용 형질에 대하여 범용으로 사용할 수 있는 SNP 마커를 개발하였다.

(가) 잎 털

2차 년도(2014년)에는 형질에 관련된 문헌조사를 하였으며 그 결과 유채 유전 정보 사이트 TAIR로부터 기 보고된 털(trichome) 관련 유전 정보를 배추에 서로 같은 (homolog한) 부위를 배추 유전체 데이터베이스인 BRAD에서 확인하여 그 정보를 수집하였다.

3차 년도(2015년)에는 2차 년도 문헌조사를 통해 수집된 마커와 함께 추가로 문헌 조사를 하여 1개의 마커를 추가 수집하였고, SNP 마커 2개를 개발하였다(표 31). 마커의 검정력 평가는 새로 개발한 SNP 마커를 3가지의 배추를 시료로 이용하여 수행하였다. 소재정보는 다음과 같다(표 32). 마커 평가 수행 결과 표현형 조사와 마커 결과가 일치하는 것을 확인하였다(그림 6).

표 31. 털 형질 관련 신규 개발된 SNP 마커 정보

번호	프라이머명	후보유전자명
1	SB664, SB665	*TTG 1*
2	SB666, SB667	

표 32. 잎털 형질에 대한 신규마커의 검정력 평가에 사용된 배추 소재 정보

순번	소재 정보
1	지부
2	RcBr
3	권심

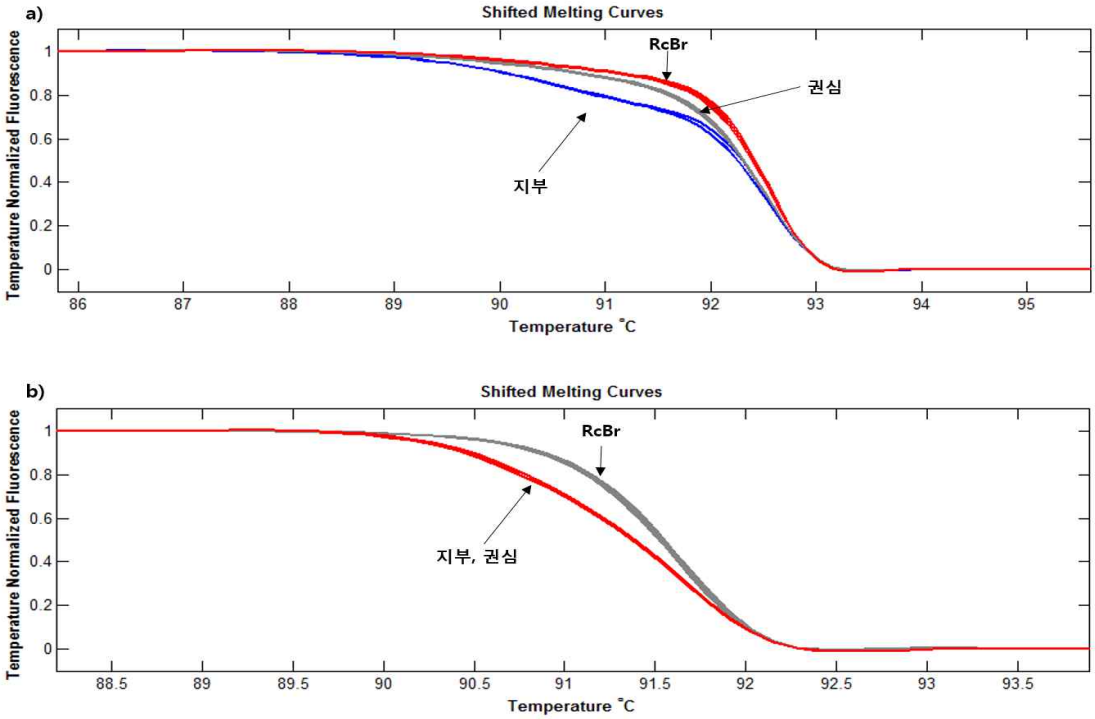


그림 6. 잎털 형질 관련 신규 개발된 SNP 마커 검정 결과.
a) SB664, SB665, b) SB666, SB667

(나) 결각(Lobation)

배추는 타가수정 식물이며, 잎의 형태, 수량성 등에 관여하는 양적형질 유전자좌가 밝혀졌다. 최근 선행연구에서 배추과 작물이며 대표적 모델식물인 애기장대에서 잎의 형태에 따라 관련된 Bolting time, Budding time, Flowering time의 3가지 양적형질 유전자가 보고되었다. 하지만 배추에서는 잎의 형태를 결정하는 양적형질 유전자좌는 발견되었지만 정확한 유전자는 밝혀지지 않았다.

선행연구에서 Bolting time, Budding time, Flowering time에 관련된 양적형질 유전자좌인 *BrFLC1*과 *BrFLC2,* early-flowering parental line에서 *BrFLC2*의 낮은 발현을 일으킬 수 있는 3개의 SNP들을 확인하였다. 또한, 잎의 lobe depth와 잎의 hairiness에 대하여 *GIBBERELLIN 20 OXIDASE 3*를 포함하는 syntenic 영역과 일치하는 하나의 주요 양적형질 유전자좌와 *BrGL1*을 포함하는 주요 양적형질 유전자좌가 확인되었다 (Li F, 2009). Erucic acid 및 Glucosinolate 함량, Flowering time, 수량 및 병저항성 형질, 잎의 형태(폭, 길이, 중륵 길이, 중륵 폭, 잎자루 길이)에 대한 양적형질 유전자좌가 보고되었으며(Li X, 2013) 특히, 잎의 모양에 따라 성분 함량과 광합성 양, Flowering time이 달라진다고 알려져 있다. 따라서 잎의 모양을 결정하는데 중요한 결각을 확인 할 수 있는 SNP 마커 3개를 개발하였다.

마커 개발에 이용한 후보유전자는 *GASA, AGL* 2가지이며 신규마커의 검정 결과 결각이 있는 RcBr과 결각이 없는 지부(chiifu) 간에 차이를 확인 할 수 있었다 (그림 7).

표 33. 결각 형질 관련 신규 개발된 SNP 마커 정보

프라이머명	후보유전자명
1번 **2번**	***GASA***
3번	***AGL***

표 34. 결각 형질 관련 신규마커의 검정력 평가에 사용된 배추 소재 정보

순번	소재 정보	소재 특징
1	RcBr	결각 있음
2	지부(chiifu)	결각 없음

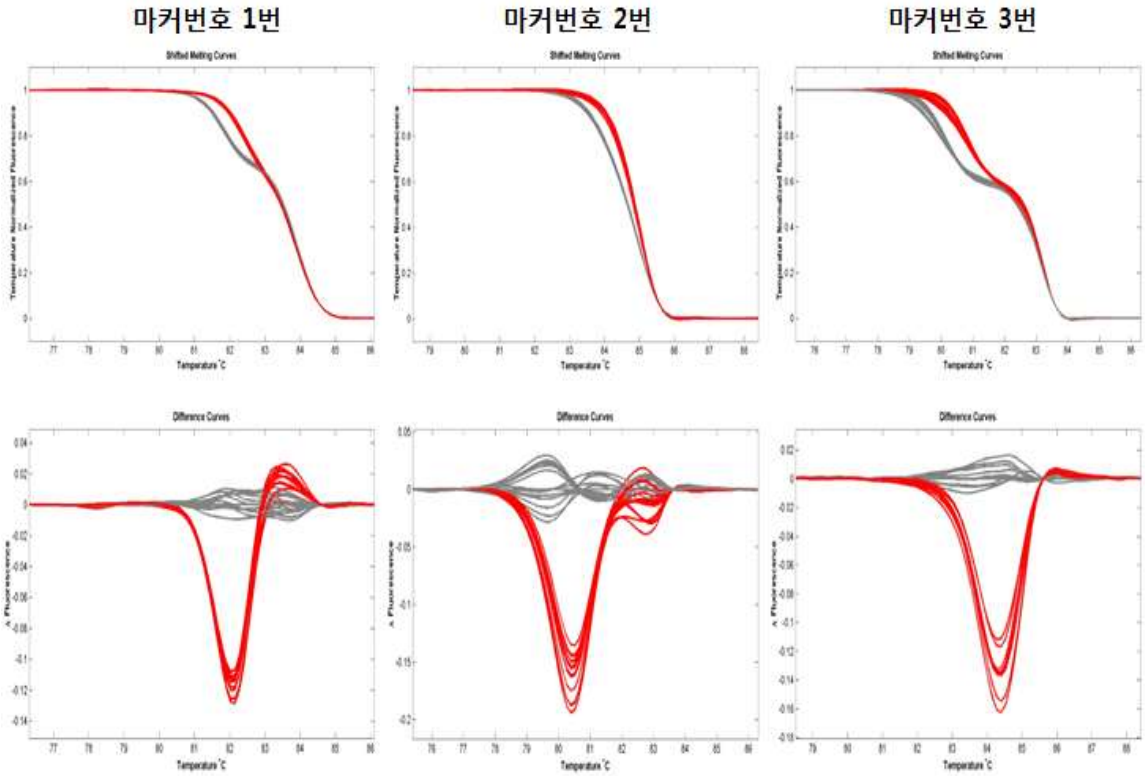


그림 7. SNP(Single Nucleotide Polymorphism) 분석 그래프
(빨간색 : Chiifu, 회색 : RcBr, 1번. 2번 : *GASA* gene, 3번 : *AGL* gene)

(다) GMO 35S promoter 표적 SNP 마커

35S promoter는 식물을 특이적으로 감염하는 Cauliflower Mosaic Virus (CaMV)의 일종으로, 이 바이러스는 뒤에 부착된 유전자를 매우 강하게 발현하는 특성이 있어 거의 모든 유전자 변형 식물에 존재하고 있다. 35S promoter 부분을 PCR방법으로 이용하여 검출함으로서 분석 시료의 GMO 생산 과정에서 이용된 외래 유전자의 형질 도입 여부를 판별 할 수 있다.

표 35. GMO 검정에 사용된 프라이머 정보

프라이머명	후보 유전자
SB0321 SB0322 SB660 SB661	CaMV 35S promoter

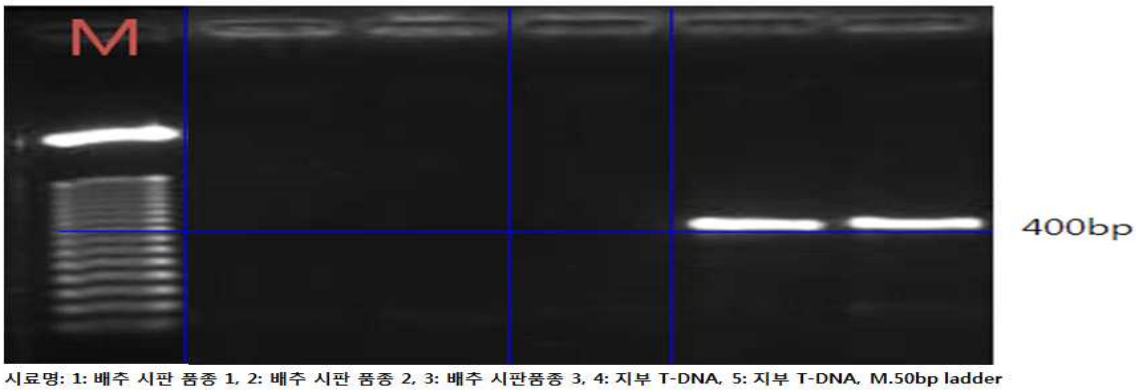


그림 8. SB0321, SB0322 마커를 이용한 GMO 검정 마커 평가 결과

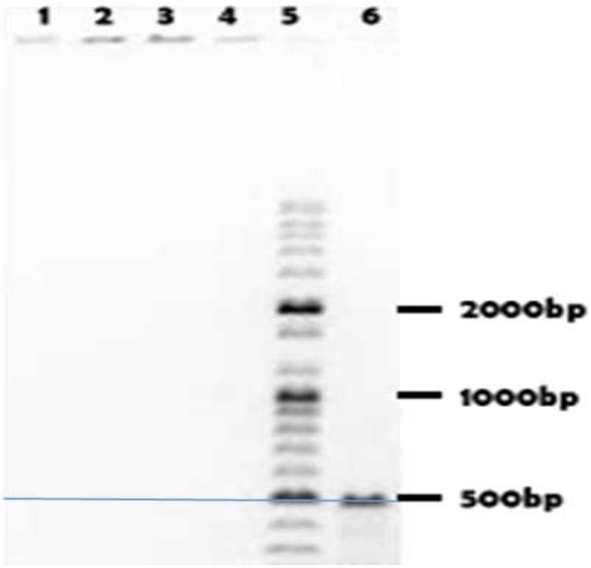


그림 9. SB660, SB661 마커를 이용한 GMO 검정 마커 평가 결과
(Lane 1 ~ 4 : 시중 판매 중인 양배추 F1 4개체, Lane 5: Size Marker, Lane 6 : 지부 T-DNA)

(4) 배추 수집단의 표현형 조사

3차 년도 과제수행 과정에서 re-sequencing 데이터를 활용한 배추 수집단의 GWAS(Genome-Wide Association Study) 연구의 기반 조성을 위하여 현재 보유하고 있는 192개의 inbred line과 9개의 엘리트라인을 대상으로 23가지 항목에 대해 표현형 변이를 조사하였다 (그림 10). 이 유용형질 표현형 조사는 충남대학교 농업 실험 포장에서 수행되었으며 각 계통이 보이는 표현형의 반복성을 입증하기 위하여 3년에 걸쳐 이루어 졌다. 조사한 항목은 초장, 무게, 엽병의 두께, 엽병폭, 구경, 엽폭, 엽병장, 구고, 엽수, 결구 내엽색 등 23가지의 항목에 대해 조사하였다. 표현형 데이터는 계통별 반복 값을 평균으로 나타내어 정리하였으며 내엽색과

같이 표현형에 구체적인 수치를 부여하기 어려울 경우, 나타나는 표현형을 분류하고 분류에 대한 계급값(indexing)을 부여하여 반복에 대한 평균값을 산출하는 방식으로 전산화하였다.

그림 10. 배추 유용 형질 표현형 조사 결과 일부

Lines	Plantheight	PlantWeight	Numberofleaves	Leaflength	Leafbladewidth	Petiolelength	Petiolewidth	Petiolethickness	Petiolecol
CHNU10066	36.0	1760.0	16.7	40.7	26.0	19.3	6.3	0.7	
CHNU10067	42.6	1596.0	20.4	41.4	22.3	24.6	5.0	0.7	
CHNU10068	38.6	1808.0	16.0	42.7	25.6	21.7	6.9	0.8	
CHNU10069	34.0	2626.7	13.3	34.3	24.5	20.0	5.7	0.6	
CHNU10070	36.7	2268.3	20.0	37.8	21.8	22.0	5.2	0.9	
CHNU10071	41.6	2290.0	18.8	62.0	25.0	24.4	6.6	1.6	
CHNU10072	46.5	1720.0	12.8	47.3	25.2	25.7	6.7	0.8	
CHNU10073	51.3	1246.7	14.3	50.0	22.1	31.1	4.6	0.7	
CHNU101048	33.3	1860.0	17.7	39.5	21.2	25.0	6.5	0.7	
CHNU120006	41.8	613.3	7.7	39.2	22.3	19.0	5.2	0.7	
CHNU120012	41.0	1523.3	9.3	41.7	19.7	23.3	7.3	1.1	
CHNU120027	107.7	383.3	11.7	29.5	13.6	13.8	1.2	3.2	
CHNU120031	55.7	733.3	71.5	57.7	18.9	38.7	2.2	3.9	
CHNU12013	42.5	1755.0	13.3	36.3	21.7	20.5	5.3	0.8	
CHNU12014	34.8	1760.0	15.7	30.7	22.4	18.2	5.7	1.0	
CHNU12015	24.3	1133.3	8.7	28.0	20.2	12.0	3.8	0.6	
CHNU25079	33.6	1876.0	13.3	34.4	22.4	15.8	6.8	1.1	
CHNU25083	24.3	1006.7	12.3	24.7	13.3	10.7	5.7	1.3	
CHNU25084	26.3	785.0	44.7	32.3	18.9	16.2	3.0	8.9	
CHNU26013	30.3	2155.0	19.0	32.5	18.8	143.8	5.6	1.0	

그림 11. 전산화 작업을 거친 배추 유용 형질 표현형 데이터의 일부

제2절 배추의 육종 특화 데이터베이스(DB) 구축을 위한 생물정보의 생산과 재가공

1. 배추 표준 유전체 서열과 유전자 정보의 수집

가. 배추 표준 유전체 정보의 확보

(1) 배추 표준 유전체의 기본 annotation 정보

1차 년도에는 배추의 육종 특화 데이터베이스 구축을 위한 배추 유전체의 기본 정보를 외부 데이터베이스로부터 확보하였다. BGI와 Phytozome에서 각각 *Brassica rapa* reference genome 및 annotation 정보를 Brassica Database(Wang et al, 2015)에서 확보하였으며, 1.5 version의 genome data를 기준으로서 Genome Browser (Tyner et al, 2017)의 구성에 필요한 track data를 추출하였다.

표 1. 배추 표준 품종 지부 유전체 및 유전자 서열

Data	길이 및 개수	버전	출처
Genome	약 284Mb, 10 chromosome	v1.5	BGI
Gene	41,020개	v1.5	BGI
Annotation info	41,019개 annotated genes	v9	Phytozome

(2) 배추 표준 유전체의 re-sequencing을 통한 전장 염기서열 정보의 재생산

1차 년도에는 향후, 다양한 계통에서의 변이 정보 재생산에는 높은 정확도를 갖는 reference genome의 물리지도가 필요하기 때문에 지부(chiifu)의 high-coverage re-sequencing을 수행하였다. Illumina Hiseq 2000 platform을 이용(Minoche et al, 2011)하여 insert size가 500bp인 paired-end 형태로 얻은 표준 유전체 서열을 read quality에 따라 quality check 및 trimming을 수행하였다. 이후 reference genome data와 비교 정렬을 통해 높은 genome coverage를 갖는 표준 유전체 서열을 얻었다.

표 2. 배추 표준 품종 지부 재분석 염기서열 raw data의 정보

File name	Num. of Reads	Avg. length (bp)	Total length (bp)	*Genome coverage
TN1309D4207_1.fq	195,071,034	100	19,507,103,400	≒137.45X
TN1309D4207_2.fq	195,071,034	100	19,507,103,400	

*Genome coverage : 각 샘플의 총 read 길이를 reference genome 의 총 길이로 나눈 값.

표 3. 지부의 re-sequencing을 통해 얻은 Reference genome의 assembly 결과

Number	Total Length (bp)	min	max	N50	avg.	N count
40,550	283,822,784	100	10,813,983	1,971,137	6,999	10,722,374

2. 수집한 배추 우수 계통의 변이 정보의 생산과 재가공

가. 배추 우수 계통의 re-sequencing을 통한 변이 정보의 대량 생산

(1) 배추 RIL (Recombinant Inbred Line)의 변이 정보 생산

1차 년도에서는 자체적으로 보유한 배추 RIL 26 계통(지부(chiifu)와 권심(kenshin)의 교배 집단)의 DNA를 추출한 후 Illumina HiSeq 2000 platform 방식으로 sequence read 라이브러리를 생산한 이후 reference genome에 대한 re-sequencing(Li et al, 2009)을 통해 RIL 26 계통의 염기서열을 생산하여 수집 계통의 변이 정보 재생산과 분석절차의 기반을 확립하였다.

표 4. RIL 26 계통의 re-sequencing 결과

Line name	Library number	Read length	Read coverage
279008	14,077,100	1,407,710,000	9.92X
279011	12,676,625	1,267,662,500	8.93X
279018	15,222,457	1,522,245,700	10.73X
279030	11,540,457	1,154,045,700	8.13X
279053	13,953,197	1,395,319,700	9.83X
279055	12,594,176	1,259,417,600	8.87X
279064	15,622,970	1,562,297,000	11.01X
279066	16,201,530	1,620,153,000	11.42X
279067	15,320,888	1,532,088,800	10.80X
279081	13,661,343	1,366,134,300	9.63X
279085	15,052,460	1,505,246,000	10.61X
279087	14,317,694	1,431,769,400	10.09X
279088	13,745,935	1,374,593,500	9.69X
279090	18,477,291	1,847,729,100	13.02X
279093	19,563,087	1,956,308,700	13.79X
279161	14,286,290	1,428,629,000	10.07X
279137	19,146,337	1,914,633,700	13.49X
279099	18,326,960	1,832,696,000	12.91X
279140	20,263,164	2,026,316,400	14.28X
279150	18,773,489	1,877,348,900	13.23X
279106	19,015,284	1,901,528,400	13.40X
279015	17,159,203	1,715,920,300	12.09X
279002	18,019,445	1,801,944,500	12.70X
279022	18,070,692	1,807,069,200	12.73X
279012	16,214,727	1,621,472,700	11.43X
279101	15,166,036	1,516,603,600	10.69X

(2) 배추 RIL (Recombinant Inbred Line) 변이 정보의 추가 생산

2차 년도 과제수행 과정에서 배추 RIL의 유전체 정보의 생산을 지속하였다. Illumina HiSeq 2000 platform에서 배추 교배친 및 우수자원 96 계통의 유전체 염기서열을 확보하기 위하여 96개 계통을 대상으로 paired-end read의 insert size를 350-550bp 수준으로 설정하여 sequence 라이브러리를 작성하였다. 이를 통해 배추 96 계통의 re-sequencing을 위한 sequence read data를 얻을 수 있었다.

표 5. 배추 RIL 96 계통의 re-sequencing 데이터 생산을 위한 read quality check 결과

Sample name	Read number(M)	Base number(Mb)	GC(%)	Q20(%)	Q30(%)
SSD073	33	2871.41	38.42	98.35	95.21
SSD350	33.82	2925.6	38.02	98.39	95.38
SSD186	28.71	2469.42	38.32	98.18	94.9
SSD357	29.5	2566.31	38.08	98.32	95.27
SSD129	29.53	2569.11	38.26	97.52	93.36
SSD351	32.01	2752.57	37.94	98.38	95.38
SSD093	17.5	1531.04	38.11	97.48	93.28
SSD287	32.66	2824.69	37.99	98.14	94.77
SSD068	39.13	3365.14	37.94	98.29	95.07
SSD044	31.24	2702.03	38.85	98.22	94.86
SSD042	41.6	3640.34	38.29	98.18	94.83
SSD086	32.42	2787.89	38.52	98.3	95.13
SSD377	29.65	2549.63	37.87	98.38	95.4
SSD344	36.63	3205.11	37.83	98.31	95.25
SSD232	28.37	2468.24	37.98	98.02	94.52
SSD309	30.57	2659.42	38.11	98.3	95.23
SSD110	37.28	3262.36	38.59	97.77	93.86
SSD064	31.98	2798.32	38.01	98.19	94.86
SSD080	36.7	3211.09	38.77	98.2	94.85
SSD051	33.61	2890.65	38.44	98.3	94.86

나. 배추의 계통 특이적 변이 데이터의 재생산과 분석

(1) 배추 유전체의 변이 데이터 생산을 위한 배추 우수 계통 re-sequencing 분석 현황

1차 년도 과제 수행 시 배추 표준 품종인 지부의 re-sequencing을 수행한 결과를 분석하였다. 염기서열 raw data의 총 sequence read의 수는 195,071,034개이며, 이의 총 길이는 약 18.2 Gbp이다. 이를 통해 생산한 re-sequencing data의 genome coverage 약 137X임을 확인함을 통해 이를 기준으로 다른 배추 계통의 변이 데이터를 안정적으로 생산할 수 있는 reference를 확보하였다. 또한 Illumina Hiseq 2000 platform 방식으로 염기서열 재분석(re-sequencing)을 통해 우수자원 26계통의 re-sequencing data를 생산하였다.

2차 년도 과제 수행 시 1차 년도에 이어 배추 RIL 26계통을 포함한 현재 보유 중인 192개 inbred line 계통의 대량 SNP 분석을 수행하였다. 이전 re-sequencing한 9개 계통과 192개 계통을 분석하여 총 201개 계통 사이의 MAF(minimum allele frequency)를 계산하여 5%이상의 MAF를 갖는 2,061,167개의 SNP를 얻었으며 20% 이상의 MAF를 갖는 272,014개의 SNP들을 얻었다.

(2) 배추 RIL의 re-sequencing 정보 분석을 통한 대량 SNPs 발굴

과제 수행 기간 동안 배추 RIL의 총 192개의 inbred line을 Illumina technology의 HiSeq 2000 platform을 이용하여 염기서열을 분석하였다. 분석 결과 NICEM에서 분석된 96개의 line의 전체 reads의 수는 대략 1,073,895,969개로 이는 38.12%의 평균 GC%를 갖는 108,463,492,869bp를 대상으로 하였으며 SEEDERS사에서 분석된 96개 line은 3,228,852,090 reads에서 322,885,209,000bp를 얻었다. 192개 계통의 re-sequencing data는 개별적으로 fastqc로 read의 quality를 phred score를 출력하여 확인하였으며 phred scrore가 30 미만으로 나타난 염기서열은 fastx-toolkit의 fastq-trimmer로 quality trimming을 수행하였다. 이와 같은 전처리 과정을 통해 chiifu whole genome sequence를 reference로 활용한 계통별 유전체 정보를 얻기 위한 read assembly와 192개 계통이 이루는 집단 내에서 나타나는 SNP calling을 위한 총 268,135,409,319bp 달하는 reads을 얻었다.

BWA aligner와 samtools로 chiifu whole genome을 reference로 이용하여 계통별 read sequence data를 alignment 및 mapped sequence data의 format을 구축한 파이프라인 내에서 변경하였다. GATK와 Picard tool은 중복된 read sequence data를 전체 data pool내에서 제거하고 정확성 평가, 정렬, 고정 그리고 InDel realignment의 수행에 이용되었다. 최종적으로 bcftool과 vcftool을 통해 각 계통에서 나타나는 SNP와 InDel들에 대한 reference genome을 기준으로 한 위치 정보를 생산하였다. Re-sequencing이 수행된 192개 계통과 이전 re-sequencing한 9개 계통을 더해 총 201개의

inbred line을 이용하여 분석한 결과 1차적으로 총 4,526,911개의 SNP를 얻었다. 이 SNP data pool을 high stringent condition으로 (quality value >=10, sequence depth >=3, mapping quality >=10, genotype quality >= 20, 최소 3계통이 homozygous SNP 보유) filtering하여 4,327,270개의 정제된 SNP data pool을 산출하였다. 201개 계통 사이의 MAF(minimum allele frequency)를 계산하여 5%이상의 MAF를 갖는 2,061,167개의 SNP를 얻었으며 20% 이상의 MAF를 갖는 272,014개의 SNP를 얻었다. 이와 같은 결과를 토대로 배추 reference genome내의 intron과 exon 위치를 산출하기 위한 최신버전(BRAD V. 1.5)의 gff 파일을 사용하여 유전체 상에 mapping이 되지 않은 scaffold는 제외하고 염색체의 각 위치를 기반으로 한 유전자에 대한 5% 이상의 MAF를 갖는 SNP들을 mapping하기 위해 새로운 Perl script를 개발하였다. 이를 활용하여 genic region상의 1,993,300개의 SNP들 중에서 574,449개의 SNP를 mapping한 결과, exon 영역과 intron 영역에 각각 340,135개, 234,314개의 SNP들이 위치하며 유전자간 영역에 1,418,851개의 SNP들이 분포하고 있음을 확인하였다.

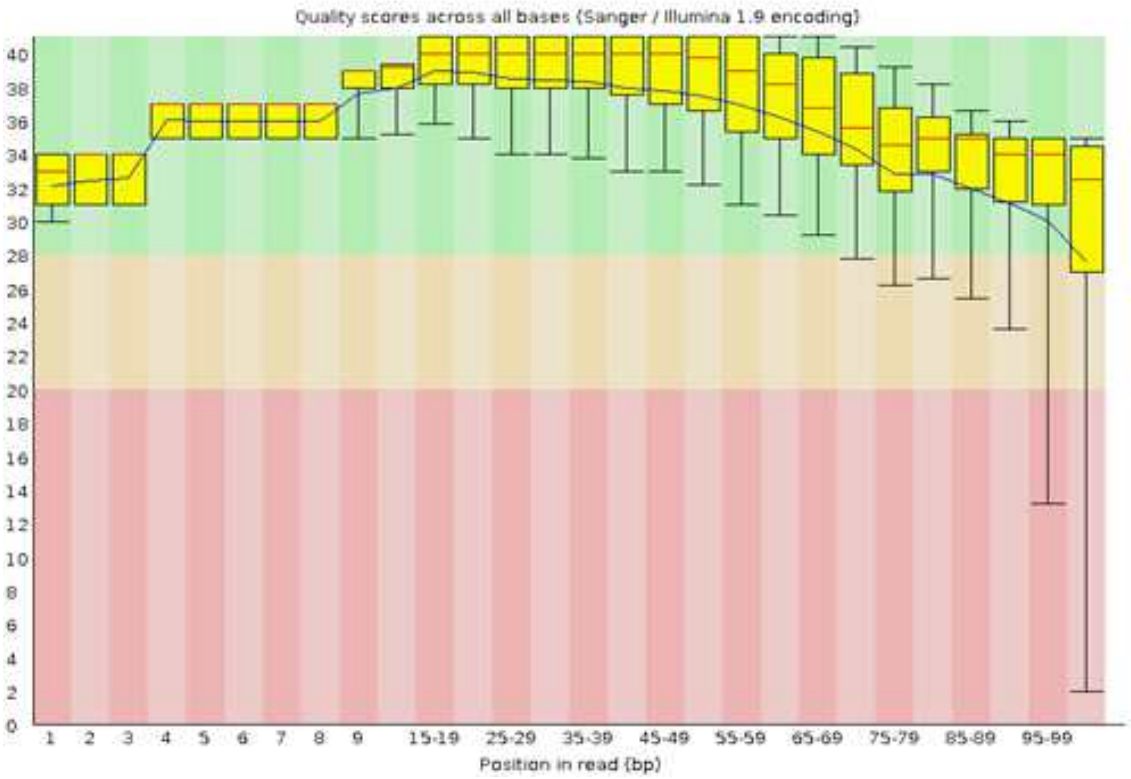


그림 1. Fastqc에 의한 배추 RIL 집단내 단일 계통의 quality check 수행 결과

다. 배추 수집단 re-sequencing 데이터와 표현형 데이터를 활용한 GWAS 분석

(1) 배추 수집단의 GWAS 분석을 위한 raw data 구성

3차 년도에는 보유하고 있는 192개의 inbred line (genome coverage: 3~5X)과 9개의 엘리트라인을 (2계통의 genome coverage: 30X, 7계통의 genome coverage: 10X) chiifu reference genome을 기준으로 대량으로 SNP정보를 산출하였다. 이를 기반으로 배추 수집단내의 계통들이 보이는 표현형과 유전적 변이의 연관 관계 파악을 위한 유전체 비교 연구를 집단내의 SNP 정보와 계통들의 표현형 정보를 활용하여 수행하였다. 23가지 표현형 항목은 충남대학교 농업 실험 포장에서의 3년간의 배추 수집단의 계통들을 재배하고 이에 대한 변이를 연차 및 개체 별 반복을 조사한 결과를 이용하였다.

GWAS 분석을 위한 SNP선발을 위해 Minimum Frequency Allele (MAF) 값에 제한을 두었을 때, MAF 5 이상을 만족하는 최종 SNP 수는 2,061,167개였다. 여기에서 필터링된 SNP matirx를 TASSEL ver 5.0 소프트웨어를 이용하여 맨하탄 플롯을 작성하였다 (그림 3). 이후 유전체 정보 분석으로 확보한 SNP를 대상으로 표현형 데이터와 연관된 SNP 분석을 수행하였다. 이 중 확인한 표현형은 엽수, 초장에 관한 내용만 본 보고서에 기술하였다.

표 6. Variants calling 과정의 조건별 SNP의 수

SNP 산출 과정의 조건	산출된 SNP의 개수
SNPs above 5 Minor Allele Frequency	2,061,167
SNPs removed which found in non-anchored scaffolds	172,498
Final SNPs considered for GWAS and marker development	1,888,669
SNPs mapped on BRAD Genes	613,473
SNPs mapped on Exons	360,937
SNPs mapped on Introns	252,535

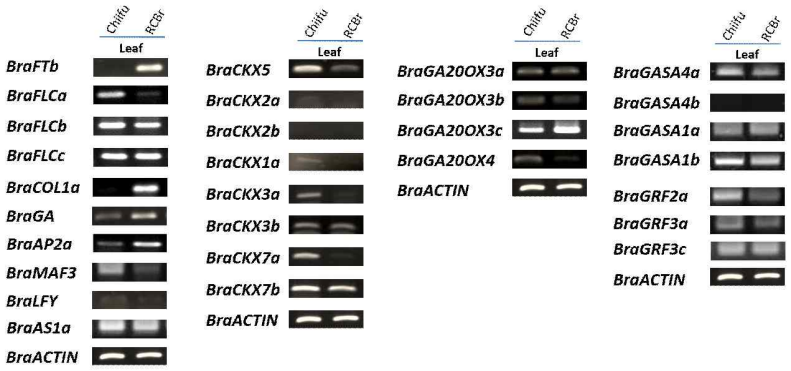


그림 2. 23개 표현형에 대한 중요 candidate gene의 RT-PCR을 통한 발현량

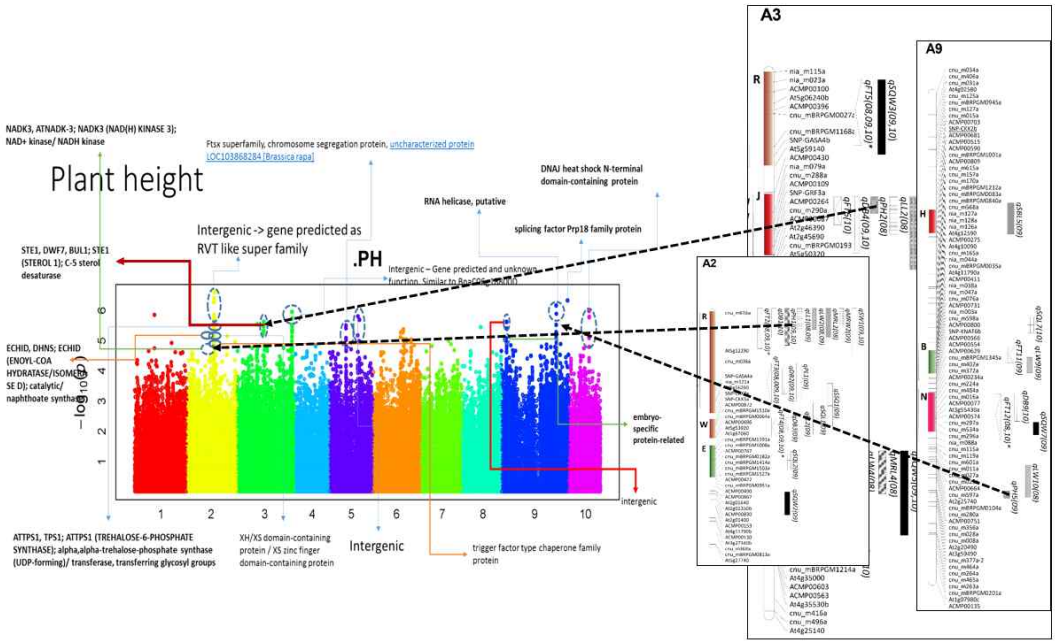


그림 3. 초장에 대한 QTL 문헌정보와 GWAS 분석 결과의 비교

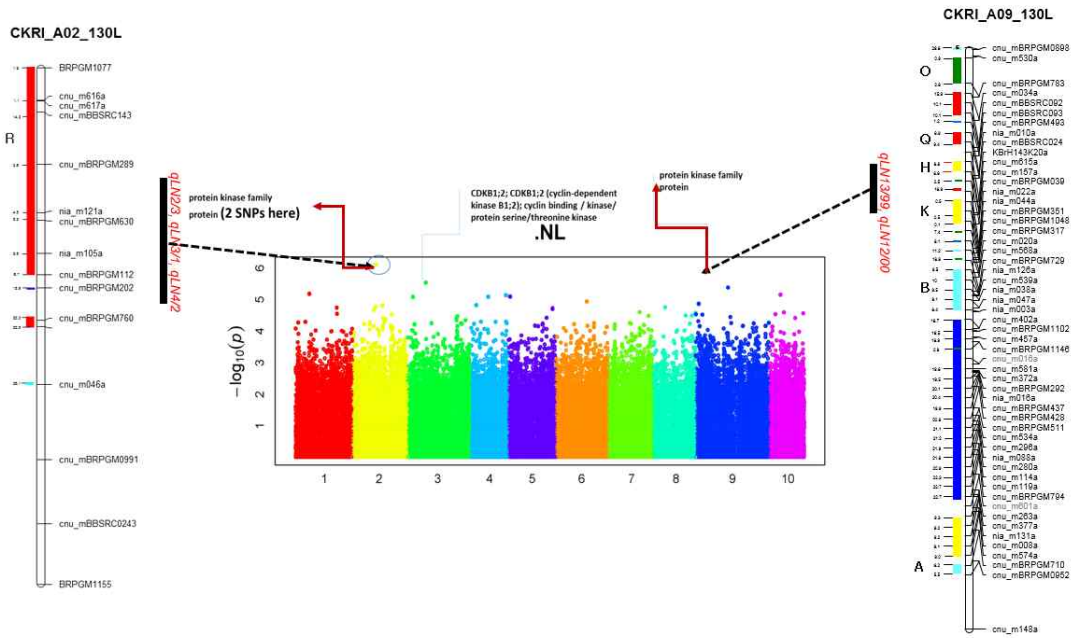


그림 4. 엽수에 대한 QTL 문헌정보와 GWAS 분석 결과의 비교

3. 배추 조직 특이별 유전자 식별을 위한 전사체 정보의 생산과 재가공

가. RNA-seq을 통한 배추의 조직별 transcriptome 데이터의 생산

(1) 지부(chiifu)의 꽃과 뿌리 조직의 transcriptome 데이터 생산

1차 년도에는 지부의 꽃과 뿌리에서 RNA를 추출하여 RNA-seq(Wang et al, 2009)을 이용하여 transcriptome 라이브러리를 생산하였으며 이를 sequencing하여 read 데이터를 생산하였다. 이를 reference genome 상에 align하여 유전자 coding 영역에 mapping된 read의 개수를 세어 유전자별 조직 특이적인 발현량을 얻을 수 있었다.

표 7. 배추(지부) 2개 전사체 샘플 (raw data) 통계치

File name	Sample description	No. of reads	Avg. length	Total length (bp)
TN1310R4460_1.fq	Flower	23,757,981	100	2,375,798,100
TN1310R4460_2.fq		23,757,981	100	2,375,798,100
TN1310R4515_1.fq	Root	21,943,854	100	2,194,385,400
TN1310R4515_2.fq		21,943,854	100	2,194,385,400

표 8. 배추 2개 전사체 샘플의 서열 퀄리티에 따른 전처리 (trimming) 통계치

File name	Sample description	No. of reads	Avg. length	Total length (bp)	Trimmed/ Raw (%)
TN1310R4460_1.fq	Flower	21,889,750	92.68	2,028,713,154	85.39%
TN1310R4460_2.fq		21,889,750	89.07	1,949,713,347	82.07%
TN1310R4515_1.fq	Root	20,260,241	92.73	1,878,771,170	85.62%
TN1310R4515_2.fq		20,260,241	89.24	1,808,034,622	82.39%

표 9. 배추 2개 전사체 샘플의 mapping 통계치

File name	Sample description	Total Reads	Mapped Reads	Unmapped Reads	Mapping Rate
TN1310R4460_1.fq	Flower	21,889,750	16,537,548	5,352,202	75.55%
TN1310R4460_2.fq		21,889,750	16,631,656	5,258,094	75.98%
TN1310R4515_1.fq	Root	20,260,241	15,798,873	4,461,368	77.98%
TN1310R4515_2.fq		20,260,241	15,880,415	4,379,826	78.38%

(2) 지부(chiifu)의 세분화된 잎 조직의 전사체 데이터 생산

2차 년도 과제 수행 시 1차 년도에 이어 RNA-seq을 이용하여 배추 지부 품종의 어린 잎(4주), 결구 후 외엽, 결구 후 내엽의 전사체 서열을 생산하였다. 아래의 일련의 표 10-12 에서는 1차 년도에 조사한 조직을 포함한 5개의 조직의 전사체에서 얻은 평균 100bp 길이의 paired-end 방식의 RNA-seq의 read 서열에 대한 정보를 나타낸다.

표 10. 배추(지부) 5개 전사체 샘플 (raw data) 통계치

File name	Sample description	No. of reads	Avg. length	Total length (bp)
TN1310R4460_1.fq	Flower	23,757,981	100	2,375,798,100
TN1310R4460_2.fq		23,757,981	100	2,375,798,100
TN1310R4515_1.fq	Root	21,943,854	100	2,194,385,400
TN1310R4515_2.fq		21,943,854	100	2,194,385,400
TN1312R4974_1.fq	Inner Leaf of Head	22,928,264	100	2,292,826,400
TN1312R4974_2.fq		22,928,264	100	2,292,826,400
TN1312R4977_1.fq	Outer Leaf of Head	23,995,399	100	2,399,539,900
TN1312R4977_2.fq		23,995,399	100	2,399,539,900
TN1312R4978_1.fq	Young Leaf	24,478,869	100	2,447,886,900
TN1312R4978_2.fq		24,478,869	100	2,447,886,900

표 11. 배추 5개 전사체 샘플의 서열 퀄리티에 따른 전처리 (trimming) 통계치

File name	Sample description	No. of reads	Avg. length	Total length (bp)	Trimmed/ Raw (%)
TN1310R4460_1.fq	Flower	21,889,750	92.68	2,028,713,154	85.39%
TN1310R4460_2.fq		21,889,750	89.07	1,949,713,347	82.07%
TN1310R4515_1.fq	Root	20,260,241	92.73	1,878,771,170	85.62%
TN1310R4515_2.fq		20,260,241	89.24	1,808,034,622	82.39%
TN1312R4974_1.fq	Inner Leaf of Head	21,328,689	92.49	1,972,777,586	86.04%
TN1312R4974_2.fq		21,328,689	89.65	1,912,040,058	83.39%
TN1312R4977_1.fq	Outer Leaf of Head	22,243,406	92.67	2,061,293,739	85.90%
TN1312R4977_2.fq		22,243,406	90.07	2,003,469,969	83.49%
TN1312R4978_1.fq	Young Leaf	22,638,293	92.62	2,096,733,472	85.65%
TN1312R4978_2.fq		22,638,293	89.86	2,034,362,739	83.11%

표 12. 배추 5개 전사체 샘플의 mapping 통계치

File name	Sample description	Total Reads	Mapped Reads	Unmapped Reads	Mapping Rate
TN1310R4460_1.fq	Flower	21,889,750	16,537,548	5,352,202	75.55%
TN1310R4460_2.fq		21,889,750	16,631,656	5,258,094	75.98%
TN1310R4515_1.fq	Root	20,260,241	15,798,873	4,461,368	77.98%
TN1310R4515_2.fq		20,260,241	15,880,415	4,379,826	78.38%
TN1312R4974_1.fq	Inner Leaf of Head	21,328,689	15,994,012	5,334,677	74.99%
TN1312R4974_2.fq		21,328,689	16,064,209	5,264,480	75.32%
TN1312R4977_1.fq	Outer Leaf of Head	22,243,406	17,964,791	4,278,615	80.76%
TN1312R4977_2.fq		22,243,406	18,040,484	4,202,922	81.10%
TN1312R4978_1.fq	Young Leaf	22,638,293	18,527,449	4,110,844	81.84%
TN1312R4978_2.fq		22,638,293	18,598,360	4,039,933	82.15%

(3) 배추 long shelf life 관련 계통의 RNA-seq 과 gene expression profiling

3차 년도에는 배추의 저장성에 관여하는 candidate gene의 탐색을 목표로 배추 수집단 내에서 표현형 검정을 거쳐서 선발된 배추 2품종 (수확 후 느린 노화: 27,142, 수확 후 빠른 노화: 27,160)에 대한 전사체 분석을 수행하였다.

선발된 두 계통의 잎 조직으로부터 RNA를 추출하고 RNA-seq을 거쳐 수확 후 저장 가능 기간에서 차이를 보이는 두 계통의 전사체에 대한 insert size 500, paired end 형식의 read 라이브러리를 구축하였으며. 이 read data는 Illumina Hiseq 2000기기를 이용하여 sequence 데이터를 생산하였다.

표 13. RNA sequencing을 통해 생산된 short reads 통계치

File name	Line name	No. of reads	Avg. length	Total length (bp)
TN1503R1438_CAGATC_R1.fastq	27160	35,549,062	100	3,554,906,200
TN1503R1438_CAGATC_R2.fastq		35,549,062	100	3,554,906,200
TN1503R1439_ACTTGA_R1.fastq	27142	36,588,343	100	3,658,834,300
TN1503R1439_ACTTGA_R2.fastq		36,588,343	100	3,658,834,300
Total	2 ea	144,274,810	100	14,427,481,000

생산된 두 계통의 Raw sequence 데이터는 SolexaQA 패키지의 DynamicTrim과 LengthSort를 이용하여 phred score 20이하의 bad quality 염기서열 제거와 25bp 이하의 짧은 read를 제거하는 전처리 과정을 수행하였다(Cox et al, 2010). 분석한 두 계통에서 평균 87.57bp의 133,194,178 read를 얻어 11,663,818,580bp의 염기서열을 확보하였다.

표 14. 노화 형질 관련 계통들의 RNA-seq을 통해 얻은 최종 염기서열 통계치

File name	Line name	No. of reads	Avg. length	Total length (bp)
TN1503R1438_CAGATC_R1.fastq	27160	32,898,882	90.54	2,978,785,150
TN1503R1438_CAGATC_R2.fastq		32,898,882	85.03	2,797,326,990
TN1503R1439_ACTTGA_R1.fastq	27142	33,698,207	90.33	3,043,877,124
TN1503R1439_ACTTGA_R2.fastq		33,698,207	84.39	2,843,829,316
Total	2 ea	133,194,178	87.57	11,663,818,580

Quailty check와 trimming이 완료된 read data는 배추 표준 유전체를 reference로 이용한 tophat으로 alingment를 수행하고 DESeq 파이프라인으로 두 계통간의 유전자 발현량을 산출하고 DEG를 추출하였다(Anders and Huber, 2010). 또한 두 계통에 대해 DEG로 나타난 mRNA의 구조를 비교하기 위해 Trinity를 통해 de novo assembly를 수행하여 후속 연구를 진행 중이다(Haas et al, 2013).

나. 배추의 조직 특이적 전사체 발현 정보의 재생산과 분석

(1) 배추의 조직 특이적 DEG (Differentially Expressed Gene)의 식별

2차 년도에는 RNA-seq을 수행한 5 조직 간에서 발현의 차이가 유의하게 나타나는 유전자들을 선별하고 조직별 발현량에 대한 분석을 수행하였다. DESeq R package를 이용하여 Control로서 선정한 조직인 어린잎에서 나타난 유전자의 발현량이 타 조직에서의 발현량과 2배 이상 차이가 나타나며 비교하고자 하는 샘플들의 평균 발현의 정도(mapping read의 값)가 200 이상, Binormial test에 의한 adjust p-value가 ≤ 0.01를 만족하는 유전자를 DEG로 판별하였다.

표 15. DEG 선발 결과 유전자의 수

비교대상 (control vs treatment)	Regulation pattern	Num. of DEGs	Num. of annotated DEGs
Young Leaf vs. Inner Leaf of Head	Up	1,477	1,477
	Down	2,400	2,400
Young Leaf vs. Outer Leaf of Head	Up	1,844	1,844
	Down	1,904	1,904
Young Leaf vs. Root	Up	2,695	2,695
	Down	4,191	4,191
Young Leaf vs. Flower	Up	2,335	2,334
	Down	3,793	3,793

*** Control은 배추 young leaf 샘플을 사용**

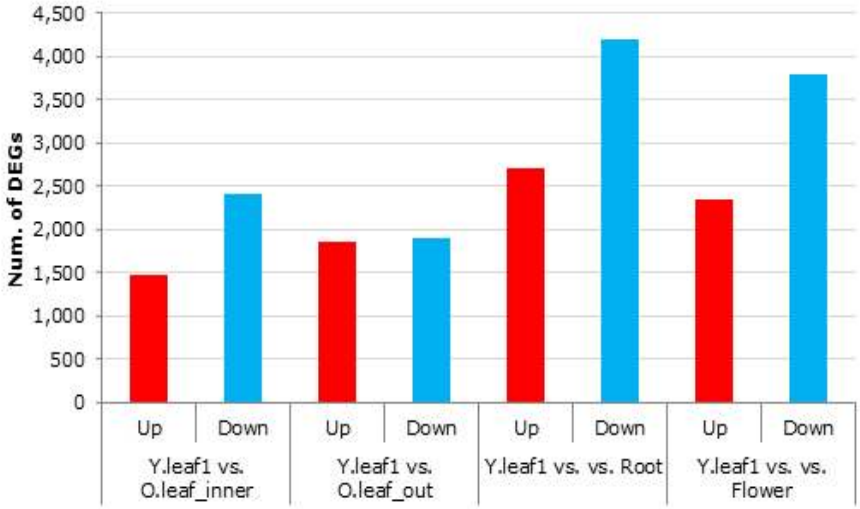


그림 5. DEG 선발 결과 유전자의 수

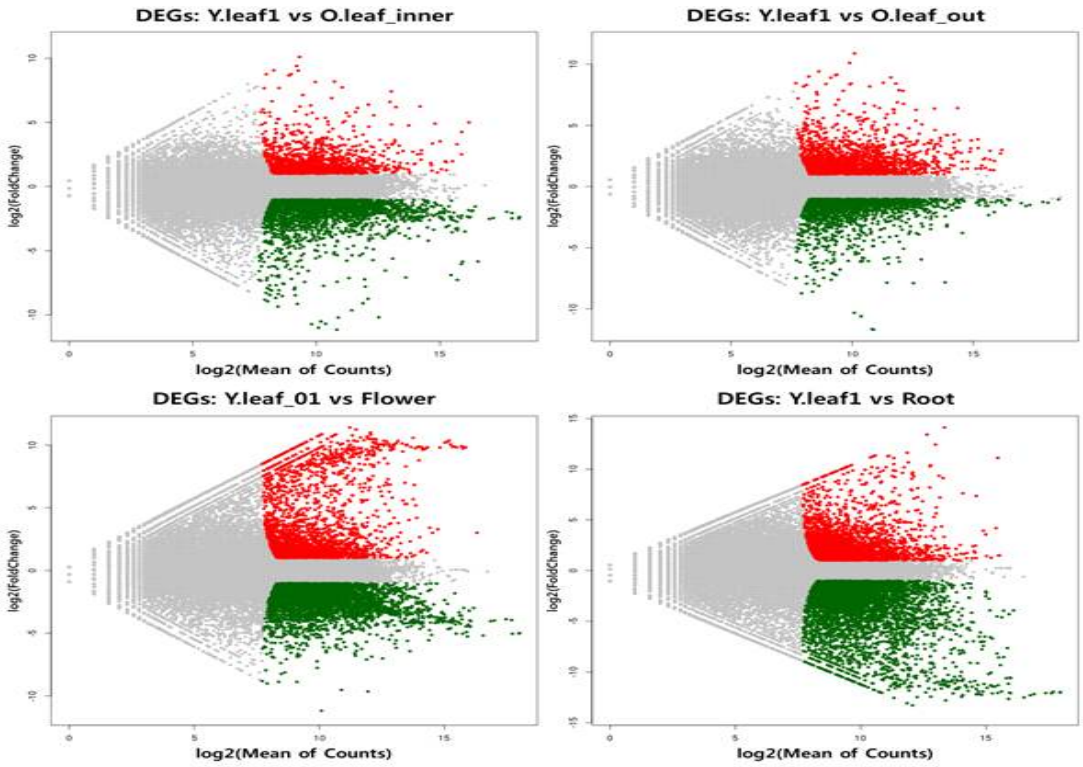


그림 6. Young leaf vs Old leaf의 유전자 발현 DEG를 시각화한 MA plot

(2) 배추 유전체의 유용 형질 관련 유전자 구조 정보 도식화

배추 유전자와 ortholog 관계를 갖는 Arabidopsis gene을 input 데이터로 준비하고 이를 DAVID에서 제공하는 Pathway Search Tool을 사용하여 배추 유전자에 대한 KEGG pathway 정보를 추출하였다(Hwang et al, 2009). Thresholds는 기능별 유전자 count ≥ 5, EASE ≤ 0.1 이다. Young_leaf_vs_Flower의 DEGs의 KEGG 일부 결과는 아래 표와 같다.

표 16. 비교 대상 간 DEGs의 KEGG pathway 예시 (young leaf vs flower)

Accession	KEGG Term	TAIR count	Gene count	top hit
ath00710	Carbon fixation in photosynthetic organisms	40	83	75
ath00195	Photosynthesis	35	71	73
ath00196	Photosynthesis	14	30	19
ath00260	Glycine, serine and threonine metabolism	23	35	44
ath00010	Glycolysis / Gluconeogenesis	37	58	86
ath00620	Pyruvate metabolism	30	46	65
ath00970	Aminoacyl-tRNA biosynthesis	22	31	47
ath00910	Nitrogen metabolism	20	37	42
ath00051	Fructose and mannose metabolism	21	36	45
ath00053	Ascorbate and aldarate metabolism	16	25	31
ath00500	Starch and sucrose metabolism	34	53	88
ath00860	Porphyrin and chlorophyll metabolism	14	22	26

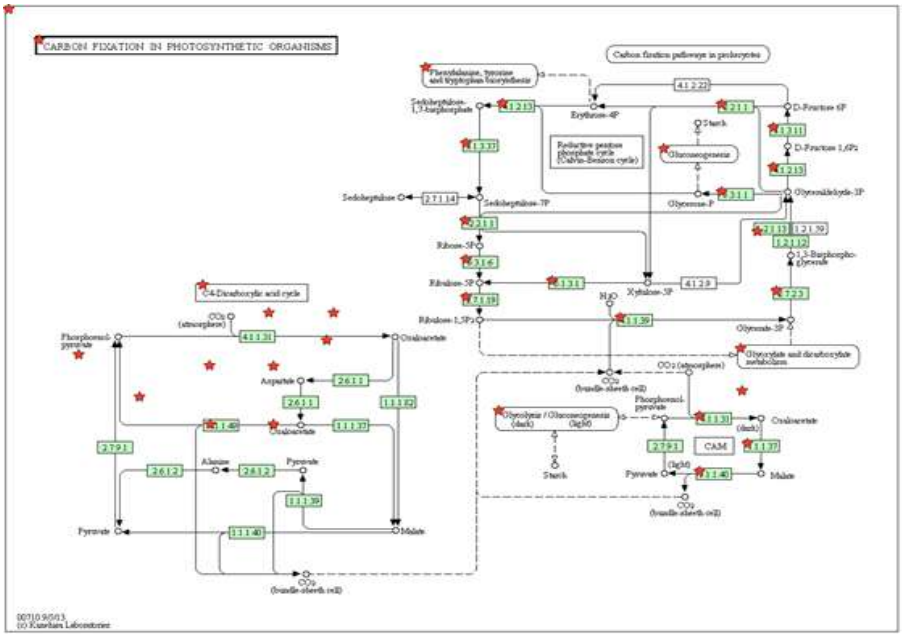


그림 7. 비교 대상 간 DEGs KEGG pathway 예시
(Carbon fixation in photosynthetic organisms)

(3) 배추 조직별 전사체 정보를 이용한 형질 관련 유전자 발굴

배추 assembled transcripts의 full-length 분석을 위해 phytozome에서 제공되는 *Brassica rapa*의 유전자의 서열을 비교한 결과 총 73,544개의 transcripts 중 31,877개 transcripts (Locus. 18,437개)가 reference genome에 90% 이상 cover되면서 reference gene 5'UTR, 3'UTR을 보유하고 있어 full-legth gene임을 확인하였다. Assembled transcripts와 reference gene을 비교하여 assembly된 배추 transcripts가 예측된 유전자를 어느 정도 대변하는지 확인하고, 예측된 유전자의 정확성을 판단하기 위해 배추 assembled transcripts을 genome에 align을 하였다. Alignment을 위해 사용한 tools은 BLAT이며, 기본옵션을 배추 transcripts을 draft genome에 mapping 한 후 100kb 이내에 assembly 된 transcripts의 50% 이상이 존재하는 경우 draft genome에 align을 하였다.

표 17. Reference gene과 orthologous인 transcripts의 full-length 분석

	Locus	Transcripts
전체 서열의 수	38,366	73,544
Full-length 로 판정되는 서열의 수	**18,437**	**31,887**

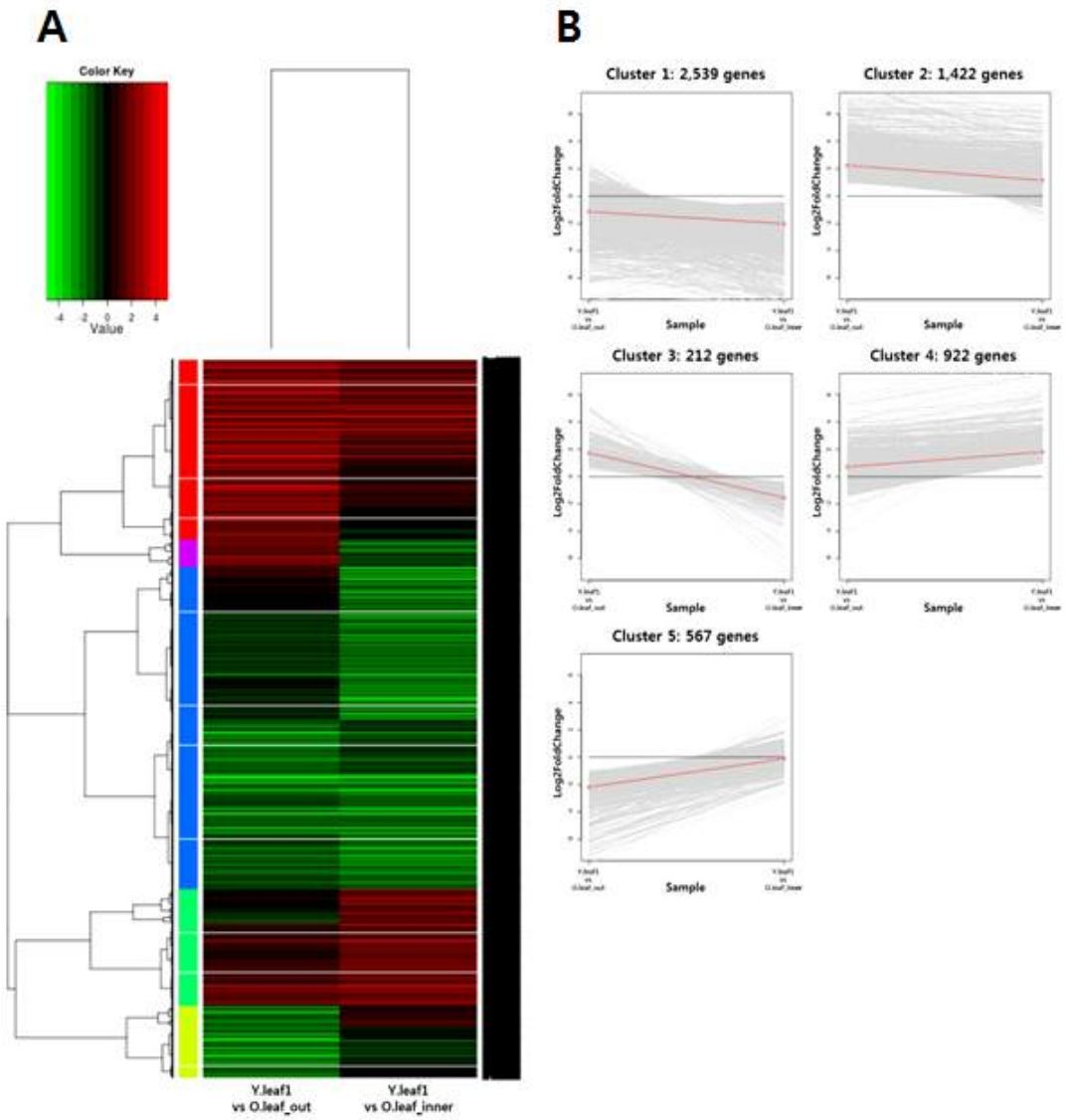
(4) 유전체 정보를 이용한 형질 관련 배추 transcriptome의 발현량 정보 도식화

2차 년도에는 DEGs를 산출한 결과 중 Y.leaf1 vs O.leaf_out, Y.leaf1 vs O.leaf_inner에서 DEGs로 산출된 유전자(5,662개)를 모두 모아 clustering 분석을 수행하였다. clustering 분석은 유전자간의 발현 패턴의 유사도를 pearson's correlation으로 계산하였으며, 발현 패턴이 유사한 유전자간의 결합 방식은 complete 방식을 사용하였다. 본 분석의 clustering 결과는 총 3개의 cluster set으로 구성되어 있다.

조직별 유의한 특이적 발현 유전자의 clustering 결과 clustering의 유전자 발현 패턴은 Y.leaf1 vs O.leaf_out, Y.leaf1 vs O.leaf_inner의 순으로 column을 정렬하였다. 각 Cluster의 발현 패턴이 유사한 유전자의 수는 Cluster 1에서 2,539개로 가장 많았고 Cluster 3에서 212개로 가장 적었다.

표 18. 각 Cluster에 포함된 DEG의 개수

Cluster #	Num. of DEGs in cluster
Cluster 1	2,539
Cluster 2	1,422
Cluster 3	212
Cluster 4	922
Cluster 5	567
Total	**5,662**



A) Heatmap은 왼쪽부터 Y.leaf1 vs O.leaf_out, Y.leaf1 vs O.leaf_inner 순으로 데이터를 표현함. B) Line plot은 heatmap에 표현된 cluster를 표현.

그림 8. Clustering분석 결과의 Heatmap과 Line plot

Gene ID	old_leaf 01_1	old_leaf 01_2	old_leaf 02_1	old_leaf 02_2	young_leaf 01_1	young_leaf 01_2	Desc
Bra021464	8574	8558	14501	14621	1202	1204	CONSTANS-like 2
Bra008668	7125	7155	13101	13181	2043	2029	CONSTANS-like 1
Bra011711	5442	5461	245	252	2681	2750	PQ-loop repeat family protein / transmembrane family protein
Bra004491	5381	5374	29905	30124	3288	3282	cytochrome P450, family 709, subfamily B, polypeptide 2
Bra009007	5249	5259	6669	6701	4808	4891	Eukaryotic aspartyl protease family protein
Bra021734	4795	4750	8018	7807	514	491	salt tolerance homologue
Bra018249	4048	3997	7896	7680	611	575	salt tolerance homologue
Bra031080	3441	3485	5697	5876	1208	1205	Erythronate-4-phosphate dehydrogenase family protein
Bra039070	3220	3413	4469	4912	2064	2069	photosystem II light harvesting complex gene 2.3
Bra033022	3096	3272	4045	4333	1880	1881	photosystem II light harvesting complex gene 2.3
Bra000391	3067	3101	3293	3309	1149	1133	cytochrome P450, family 704, subfamily A, polypeptide 2
Bra034163	2903	2938	2145	2185	1162	1159	receptor like protein 34

그림 9. 형질 관련 배추 transcriptome의 발현량 정보(Old leaf vs Young leaf)

4. 배추 뿌리혹병 (*Plasmodiophora brassicae*) 유전체의 염기서열 해독

뿌리혹병은 배추를 포함한 십자화과 작물에서 뿌리 조직에 기생 후 근계에 gall을 형성하여 식물의 양분 및 수분 흡수를 저해시켜 정상적인 생육을 저해한다. 이는 작물의 예상 수량을 저하시키고 병 방제에 대한 비용의 소모를 강요하여 국내에서는 배추 농가의 소득 저하에 한 원인으로 지목되고 있다. 배추의 뿌리혹병을 매개하는 rhizaria에 속하는 미생물로 알려진 *P. brassicae*는 과거로부터 전세계적으로 재배되고 있는 십자화과 작물에서 많은 피해 사례의 원인으로 알려졌지만 이 미생물 자체의 유전적 연구는 큰 진전이 없는 상태였다. 3차 년도에는 해외 연구팀과의 협업으로 *P. brassicae*의 염기서열을 해독하여 유전체 물리 지도를 구성하였다. 그리고 유전체 상의 유전자를 예측하고 그에 대한 annotation을 수행하였으며 유전체가 해독된 근연종 연관관계를 확인하여 *P. brassicae*의 진화상의 위치를 특정하였다. 본 연구의 결과는 "The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases" 라는 제목으로 2015년 SCIENTIC REPORTS 5:11153에 게재되었다(Schwelm et al, 2015).

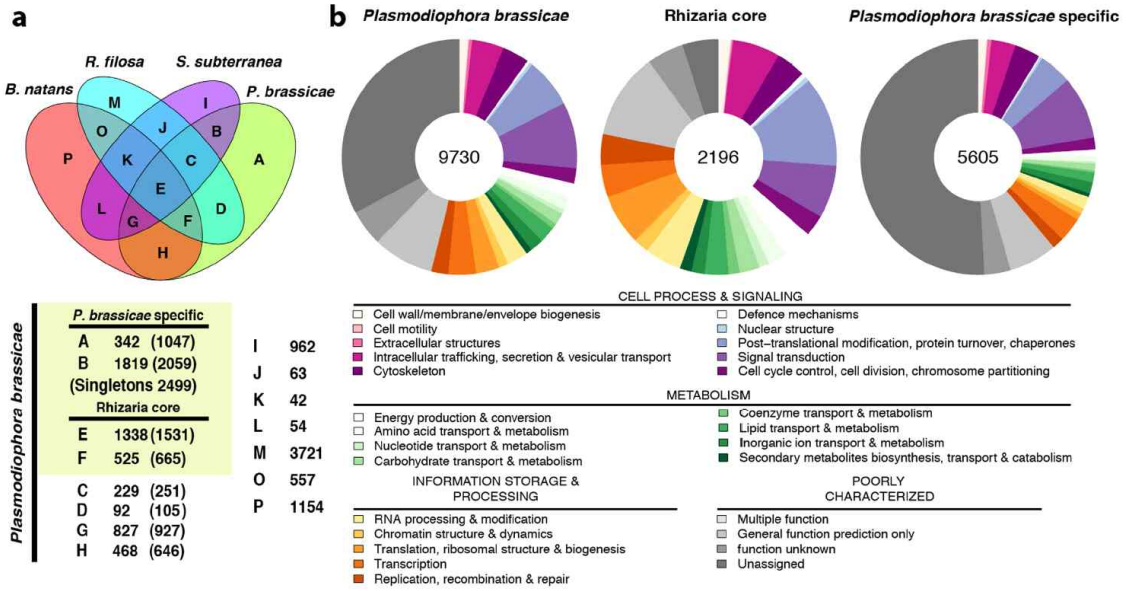


그림 10. *P. brassicae*와 근연종간의 유전자의 기능을 비교

(a) *P. brassicae, S. subterranea*와 감염원이 아닌 Rhizarian *B. natans*와 *R. filosa*의 OrthoMCL protein families 벤다이어그램. (b) KOG functional categories를 이용한 *P. brassicae* proteins의 기능분석을 통해 *P. brassicae*가 갖는 고유 유전자를 확인.

제3절 생산 및 수집한 생물정보에 기반한 배추의 육종 특화 데이터베이스(DB) 구축과 운영

1. 배추 분자육종 활성화를 위한 특화된 데이터베이스 구축

가. 배추 분자육종을 위한 특화 데이터베이스의 구축

(1) 배추 유전체 정보 기반 데이터베이스의 디자인

1차 년도에는 배추 분자육종 활성화를 위한 특화된 데이터베이스를 구축하기 위하여 구성요소, 제공 내역, 사용자 편의성을 고려하여 유전체 기반의 데이터베이스를 디자인하였다. Resource (수집된 배추 유전자원), Genome browse, Annotation, MAB, Tissue Specific Gene, KEGG, BLAST로 Web DB 메뉴를 구성하고, 심플한 UI로 데이터베이스를 구현하였다.

배추 분자육종 특화 데이터베이스는 http://168.188.15.201/cabbage/의 웹상의 주소로 접속이 가능하며 과제 수행과정에서 생산 및 수집한 표준 유전체 정보와 연동되며 배추의 표준유전자 정보를 다양한 공개 annotation DB 정보와 연결하여, 사용자에 의해 선택된 유전자의 annotation 및 육종 관련 정보의 시각화를 목표로 설계되었다.

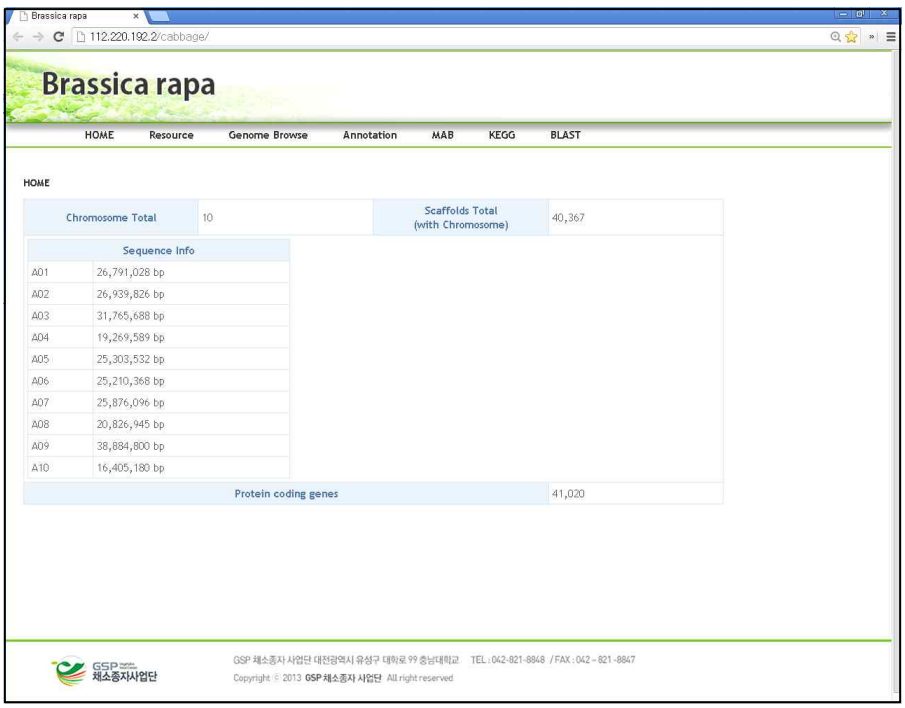


그림 1. 배추의 육종 특화 web DB의 메인페이지

(2) 배추의 모델 식물인 지부(chiifu)의 전사체 데이터를 이용한 분자육종 지원

1년차 과제 수행에서 생산한 배추의 모델 식물인 지부(chiifu)의 꽃(flower)과 뿌리(root)에서 RNA-seq을 통해 전사체(transcriptome) 염기서열을 생산하고 이를 분석한 결과를 데이터베이스에 연동하였다.

사용자가 조직특이성을 보이는 유전자를 (tissue specific gene)을 쉽게 확인할 수 있게 배추 전사체 분석 페이지를 디자인하였으며 선택한 유전자의 annotation 정보 또한 확인할 수 있도록 annotation page로의 하이퍼링크를 구성하였다.

나. 배추 표준 유전체에서 예측된 유전자의 annotation의 게시

(1) 배추 표준 유전체 annotation 페이지의 구성

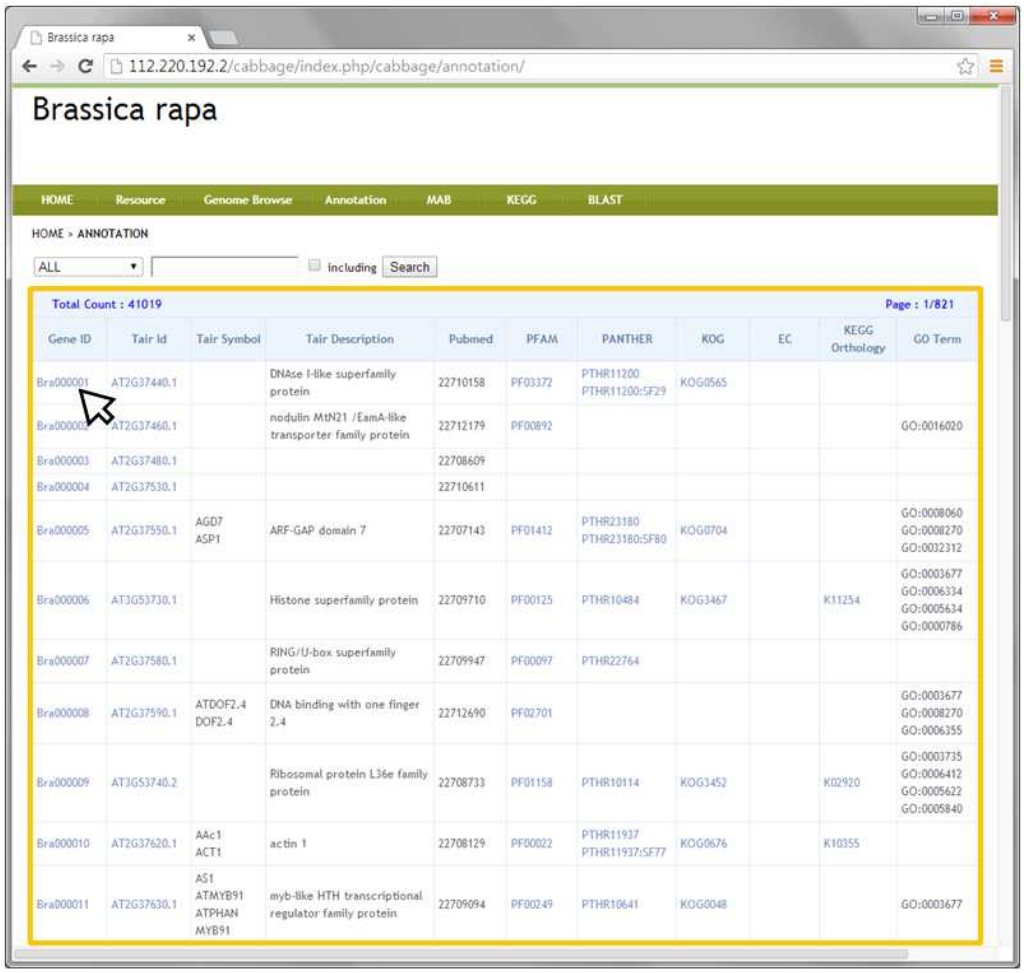


그림 2. Annotation 메뉴의 유전자 Table

배추 gene annotation 메뉴는 유전자 정보를 Table 형식으로 제공하며. 배추의 육종 특화 데이터베이스내에 별도로 구축한 genome browser에서도 이 annotation을 기준으로 선택한 유전자에 대한 정보를 열람 가능하다. 또한 annotation page의 table 상의 배추의 Gene ID를 클릭하면 genome browser로 연결되어 데이터베이스내의 두 플랫폼 사이의 상호 참조가 가능하다. 각 annotation된 ID를 클릭하면 해당 annotation이 제공된 DB의 web page로 연결되어 annotation의 원문 검색이 가능하다.

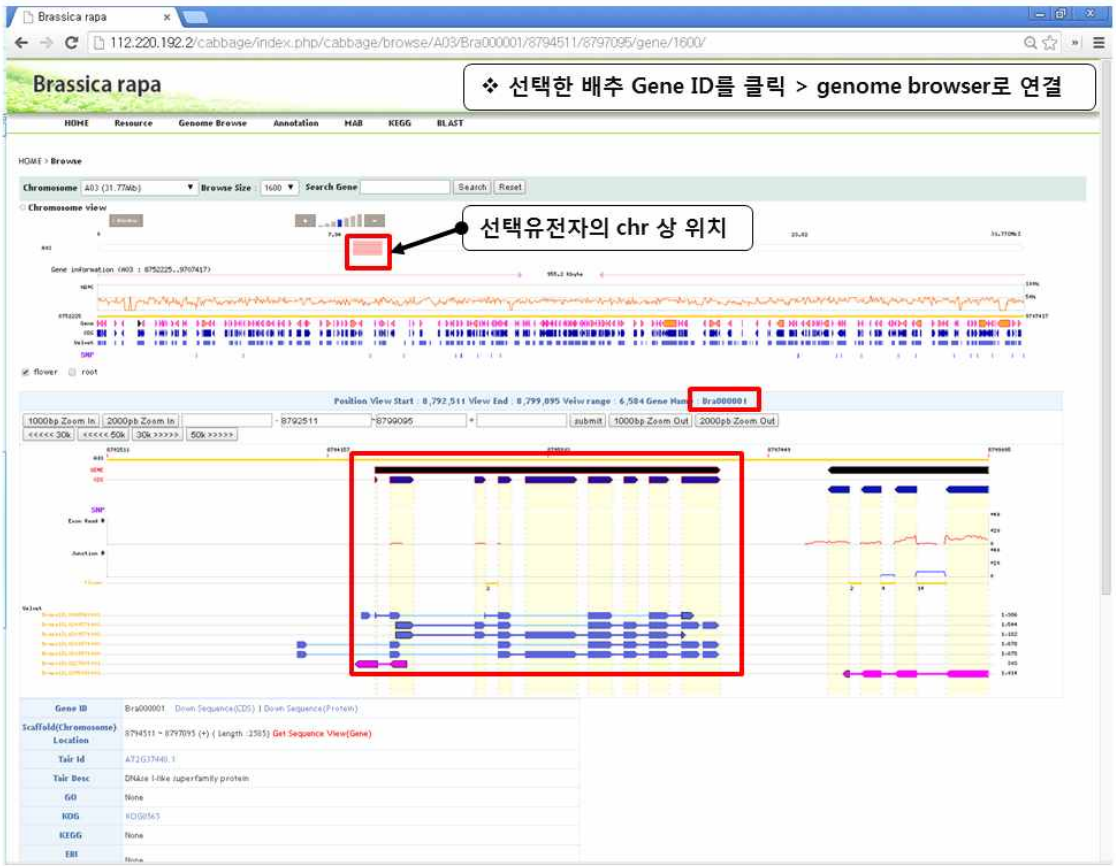


그림 3. Annotation 메뉴와 genome browser와 연동

다. 배추 표준 유전체 기반 Web-BLAST 시스템의 구현

(1) 배추 표준 유전체의 예측 유전자 서열을 활용한 BLAST 시스템

배추 유전자 서열의 유사도 검사를 위한 BLAST 시스템을 구축하였다. 배추의 표준 유전체 서열(draft genome)과 유전자(cds) 서열을 nucleotide, proteins 2가지 form으로 BLAST DB로 구축하여 사용자가 소유하고 있는 서열을 메뉴의 염기서열 입력란에 입력 후 분석을 진행하여 예측된 배추의 41,020개 유전자 중에서 가장 서열 유사도가 높은 유전자 ID 및 수반 정보가 출력되어 이를 이용할 수 있다.

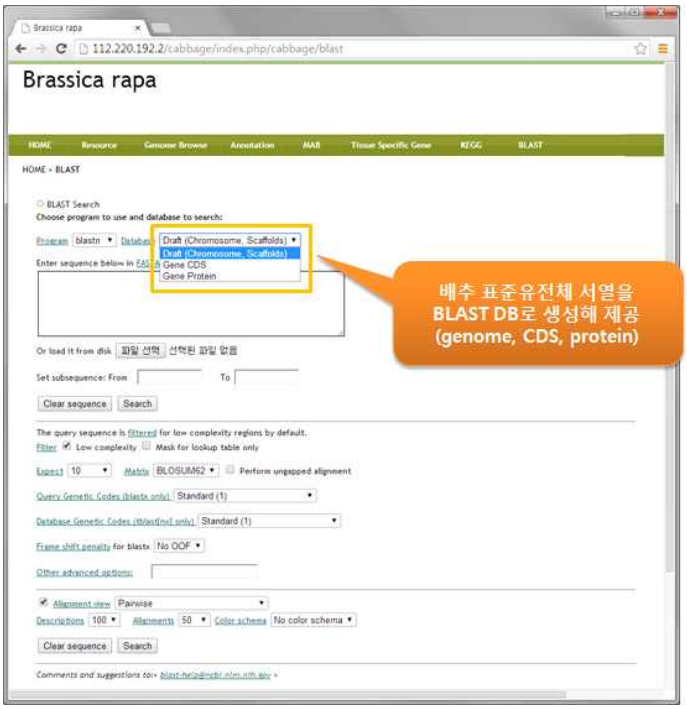


그림 4. BLAST 메뉴 구성

라. Genome Browser를 통한 배추 유전체 기반 생물정보의 통합적 시각화

(1) 배추 표준 유전체 기반 Genome Browser

배추 표준 유전체의 genome 서열, track data와 유전자 정보를 확인하여 genome browser로 구축하였다. 원하는 유전자 혹은 부위를 검색할 수 있도록 2가지 접근 방식이 가능하며, 첫 번째 방식은 chromosome을 선택하고 원하는 물리적 부위를 직접 클릭을 통하여 선택하는 방식이고 두 번째 방식은 유전자 search 창을 통해서 원하는 유전자를 검색하여 직접 접근할 수 있다.

Genome Browser 출력 섹션에서 사용자가 chromosome 상의 유전적 영역을 자유롭게 선택할 수 있고, Genome Browser page 접속 초기에는 chromosome 1번이 출력된다. 1번 염색체 전체의 길이를 확인할 수 있고, 한 화면에 보여 지는 영역이 1번 염색체 중 어느 위치인지를 분홍색 네모박스로 표시함으로써 현재 확인하고 있는 위치를 알 수 있다. 또한 사용자의 필요에 맞도록 browser size(해상도)를 조절할 수 있다.

그림 5. 배추 web DB의 Genome Browser 메뉴

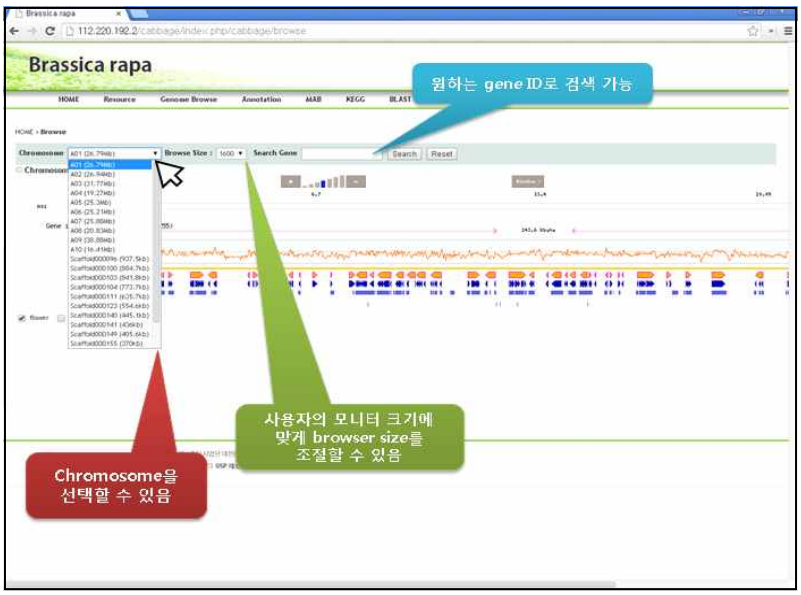


그림 6. 배추 web DB의 chromosome 및 scaffold. 선택 기능

Genome Browser는 사용자가 얻고자 하는 정보를 집약시킨 기본 인터페이스이며, 기본적으로 유전자의 배추 유전체상의 물리적 위치와 CDS, SNP marker 분포를 표시한다.

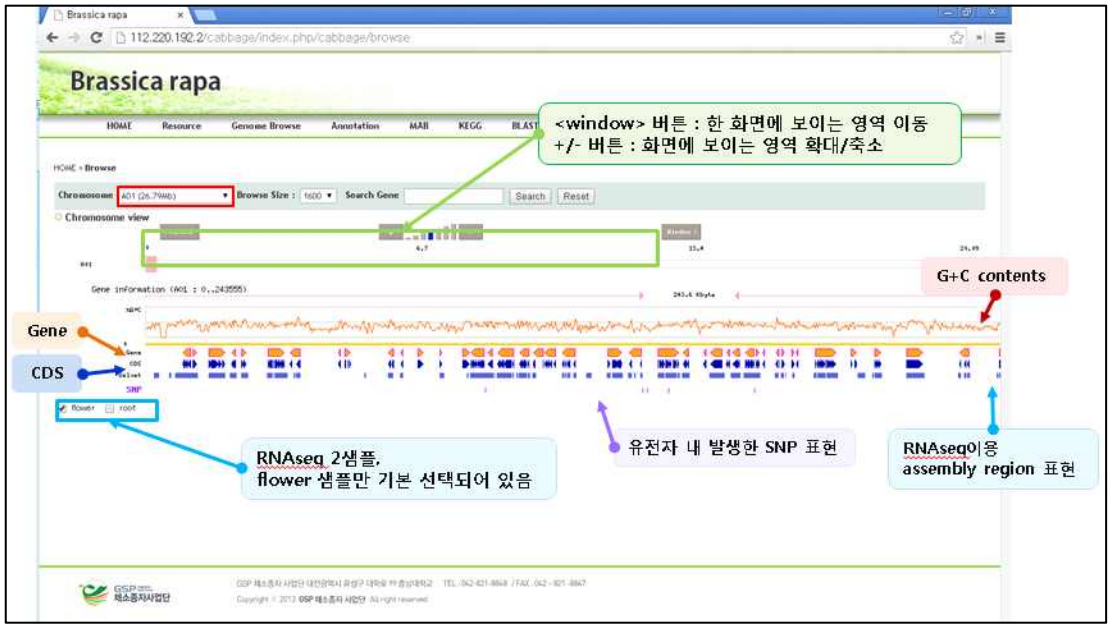


그림 7. 배추 web DB의 Genome Browser 구성

(가) Gene: 유전자의 boundary를 표시하며, 클릭 시 해당 유전자의 자세한 정보를 아래의 창에 확대하여 볼 수 있다. gene의 방향을 표기한다.

(나) CDS(Coding Gene Sequence): gene 내부의 coding sequence 영역 및 구성을 나타내고 유전자 발현의 방향 또한 나타낸다.

(다) SNP: 배추 RIL 집단의 re-sequencing 데이터를 이용하여 분석된 SNP의 배추 유전체상에서의 전체적인 분포를 나타낸다. 이를 통해 사용자가 원하는 영역 내의 SNP의 분포 또한 확인할 수 있으며 원하는 유전자 주변에 SNP의 존재 여부를 확인할 수 있다.

(라) RNA-seq assembly : 배추 RNAseq 데이터를 이용하여 수행한 assembly 결과 transcript를 표기한다.

(마) G+C conents : 유전체 전체의 G+C contents를 측정하여 꺾은선으로 표현한다.

(바) Window 버튼 : + 혹은 - 를 클릭함으로서 보고자하는 범위를 쉽게 조정(확대/축소)하여 설정할 수 있도록 한다.

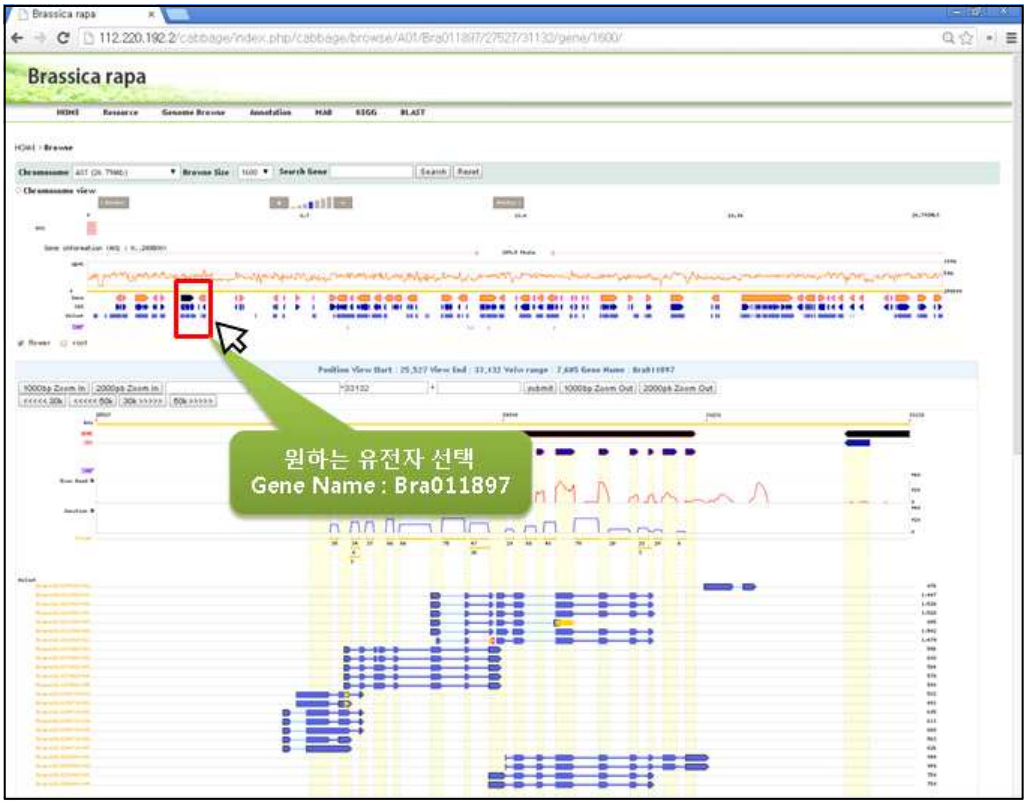


그림 8. Genome Browser 상의 배추 유전자 직접 선택을 통한 정보 인출 기능

마. Genome Brower에서의 배추 전사체 기반 데이터의 시각화

배추의 조직별 RNA-seq를 이용한 transcriptome assembly를 수행한 뒤, assembled transcripts를 표준 유전체에 mapping 하고, 그 결과를 그래픽 적으로 확인할 수 있도록 genome browser에 연결하여 구성하였다. genome browser의 gene model에서 화살표와 선으로 구성된 도형은 assembled transcripts를, 오른쪽 방향 도형은 forward 방향, 왼쪽 방향 도형은 reverse 방향을 의미한다.

RNA-seq 샘플별 assembly 결과를 그래픽 적으로 확인함으로써 발현량 결과와 함께 Alternative Splicing form 예측이 가능하다. 상단의 Zoom In/Out 버튼을 통해 브라우저의 해당영역을 더 자세히 혹은 넓게(간략히) 확인 할 수 있다. 1000bp는 좌우 500bp 씩, 2000bp는 좌우 1000bp씩을 의미한다. 또한 숫자 값을 직접 입력하고자 할 때는, 좌우 숫자 값 입력 후에 submit 버튼을 누르면 적용된다.

(1) RNA-seq에 의한 배추 조직별 발현량 데이터의 연동

Genome Browser 상에서 사용자가 원하는 유전자가 위치하는 것으로 예측된 유전적 영역을 선택 후 시각화된 track data 상의 유전자를 선택함으로써 해당 유전자에 대하여 생산된 조직별 유전자 발현량을 열람할 수 있다. RNA-seq 이후 조직별 발현량 데이터는 샘플별로 다르게 나타날 수 있는 sequence coverage 및 quality를 보정하기 위해 DESeq package을 통해 normalization이 된 발현량을 조직별 발현량의 상호 비교를 위해 채택하였다.

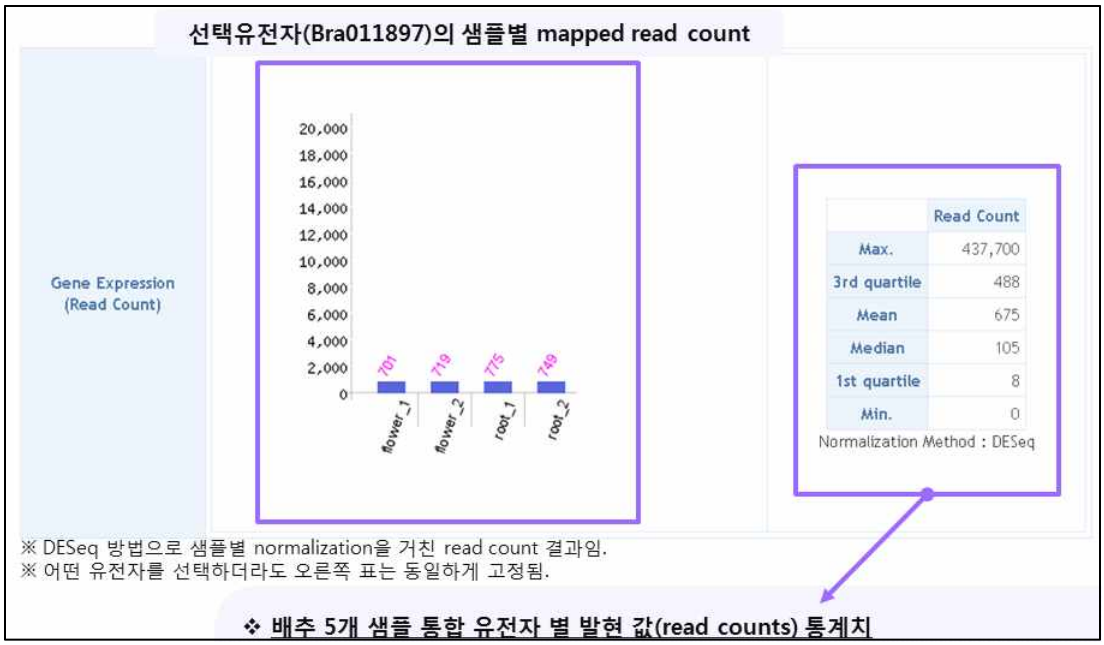


그림 9. 배추 web DB의 유전자 발현량 정보 제공

표 1. 유전자 발현량 통계정보의 내역

항 목	설 명
Max	최대(maximum) 발현 값 (즉, 어느 한 유전자가 437,700개의 mapped reads를 가짐)
3rd quartile	제 3사분위수 (발현 값을 정렬하였을 경우, 75% 위치에 해당하는 값)
Mean	평균값
Median	중앙값 (제 2사분위수와 동일 발현 값을 정렬하였을 경우, 50% 위치에 해당하는 값)
1st quartile	제 1사분위수 (발현 값을 정렬하였을 경우, 25% 위치에 해당하는 값)
Min	최소(minimum) 발현 값 (0값은 배추 유전자에 붙지 않는 reads가 없는 것을 의미)

(2) 조직별 RNA-seq의 alternative splicing의 예측 및 시각화

배추 RNA-seq를 이용한 transcriptome assembly를 수행한 뒤, assembled transcripts를 표준 유전체에 mapping 하고, 그 결과를 그래픽 적으로 확인할 수 있도록 genome browser에 연결하여 구성하였다. genome browser의 gene model에서 화살표와 선으로 구성된 도형은 assembled transcripts를, 오른쪽 방향 도형은 forward 방향, 왼쪽 방향 도형은 reverse 방향을 의미한다.

샘플별 assembly 결과를 시각화하고 이를 발현량 결과의 출력과 함께 유전자로부터 발현한 mRNA가 나타내는 Alternative Splicing form의 예측이 가능하다 (Trapnell et al, 2009). 상단의 Zoom In/Out 버튼을 통해 브라우저의 해당영역을 더 자세히 혹은 넓게(간략히) 확인 할 수 있다. 1000bp는 좌우 500bp 씩, 2000bp는 좌우 1000bp씩을 의미한다. 또한 숫자 값을 직접 입력하고자 할 때는, 좌우 숫자 값 입력 후에 submit 버튼을 누르면 적용된다.

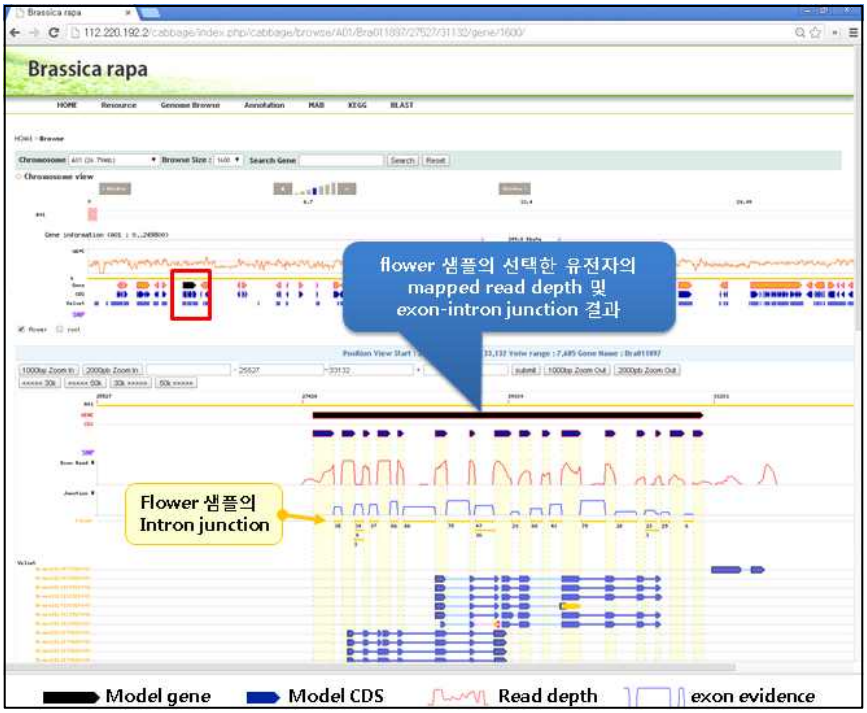
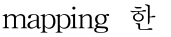


그림 10. 조직별 RNA-seq assembly 결과의 Alternative splicing의 시각화

그림 11. 배추 전사체의 transcript assembly 정보의 시각화

Genome Browser를 이용하여 assembled transript의 배추 표준 유전체상의 위치 정보를 확인 할 수 있다.

Assembled transcripts 표현 방법 : BLAT을 이용하여 reference genome에 assembled transcripts를 mapping 한 결과를 표현한다.  는 정방향의 assembled transcripts를,  는 역방향의 assembled transcripts를 표현한 것이다. Assembled transcripts의 CDS 조각 단위로 각각 genome에 mapping 되어 있다. 하나의 transcript 내의 CDS 조각들이 genome 내 CDS 와 차이(variation, position에 차이가 있는 경우 등)가 있을 경우,  와 같이 CDS 사이 연결선을 연한색으로 표현했으며, 그렇지 않다면  처럼 연결선을 CDS 조각과 동일한 색으로 표현하였다. 또한 같은 영역에 CDS가 다른 양상(position 차이, 방향차이 등)으로 mapping 된 경우,  노란색으로 표현하였다. Assembled transcripts의 CDS에 Start/end codon이 존재할 경우,  와 같이 검은색 테두리를 추가로 표현하였다.

사. Genome Brower 상의 배추 유전체의 sequence 관련 정보의 연동

(1) Genome Browser의 선택 유전자에 대한 통합적인 annotation 출력

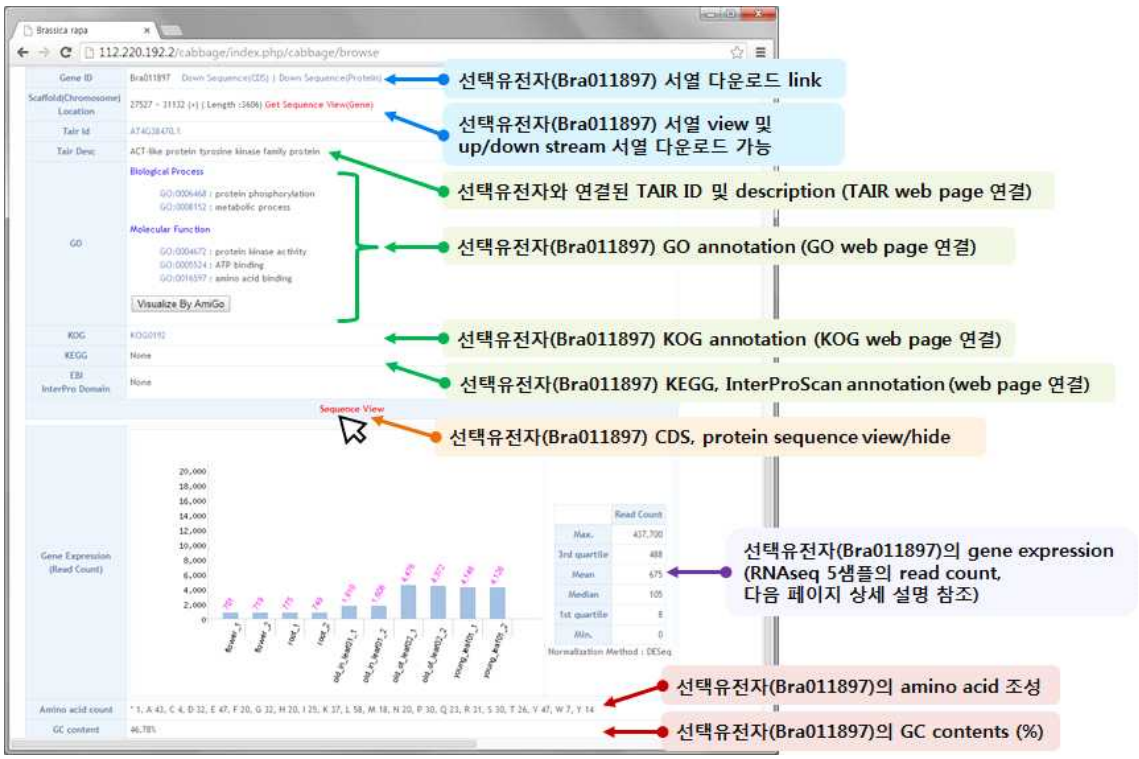


그림 12. 배추 Genome Browser의 gene annotation 정보 연결

Genome Browser의 annotation 항목 역시 현재, 공개된 annotation DB내의 해당 annotation ID에 대한 정보를 배추의 표준 유전자 정보와 연결시켰으며, 알려진 annotation DB로 TAIR, GO, KOG, KEGG, EBI InterProScan, pfam, PANTHER를 사용하였다.

(2) 표준 유전자의 서열정보 추출

Genome browser에서 배추의 유전체 상에 블록 형식으로 시각화된 유전자를 사용자가 직접 선택함으로써 선택 유전자의 nucleotide 혹은 amino acid의 조성과 G+C contents에 대한 측정치가 table의 형태로 제시된다. 또한 Genome browser는 선택된 유전자에 대한 사용자의 의도에 맞는 분석 및 재가공이 가능하도록 배추의 표준 유전자에 대한 annotation 정보 및 서열(CDS, protein)의 다운로드 기능을 제공한다. 사용자의 목적이 배추의 유전자 구조에 대한 확인을 넘어 유전자의 조절 영역의 확인일 경우를 상정하여 선택된 유전자의 구조적 서열 뿐 아니라 up/down stream의 원하는 길이만큼의 서열을 사용자가 원하는 범위만큼 지정하고 이에 대한 다운로드 기능을 제공한다.

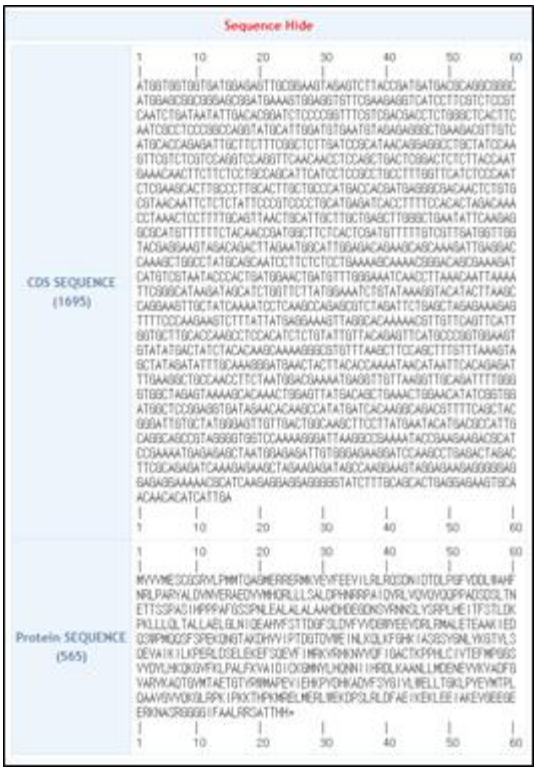


그림 13. Genome Browser에서 배추의 표준 유전자 서열 제공

(3) RIL 집단 유래 변이 데이터의 Genome Browser로의 연동

배추 육종을 목표로 하는 육종 주체가 특정 형질에 밀접하게 연관된 유전자를 식별하였을 때, 배추의 종자개발을 위한 육종 특화 데이터베이스에 구축한 Genome Browser를 통해 형질 연관 유전자의 배추 유전체 상에서의 물리적 위치를 확인할 수 있다. 또한 과제 3차 년도까지 작성한 배추 RIL 집단의 re-sequencing 결과를 활용하여 선택한 유전자 상의 변이를 시각화하는 기능을 구성하였다.

배추 Genome Browser 메뉴의 chromosome viewer에서 gene 또는 CDS를 선택한 후, 사용자가 viewer 상에 제시된 복수의 RIL 계통들을 선택할 수 있다. 이를 통해 선택된 유전자에 mapping 된 read들이 assembly된 서열을 viewer로부터 확인할 수 있으며 사용자가 관심을 가진 계통들 사이에서 특정 유전자상에서 발생한 SNP 변이가 집단 내에서 나타난 물리적 위치의 정보를 확인 할 수 있다.

3차 년도까지 수행한 작업을 기준으로 현재, RIL reference genome 정보를 포함하여 총 202 계통에 대한 SNP 변이 정보가 Genome Browser상에 반영되었다.

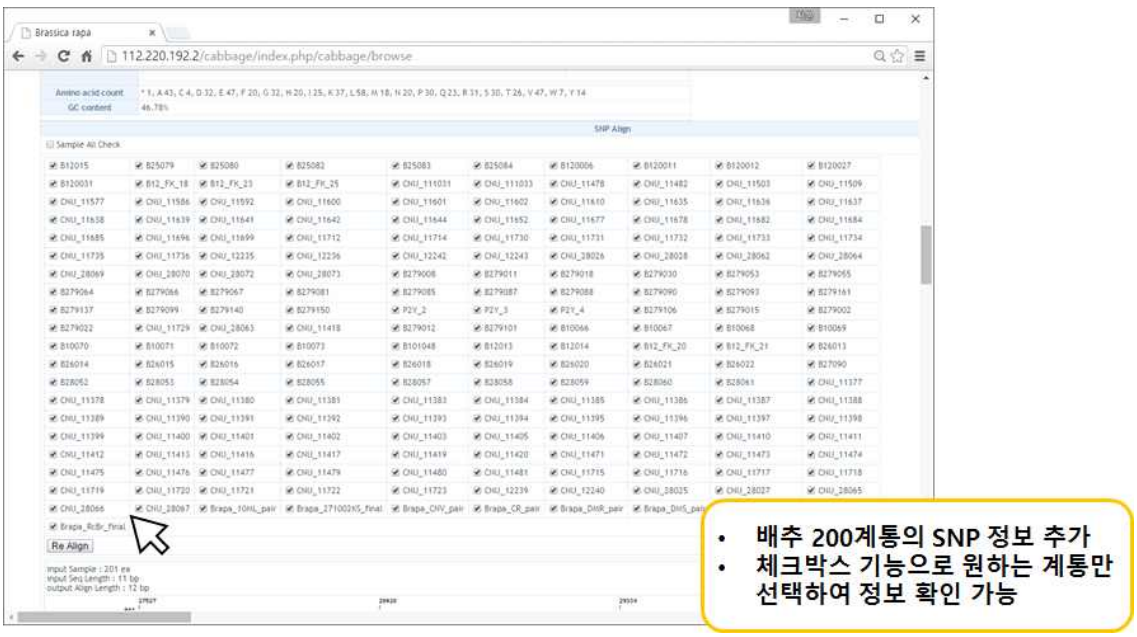


그림 14. Genome Browser에 반영된 배추 200여 계통

배추 유전자 상의 변이 정보의 출력 시 Genome Browser의 default 설정에 의해 배추 201 계통에 대한 변이 정보가 전부 출력된다. 사용자가 201 계통 전체가 아닌 일부 특정 계통들에 대한 변이 정보에 접근하는 경우를 상정하여 각 계통명 옆에 위치한 체크 박스에 대한 선택 여부를 설정할 수 있도록 하였다. 이를 통해 사용자가 원하는 계통들에 대한 변이 정보의 출력과 해당 sequence

상에서 변이가 나타난 위치를 시각적으로 다르게 표시함으로써 사용자가 계통별로 변이가 나타난 위치를 쉽게 식별할 수 있도록 하였다.

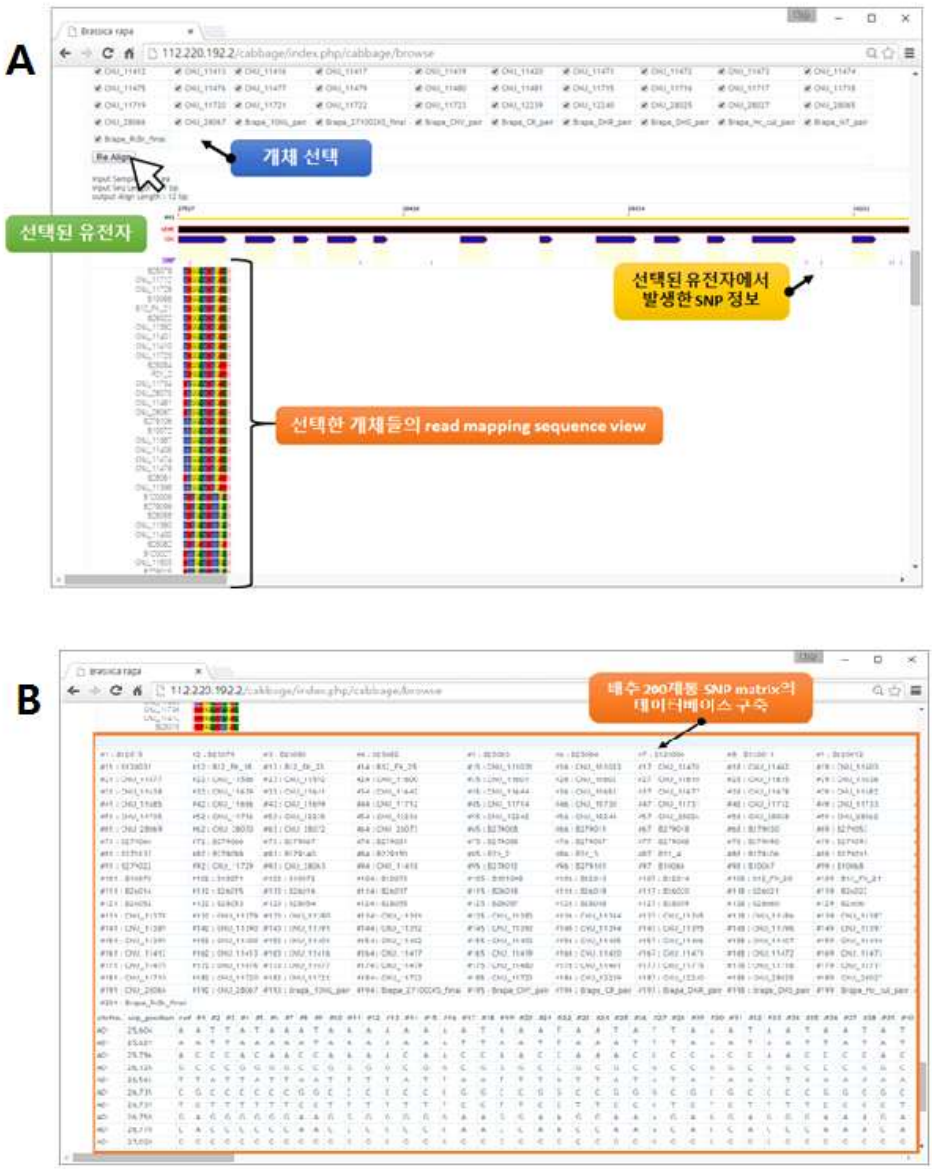


그림 15. A) 사용자 선택에 따른 제시된 201 계통내의 선별적 변이의 re-align 결과 B) 201 계통의 배추 유전자 영역 상의 SNP matrix

아. 배추의 조직별 DEG 정보와 KEGG annotation의 연동

Laccaria bicolor의 KEGG pathway map 분류를 따르도록 구성되었다. 배추 DEG 전체를 대상으로 phytozome annotation 수행 후, DAVID의 배추의 KEGG number 결과를 이용하였다. Description을 클릭하면 pathway 그림과 할당된 DEG 및 TAIR ID 정보 확인이 가능하다.

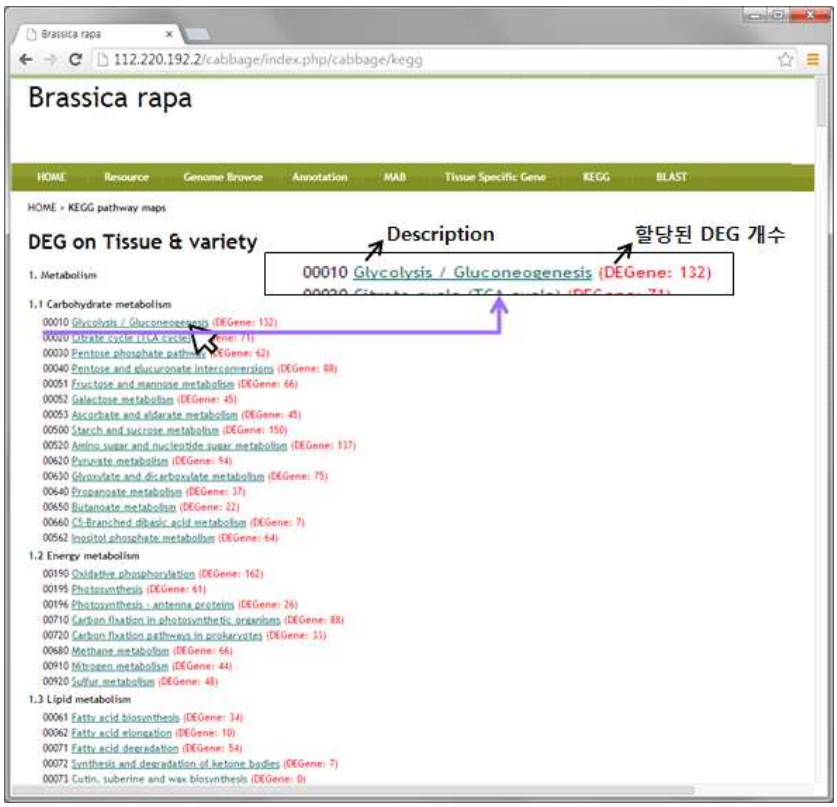


그림 16. DEG 데이터의 KEGG annotation 연동

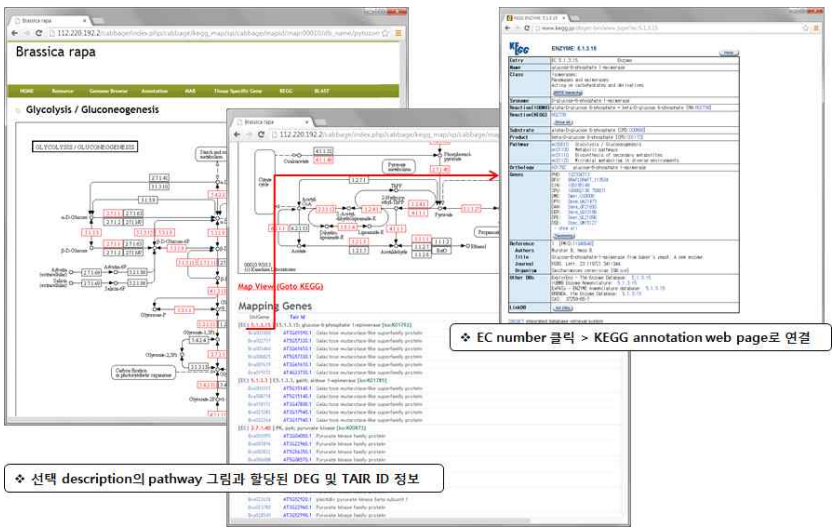


그림 17. KEGG 메뉴 내 KEGG pathway map 연결

자. 배추 RIL 집단의 genetic map DB의 구축

2차 년도 과제수행 과정에서는 1차 년도에 만들어진 배추의 종자개발을 위한 육종 특화 데이터베이스의 구성된 메뉴에 Bin Map이라는 새로운 메뉴를 추가하였다. Bin Map 메뉴 내에서는 데이터베이스내에 등재되어 있는 배추 RIL 집단을 이루는 계통들에 대한 목록이 제시되어 있으며 이용자가 열람하고자 하는 계통을 선택할 수 있다.

특정 RIL 계통을 genetic map viewer에서 선택함으로서 과제 수행과정에서 산출한 변이 데이터로부터 각 계통에서 chromosome 별로 부본과 모본 유래의 genomic partial을 서로 다른 색으로 시각화하여 (chiifu 유래: 붉은색, kenshin 유래: 파란색) RIL 집단 내의 계통들이 갖는 유전적 조성 정보를 확인 및 계통간 비교가 가능하다. 배추의 10개 염색체 중 원하는 염색체를 개별적으로 선택할 수 있으며 시각화된 염색체 도식의 길이는 'Height Size' 항목을 통하여 자유롭게 조절할 수 있다. 이용자가 정보 열람을 위한 setting을 완료한 후, search 버튼을 누르면 bin map (genetic map viewer)의 확인이 가능하다.

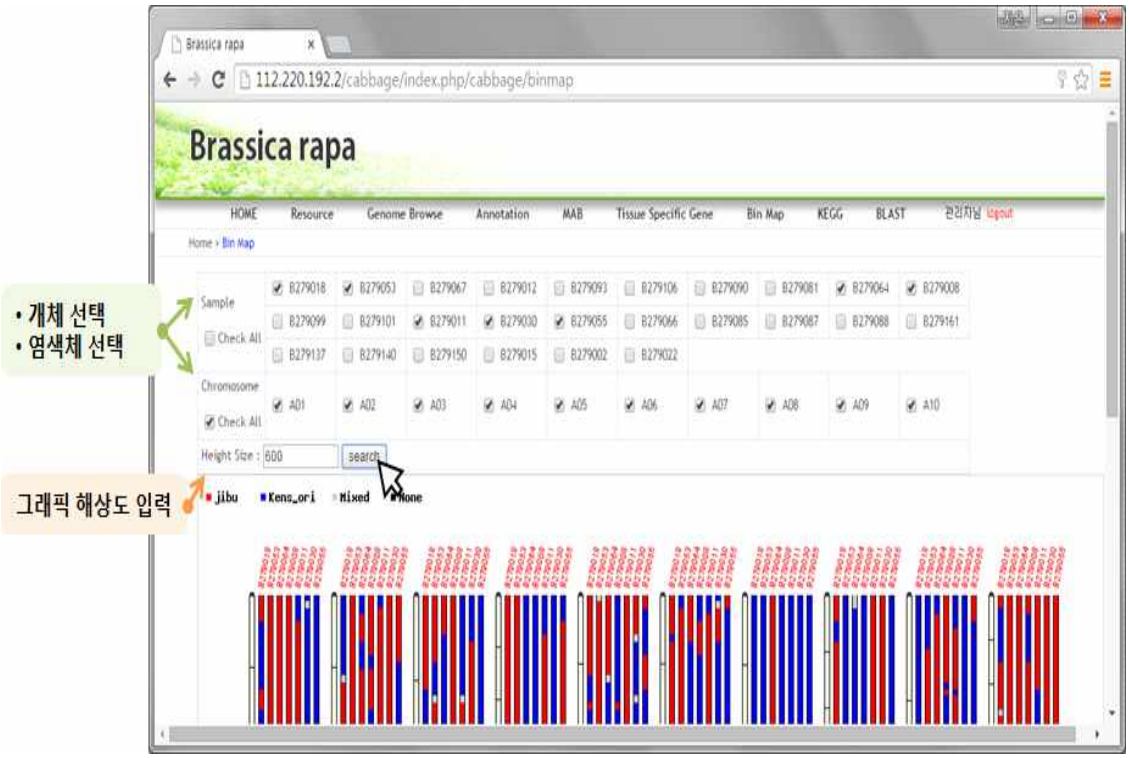


그림 18. 배추 RIL 집단의 부모본이 가진 변이를 중심으로 한 bin map의 시각화

차. RIL 집단의 Genome-wide SNPs를 이용한 여교잡 선발용 분자마커 (Marker Assisted Back-crossing) 데이터베이스의 구축

2차 년도에는 배추의 각 염색체를 물리적 길이로 동등하게 5등분된 시각화된 양식에 RIL 집단의 re-sequencing 데이터로부터 얻은 primer로 이용 가능한 SNP를 한 구획 당 3개의 SNP 마커를 선발하여 보여준다. 현재 배추의 종자개발을 위한 육종 특화 데이터베이스에 입력된 32개의 계통간의 교배조합을 선택하는 옵션을 제공하여 선택된 교배조합에 따라 자동으로 적용 가능한 SNP의 염색체 별 분포를 나타낼 수 있다.

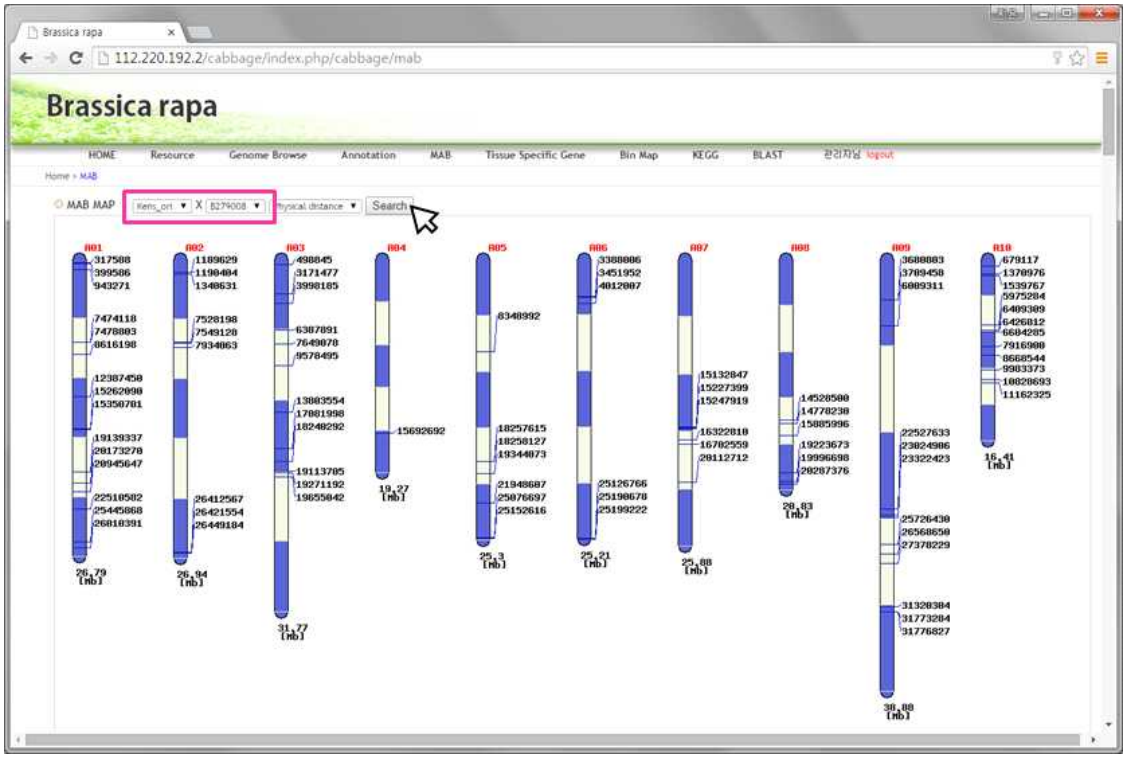


그림 19. Kenshin_ori와 B279008 교배조합 간 이용할 수 있는 데이터베이스 상에서 예측된 SNP 마커의 물리지도 상의 위치 정보를 시각화

카. 목표형질 선발용 분자마커(MAS) 데이터베이스 구축

1차 년도에는 기 보고된 뿌리혹병에 관련된 분자마커 정보를 수집하였다. 또한 이외에 유용형질 관련 분자마커와 환경스트레스에 관련된 분자마커 정보를 수집하였다. 또한 2차 년도에는 문헌상의 TuMV, 노균병 등 내병성 관련 형질 분자마커 정보를 수집하고 웅성불임, 자가불화합성, 키, 종피색과 같은 배추 유용형질 관련 분자마커 정보를 수집하였으며 3차 년도에는 수집한 마커가 지정한 유전적 영역의 위치의 배추 수집단 내에서의 SNP 정보를 확인하고 이를 SNP 범용 마커로 전환하는 시도를 하여 잎털, 결각, GMO 검정을 위한 분자마커를 개발하였다. 현재 수집한 분자마커의 정보는 table의 형태로 배추의 종자개발을 위한 육종 특화 데이터베이스의 MAS 페이지 내에 정리 되어있으며, 각 분자마커의 배추 표준 유전체 상의 물리적 위치 또한 시각화되어 제공되고 있다.

타. 형질 관련 마커 예측을 위한 eQTL 분석 시스템 구축

과제수행 과정에서 구축된 배추의 종자개발을 위한 육종 특화 데이터베이스의 Tissue Specific Gene 메뉴를 선택하면 사용자가 선택한 유전자가 나타내는 특정 조직별로의 발현량을 확인할 수 있다. 현재는 과제 1차 년도부터 생산한 5개의 조직에서 특이적으로 발현되는 transcriptome 정보가 제공되고 있다. 배추의 조직별로 배추 표준 유전자에 mapping된 read의 수 및 조직 특이적으로 발현하는 유전자들의 목록을 확인할 수 있으며 이는 선택한 배추의 조직에서 특이적으로 발현하는 유전자들을 대상으로 발현량의 내림차순으로 출력된다. 각각의 유전자들의 기능 정보는 Arabidopsis 데이터베이스(TAIR)에서 얻은 애기장대 유전자의 annotation 정보와 연동되어 있다.

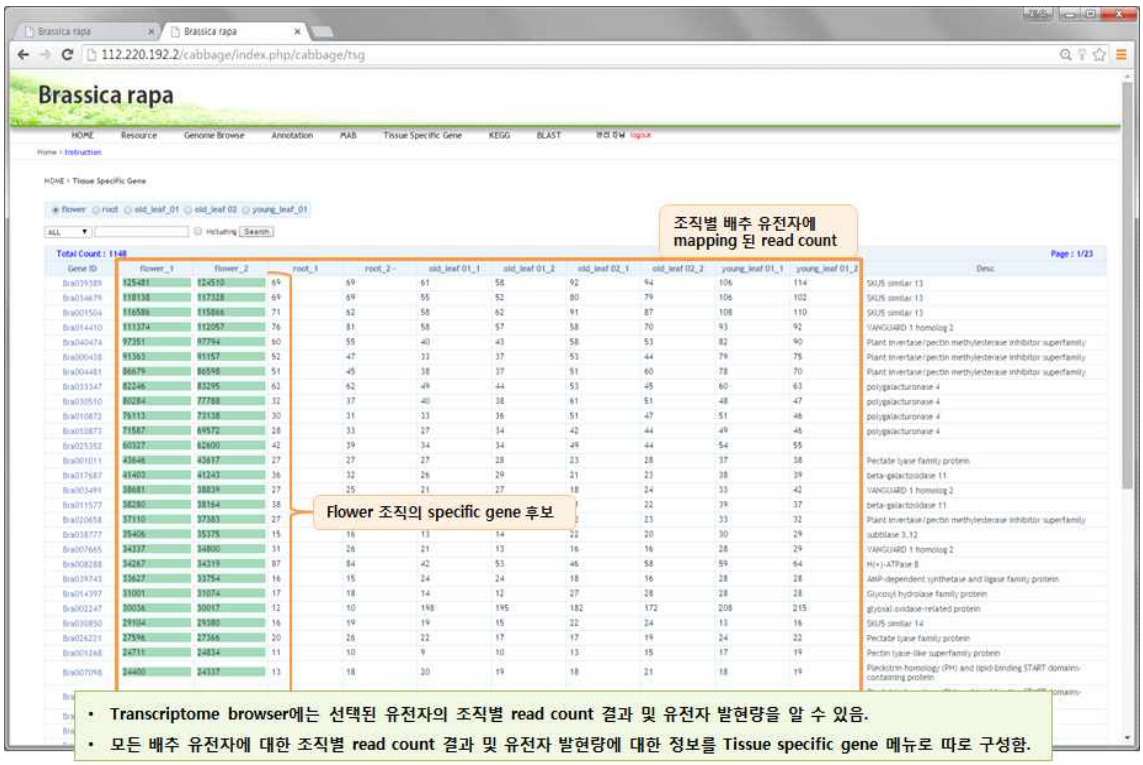


그림 20. 배추 화기(flower) 조직에서 tissue specific gene의 목록과 발현량

파. Linkage Disequilibrium(LD)을 이용한 배추 RIL 집단의 유전체 재조합 정보 분석 및 Haplotype 정보의 도식화

Re-sequencing을 통해 생산한 배추 9 계통의 read 데이터를 reference genome에 mapping한 후 SNP 정보를 추출하였다. 추출한 데이터는 reference sequence를 포함한 배추 10 계통에 해당하는 SNP matrix를 생성하였다.

하나의 SNP가 두 개의 계통 이상에서 reference genome 상에 mapping된 계통의 read가 없을 때 나타나는 'N'을 가질 경우, 해당되는 SNP를 SNP matirx에서 제거하였다. Filtering이 완료된 SNP matrix 상에서 물리적 길이 1000bp당 1-2개의 SNP를 선택하여 genome-wide SNP physical position 정보를 구성하였으며 배추 유전체 전체에 편중된 위치가 없이 고르게 분포하는 SNP matrix를 새로 구성하였다. 완성된 SNP 데이터를 LD분석 프로그램인 Haploview (Barrett et al, 2005)에 입력하여 LD를 계산하고 LD block을 생성하였다. 또한 LD 패턴을 시각화하여 그림파일로 추출하였다. 또한 배추 염색체별 LD 정보를 genome browser에 연결하였다.

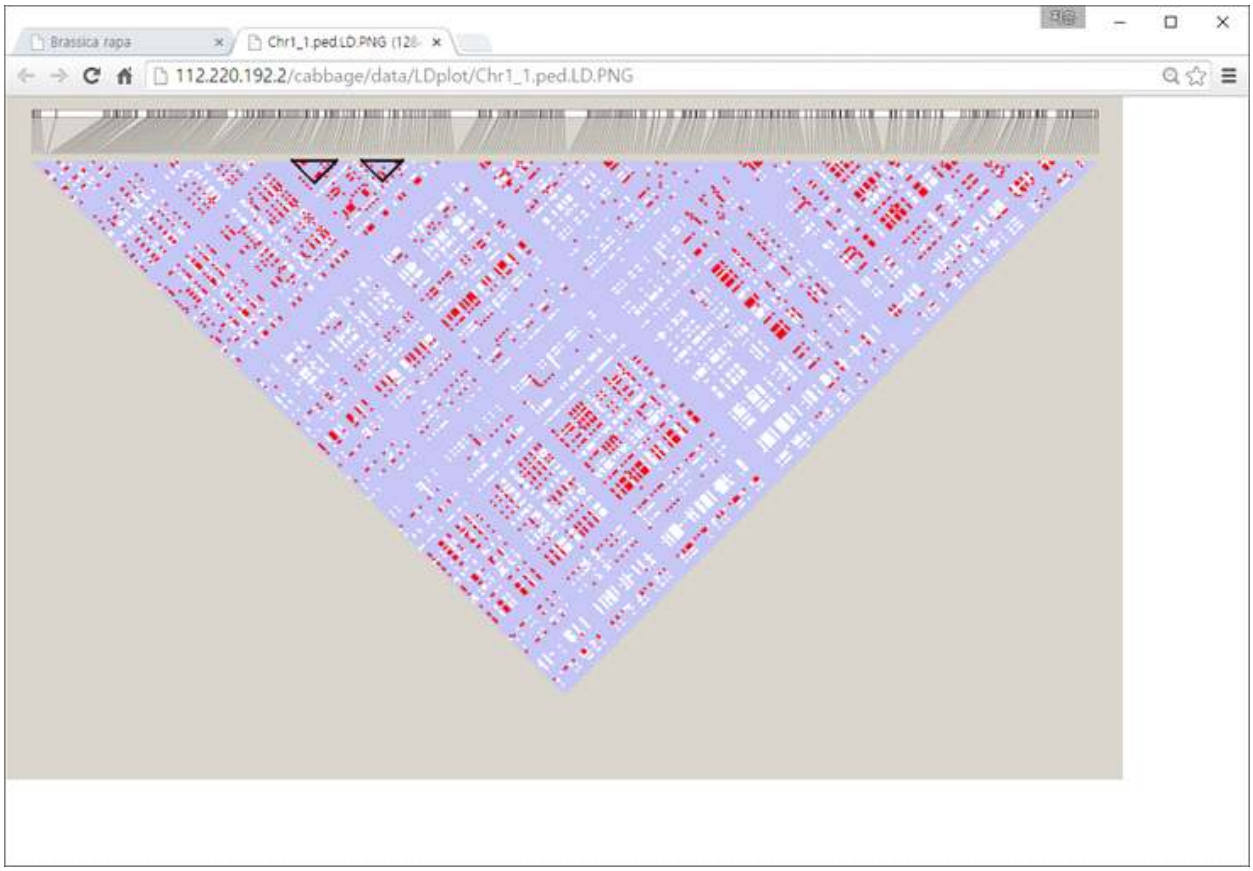


그림 21. 배추의 10개 계통에서 1번 chromosome에서 나타난 LD block

하. 5대 채소작물의 분자육종 활성화를 위한 데이터베이스의 구축

과제 3차 년도까지 구축한 배추의 육종 특화 데이터베이스를 기준으로 다른 4개 작물의 데이터베이스를 구축하였다. 향후 과제에서 각 작물의 유전체 및 육종 관련 데이터를 수집 및 생산하고 이에 입력할 계획이다.

그림 22. 5대 채소 작물의 육종 특화 데이터베이스의 링크 페이지

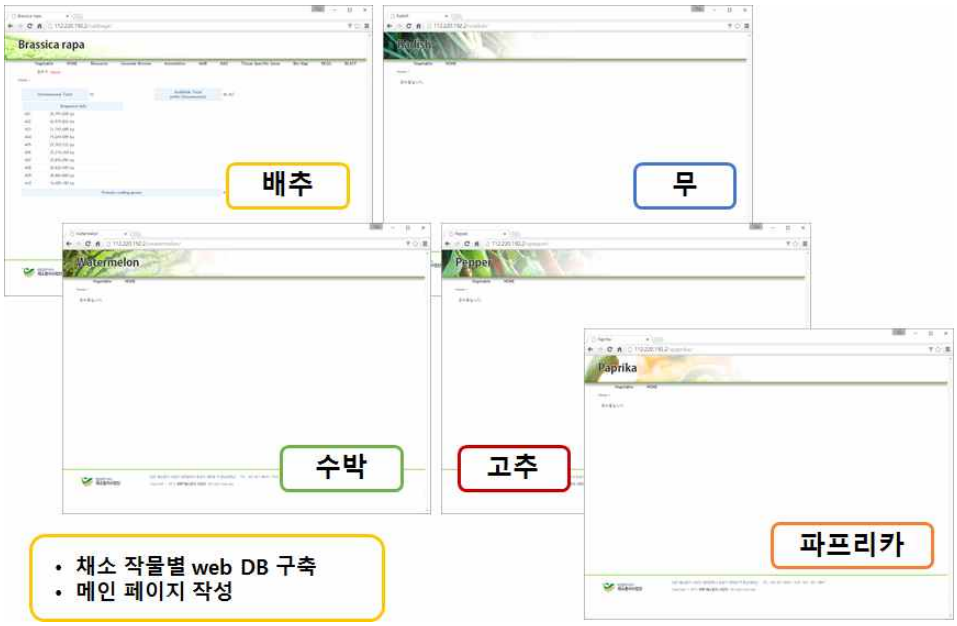


그림 23. 배추를 모델로 구성된 5대 채소 작물의 육종 특화 데이터베이스

제4절 배추의 유용 유전자 발현량 데이터베이스 (BrTED)의 구축과 운영

1. 배추 전사체 정보 분석 특화 데이터베이스의 구축 동기

배추(*Brassica rapa*)는 전 세계적으로 유통되고 있는 경제 작물로서 특히 한국을 중심으로 동아시아에서 높은 수요와 거대한 시장 규모를 나타내고 있다. 배추는 시장에서의 중요성 뿐 만이 아니라 다른 배추과에 속하는 여러 작물들과 비교하였을 때 2배체임에도 불구하고 종내에서 보이는 표현형적 다양성과 상대적으로 작은 유전체의 크기로 인하여 배추과 A genome 작물들의 유전연구 모델로서 활용되고 있다(Li et al, 2010). 또한 식물 연구의 모델 식물로 이용되고 있는 애기장대(*Arabidopsis thaliana*)와 유전자의 기능 및 구조에 높은 유사성을 보이기 때문에(Schranz et al, 2006) 이미 발표된 문헌상의 정보를 활용하여 유전자 구조와 표현형과의 관계를 쉽게 연구할 수 있는 이점을 가지고 있다. 따라서 농업적으로 유용하게 이용 가능한 배추의 다양한 형질들에 대하여 연관된 유전자들의 발현 수준에 대한 microarray 또는 RNA-seq을 중심으로 한 연구들이 진행되었다.

배추 전사체 연구에 사용된 발현량 데이터는 NCBI와 같은 거대 생물정보 데이터베이스에서 다운로드 및 재가공이 가능하다(Barrett et al, 2013). 그러나, 이에 필요한 system resource와 전문 분석 기술이 수반되기 때문에 일반적인 연구자들이 필요한 데이터를 확보하고 이를 가공하여 정보를 재생산하는 것에는 많은 어려움이 있다. 또한, 지난 과제 수행과정에서 구축한 배추의 분자육종 활성화를 위한 데이터베이스의 경우에는 201 계통의 re-sequencing 정보를 통해 배추의 genome 상의 변이를 육종가가 이용하는 데 초점이 맞추어져 있었다. 그로 인해 배추 유용 유전자 식별을 이용된 배추의 전사체 데이터는 5개 샘플의 RNA-seq으로부터 구성되어 다양한 조건하에서의 특이적 유전자를 식별하기 위한 data pool이 크게 부족하다는 한계가 있다. 이를 극복하기 위해 과제 3차 년도부터 진행한 배추 전사체 데이터베이스의 구축으로부터 전 세계의 연구자들에 의해 생산된 다양한 조건하의 배추 전사체 데이터에 대한 접근성을 높이고 자유롭게 분석할 수 있는 BrTED (Brassica Transcriptome Expression Database: 배추 전사체 특화 웹데이터베이스)를 구성하였다.

총 10개의 공개된 실험들로부터 다양한 조건을 반영한 92개의 전사체 데이터를 수집하여 조건별 유전자 발현량을 산출하고 배추의 41,020개의 유전자에 annotation을 진행하였다. 또한 산출된 DEG를 분석하는 system을 web-based 데이터베이스 플랫폼 내에 구축하였다. 이로서, 연구자가 특정 조건의 조직 내에서 발생하고 있는 생물학적 사건들을 쉽게 파악할 수 있는 것을 목표로 구축되었다. BrTED는 새롭게 생산된 발현량 데이터를 지속적으로 수집하여 배추의 형질 및 기능 유전체 연구에 유용한 tool이 될 것으로 예상된다. 이하는 3차 년도 과제수행 기간부터 구성한 BrTED에 대한 구축 과정 및 결과들을 기술하였다.

2. BrTED의 구성에 필요한 기본 정보의 생산

가. 외부 1차 데이터베이스에서의 배추 전사체 정보의 수집

배추의 전사체 데이터베이스인 BrTED의 구축을 위해서 1차적으로 필요한 작업은 어떠한 데이터가 필요한지를 선별했고 해당 데이터의 출처를 확인했다. 1차적인 검색으로 얻을 수 없는 경우에는 확보한 데이터를 재가공하였다.

(1) 배추 전사체 데이터의 수집

BrTED의 실질적 데이터인 배추(*B.rapa*) 한정으로 다양한 조건하에서의 유전자의 발현량을 얻기 위하여 전세계의 연구팀이 각자 생산한 sequencing 정보 및 Microarray 결과를 upload 하는 NCBI GEO (www.ncbi.nlm.nih.gov/geo/)에 접속하였다.

NCBI GEO에서 배추의 학명인 *Brssica rapa*를 키워드로 검색하고 RNA-seq 정보 및 microarray 데이터를 확보하여 BrTED의 구축에 활용한 NCBI상의 정보는 표 1과 같다. (EMBL에서 얻은 microarray dataset 1개 포함)

표 1. BrTED의 expression profiling으로서 이용 가능한 실험 및 sample의 현황

Method	Series ID	Title	Sample	Treatment
RNA-seq	GSE43245	Transcriptome sequencing of Brassica rapa tissues	8	Differential expressed gene
	GSE51363	Global Analysis of the Transcriptional Response of Chinese cabbage (*Brassica rapa* ssp. pekinensis) to Methyl Jasmonate Reveals JA Signaling on Enhancement of Secondary Metabolism Pathways	2	Jasmonic acid
	GSE58895	Elucidation of stress memory inheritance through epigenome alterations in *Brassica rapa* plants [RNA-seq]	28	Heat stress
	GSE69785	Transcriptome profiling of the defense response following elicitation by a bacterial pathogen or flg22	12	Pathogen Infection
	GSE74044	Transcriptome analysis of Brassica rapa near-isogenic lines carrying clubroot-resistant and -susceptible alleles in response to Plasmodiophorabrassicae during early infection	8	Pathogen Infection
	GSE75464	Physiological characterization and comparative transcriptome analysis of a developmentally retarded Chinese cabbage (*Brassica campestris* ssp. *pekinensis*) mutant	6	Radiation derived DH
	GSE77427	Comparative transcriptome of the fertile and sterile buds of genic multiple-allele inherited male sterile AB line in Chinese cabbage	2	Male sterility
Microarray	GSE23409	Comparison of root and leaf samples from *Brassica rapa*	6	Differential expressed gene between tissues
	GSE47665	Comprehensive analysis of genic male sterility-related genes in *Brassica rapa* using *Brassica rapa* 300k microarray v2.0	14	Male sterility
	E-MTAB-2386	Transcription profiling by array of 10 days old *Brassica rapa* ssp. chinensis seedlings treated with 2mM methyl jasmonate by spraying and harvesting 48 hours past treatment	6	Jasmonic acid

(2) 수집한 배추 전사체 데이터의 발현량 정보 재생산

NCBI GEO의 RNA-seq 정보를 저장한 FTP server는 연구자들에 의해 업로드된 sequence read 데이터를 sra 포맷으로 저장하고 있다. 이를 자체 보유하고 있는 분석 서버에 샘플이 포함된 실험별로 디렉토리를 구성하였다. 저장된 sra포맷의 데이터는 sra toolkit의 fastq-dump 명령어를 이용하여 fastq 포맷으로 변환하였다. 변환된 RNA-seq의 read 데이터는 다음과 같은 절차를 거쳐 특정 조건하의 전사체 데이터로서 활용될 수 있다 (그림 1).

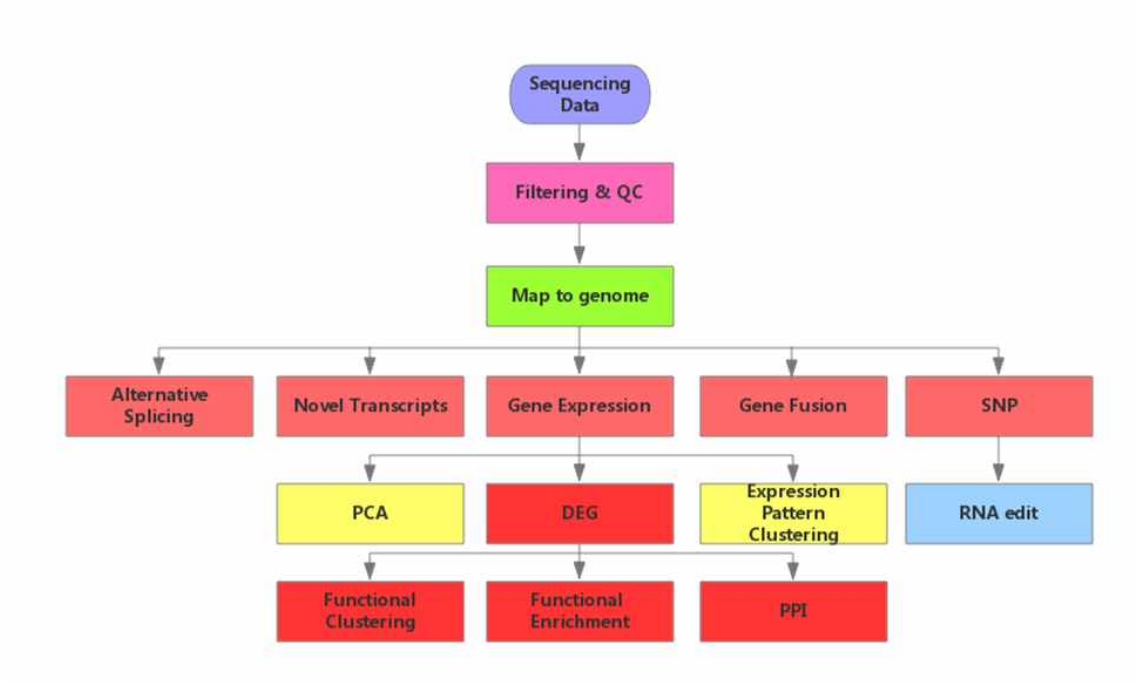


그림 1. 일반적인 RNA-seq 데이터의 처리 절차(pipeline)와 기대 가능 결과(output)

(가) RNA-seq read 데이터의 Quality Check(QC)와 Quality Trimming(QT)

mRNA를 cDNA화하여 sequencing을 수행한 RNA-seq 결과는 대략 75~150bp 수준의 길이를 갖는다. 현재 RNA-seq 분석을 위해 상용 중인 illumina의 sequencer에 의해 산출한 데이터의 경우 read 내의 bp 순서가 후반으로 갈수록 염기서열 결정에 대한 오류 확률이 높아지게 된다.

이러한 오류 확률이 높은 read에 대해 reference mapping에 사용하게 될 경우 잘못된 alignment 결과를 가져올 수 있기 때문에 alignment 이전에 QC 및 QT를 수행하게 된다. QC는 fastqc 소프트웨어를 이용하여 수행되었으며 이를 통해 read의 전반적인 오류 발생 패턴을 확인하고 read data pool에서 adapter 서열로 나타나는 overrepresented sequence를 식별하였다.

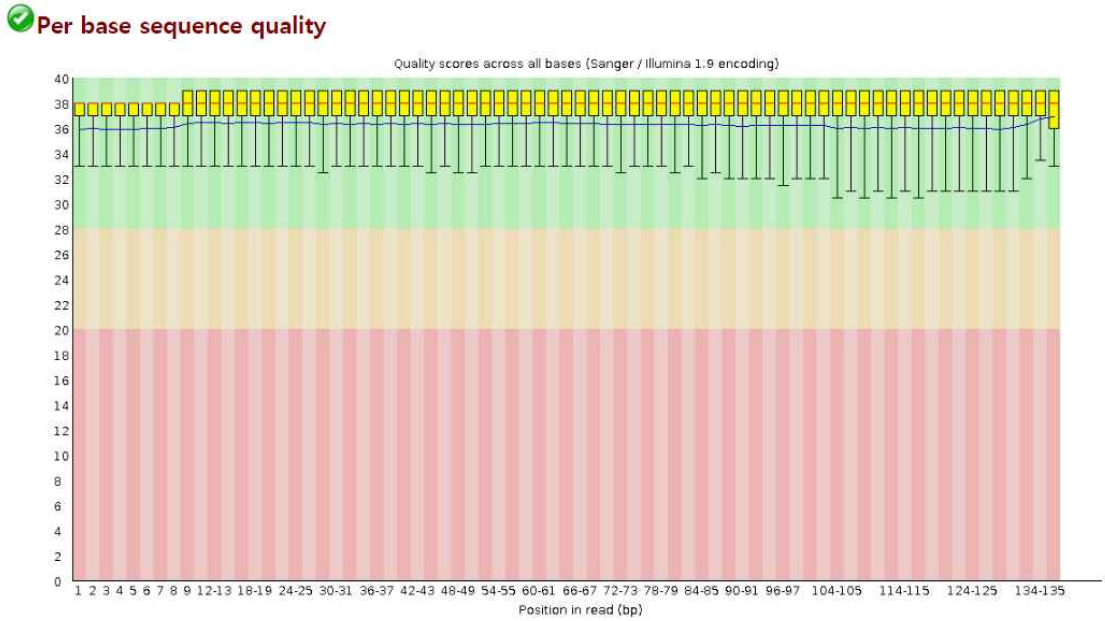


그림 2. fastqc를 통한 배추 전사체 sample의 read quality의 시각화

read의 QT는 fastx toolkit의 fastx_trimmer로 phred score 30 이상, trimming 이후 최소 read length 20 이상을 기준으로 수행되었다. fastqc 과정에서 overrepresented sequence로 나타난 adapter 서열은 cutadapt 소프트웨어를 통해 제거하였다. overrepresented sequence로 나타난 서열은 명령어상에 입력되고 trimming 이후 최소 read length가 20 이상인 것을 취하는 것을 조건으로 작업을 수행하였다. QT 이후 다시 fastqc를 통해 전사체 샘플내의 read data pool의 오류 상황을 재확인 하고 문제가 없을 경우 trimmed read를 이용한 reference mapping을 수행하였으며 QC이후 문제가 다시 발견된 경우 QT에 이용되는 명령어의 parameter에 반영되는 조건을 엄격하게 상향하여 QT를 반복 수행한 이후 QC에서 문제점이 나타나지 않았을 때 다음 단계를 수행하였다.

(나) tophat을 이용한 read의 reference genome으로의 alignment

QT가 완료된 전사체 샘플의 reference 활용 alignment를 수행하기 위한 tool로서 tophat 소프트웨어 (Trapnell et al, 2012)를 선정하였다. tophat은 reference genome(BRAD 1.2 version 기준)의 염기 서열 정보를 fasta 포맷과 서열상의 유전자의 위치 정보를 담은 gtf 포맷으로부터 입력받으며 bowtie2 소프트웨어의 지원을 받아 read 데이터를 reference 상에 mapping한다 (그림 3).

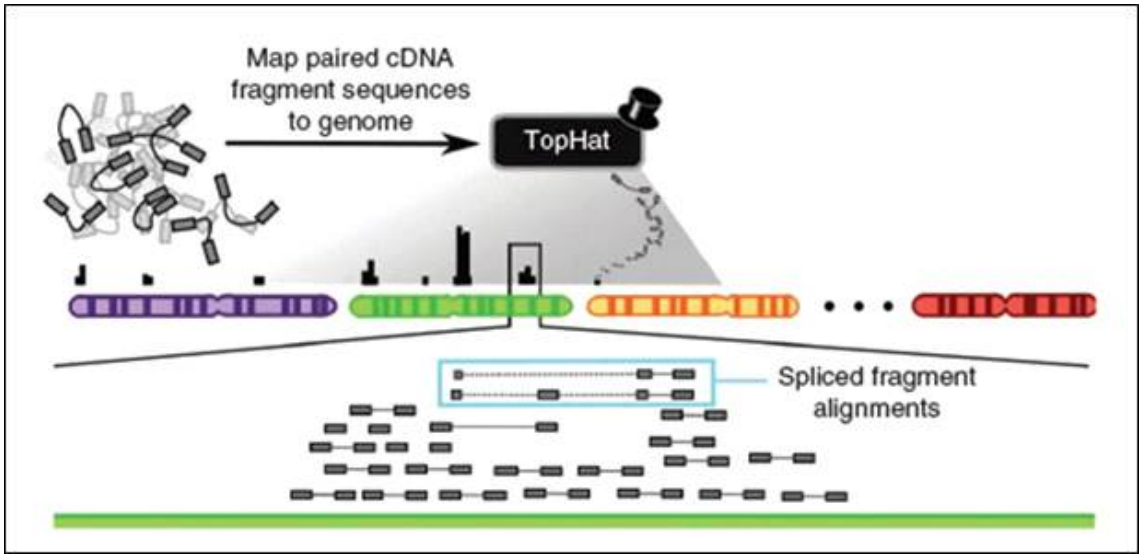


그림 3. tophat을 통한 전사체 sample의 read data의 reference genome으로의 mapping 및 splicing site의 식별

tophat을 통한 read data의 mapping 과정에서 splicing site의 위치 정보가 부가적으로 생산된다. 이 splicing site를 취합하여 unmapped read를 다시 mapping 하여 평균적으로 unmapped read의 0.1 % 수준의 read를 배추 유전체 내의 유전자 상에 추가적으로 mapping 할 수 있다. read의 mapping 결과는 binary 파일인 bam file로 명령어에서 지정한 경로에 저장되며 전사체 정보를 구성하는 실험의 조건별 샘플에 대한 유전자별 발현량 (read count 형식)의 산출과 normalization에 사용된다.

(다) DESeq 파이프라인을 통한 발현량의 산출과 normalization

DESeq manual (www.bioconductor.org/packages/devel/bioc/vignettes/DESeq.pdf)에서 제공하는 파이프라인을 통해 수집한 7개의 실험별로 RNA-seq read의 mapping 결과를 R 환경에 load 하여 배추의 표준 유전체 1.2 version 상에서 예측된 유전자의 reference 상의 위치에 mapping된 read의 수를 세어 예측된 배추 유전자에 대한 개별적인 발현량을 산출하였다.

실험별 발현량 데이터는 동일 실험내에 존재하는 전사체 발현량 양상의 비교를 위해 DESeq에서 제공하는 함수 estimateSizeFactors를 사용하여 실험별로 얻은 1차적 발현량 matirx의 normalization된 발현량 matrix를 얻을 수 있었다. 또한 sequencing coverage와 mapping quality가 서로 다른 실험 간의 발현량을 서로 비교하기 위해 TPM normalization을 현재 보유한 RNA-seq 샘플 66개에 대해 수행하여(Wagner et al, 2012) TPM normalized matrix를 얻었다.

나. 배추 표준 유전체상의 예측된 유전자에 대한 annotation 작업

BrTED의 구성에서 특정한 배추 표준 유전자를 검색하여 그에 대한 annotation을 출력하고자 할 경우나 실험내 혹은 서로 다른 실험사이에서 RNA-seq 샘플을 두 개 이상 선택하여 전사체의 발현양상을 비교하여 DEG를 산출하였다. DEG의 기본 정보를 확인하려 할 때, 배추의 표준 유전자에 대한 다양한 각도의 annotation을 수행한 결과가 있어야 선택 및 출력된 유전자에 대하여 사용자에게 최대한의 정보를 전달할 수 있다.

배추의 표준 유전자 41,020개에 대한 통합적인 annotation 작업을 통하여 이를 BrTED에서 제공하는 배추의 표준 유전자에 부여할 수 있었다 (표 2).

표 2. BrTED의 *Brassica rapa* 표준 유전자 41,020개에 대한 annotation 현황

Annotation Class	Unique Annotation	*B. rapa* gene ID with Annotation
Basic information of *B. rapa* genes (Chromosome, Location, Sequence)	41,020	41,020
Uniprot ID	24,790	37,937
TAIR ID	20,486	37,847
Functional description	36,673	37,842
Pfam ID	3,975	31,459
GO number	3,981	28,561
EC number	1,085	10,985
miRNA family	496	8,280
KEGG Pathway	132	7,836

(1) 배추 표준 유전자의 기본 정보

BRAD(brassicadb.org/brad/)에 공개된 배추의 표준 유전체 1.2 version에 대한 관련 정보를 BRAD에서 제공하는 FTP server로부터 다운로드 받아 배추의 표준 유전자별 정보를 구성하는데 활용하였다. 구성 정보로서 유전자 ID, 유전자의 염기서열. 단백질 서열 그리고 유전체 상의 물리적 위치를 얻을 수 있다.

(2) 외부 데이터베이스 유래 정보

배추 표준 유전체의 41,020개 유전자의 단백질 서열을 fasta 포맷으로 구성하고 이를 input 파일로 활용하고 NCBI FTP server(ftp://ftp.ncbi.nlm.nih.gov/)에서 확보한 현재까지 NCBI에 등재된 단백질에 대한 서열 정보를 중복 없이

기록한 non-redundant(nr) 데이터베이스를 makeblastdb 명령어를 통해 query인 배추 표준 유전자와의 아미노산 서열상의 similiarity를 상호비교할 데이터베이스로 구성하여 BLASTP를 수행하였다(-evalue 10^{-5}, -max_target_seq 1). 이를 통해 배추의 표준 유전자 ID에 nr에서 나타나는 GI number를 개별적으로 부여하였다.

Gene annotation에 반영할 기본 구성 요소의 수집은 NCBI의 ftp 서버 및 Gene ontology consortium에서 확보하였으며 이를 보유한 분석 서버 상에서 gene2go (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go: NCBI ID, Gene ontology number 와 그의 description 연관 관계 그리고 Biological process, Cellular component, Molecular function 여부), gene2info (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2info: NCBI ID 및 gene symbol 과 gene description 연관 정보 기재), ec2go (http://www.geneontology.org/external2go/ec2go: EC number와 GO ID 연관 정보 포함)라는 파일명을 부여하여 저장하였다.

각 파일들은 GI number를 기준으로 GI number에 대한 속성을 record로서 보유하고 있기 때문에 이 연관관계를 이용하여 BLASTP를 통해 배추 유전자에 부여한 GI number에 EC, GO, gene description에 대한 정보를 연결시킬 수 있다.

배추의 41,020개 유전자의 단백질 서열위에 분포하는 domain에 대한 정보를 생산하기 위해 배추 표준 유전체의 protein sequence fasta 파일을 Pfam(pfam.xfam.org/search)의 batch search 페이지에 입력하였다. domain search 결과는 사용자의 메일 계정으로 회신되고 분석서버로 이관하여 cat 명령어로 하나로 취합하였다. 배추의 표준 유전자 상에 위치한 domain의 종류와 Pfam accession number 정보를 얻을 수 있다.

(3) 모델 식물 애기장대(*Arabidopsis Thaliana*) 유전자에 대한 ortholog 정보

배추는 애기장대가 whole genome의 triplication을 거친 후 다시 2배체로 돌아가는 과정을 거치면서 나타난 A genome에 속하며 그 결과 배추의 유전체 상에 하나의 애기장대 유전자에 대한 orthologous gene이 복수로 나타나게 되었다. 배추의 orthologous gene에 대한 정보는 g:profiler(biit.cs.ut.ee/gprofiler)의 g:Convert 메뉴에서 Organisms을 *Brassica rapa,* Target Database를 TAIR LOCUS로 설정하고 input으로 배추 표준 유전자의 ID를 중복없이 입력하여 배추 유전자와 ortholog인 애기장대 유전자의 TAIR ID와 유전자의 기능에 대한 설명에 대한 정보를 얻을 수 있다(Reimand et al, 2016).

(4) Uniprot 기반 정보

BLASTP에 의해 GI number가 부여된 배추의 표준 유전자 ID들을 대상으로 얻은 GI number를 Uniprot의 ID mapping service (http://www.uniprot.org/uploadlists/)에 입력하였다(The uniprot consortium, 2017). 이를 통해 사용자에 의해 customizing된 annotation을 얻고 이를 text file의 형태로 다운로드가 가능하다. GI number에 의해 Uniprot database로부터 추출할 annotation은 Organism, PANTHER ID, Pubmed ID, Gene symbol이 선택되었으며 입력된 GI number에 대한 annotation 결과를 리눅스 분석 서버로 이관하여 SED 명령어를 통해 organisms에 대한 정보가 식물계에 포함되지 않는 행을 제거하고 남은 record만을 데이터로서 반영하였다.

(5) miRNA 정보

miRNA는 그 구조상에 존재하는 binding site의 작용으로 target 유전자의 발현에 영향을 미치는 배추의 표현형에 관계된 생리적 패스웨이 상의 중요한 인자이다. 작업의 수행을 위해 배추 유전체 내의 miRNA에 대한 분류와 동정을 수행한 연구 결과를 3개의 서로 다른 논문들로부터 수집했다(Dhandapni et al, 2011; Kim et al, 2012; Sun et al, 2015). 각 연구 결과로부터 정리된 miRNA 리스트에 대한 정보를 상호 비교하여 mature miRNA의 sequence가 서로 동일한 것으로 인정되고 유전체 상의 위치가 같은 서로 다른 연구에서 나타난 miRNA를 동일한 miRNA로 분류하여 배추 유전체상에 총 1,055 지점의 miRNA coding 지역이 있는 것을 확인하였으며 이는 총 496개의 서로 다른 miRNA family로 분류되었다(표 3).

표 3. 배추(*B. rapa*)의 miRNA 식별에 대한 문헌 및 miRNA의 문헌간 상호 비교

Author	Paper	Available miRNA in research
Dhandapani et al. (2011)	Identification of Potential microRNAs and Their Targets in *Brassica rapa* L.	105
Kim et al. (2012)	Identification and profiling of novel microRNAs in the *Brassica rapa* genome based on small RNA deep sequencing	412
Sun et al. (2015)	Impacts of Whole-Genome Triplication on MIRNA Evolution in *Brassica rapa*	675
		Total: 1,192
		Unique coding region: 1,055
		Unique miRNA family: 496

다. 배추 전사체 발현 데이터 분석 특화 데이터베이스 BrTED의 launching

(1) BrTED의 URL과 구성

BrTED는 PHP, Mysql, Apache를 개발 플랫폼으로 활용하여 구축되었으며 현재 BrTED는 도메인 brted.cnu.ac.kr을 충남대학교 정보통신원으로부터 제공받아 접속이 가능하다 (그림 4). BrTED에는 5개의 분석 및 정보 검색 메뉴가 정립되어 있으며(GENE SEARCH, DEG ANALYSIS, GROUP EXPRESSION, MIRNA, KEGG) 각 메뉴에서 사용자의 query에 의해 출력되는 검색 및 분석 결과는 서로 유기적으로 연결되어 상호 참조가 가능하다 (그림 5).

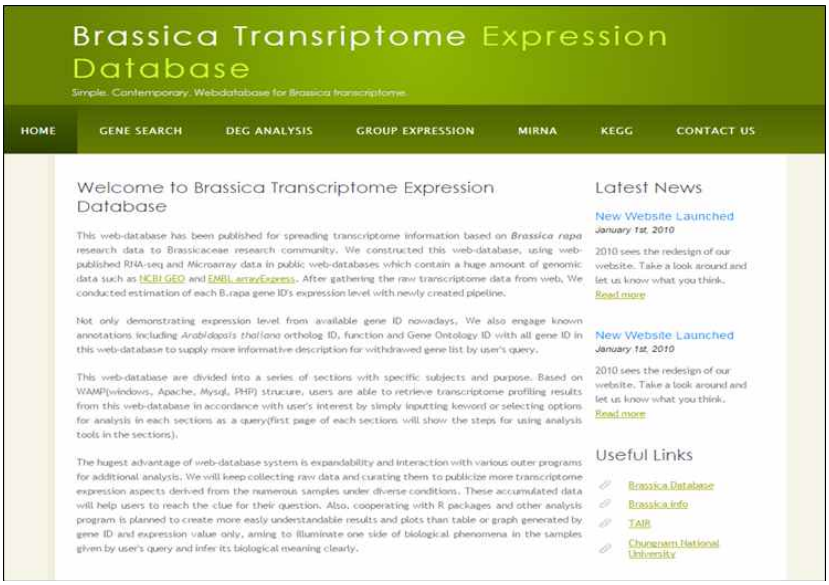


그림 4. BrTED의 메인 페이지와 5개의 분석 및 정보 검색 메뉴 URL: brted.cnu.ac.kr

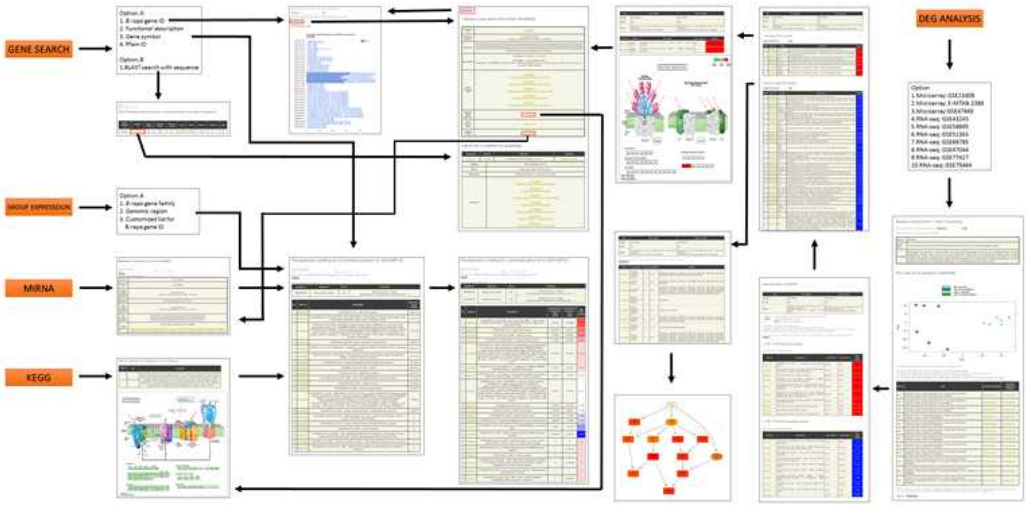


그림 5. BrTED의 5개의 분석 및 정보 검색 메뉴와 사용자 query에 의해 출력된 정보의 상호 관계

(2) GENE SEARCH

(가) keyword를 통한 배추 표준 유전자의 검색

배추 표준 유전자 정보의 검색을 지원하기 위한 검색 옵션을 구성하였다. 검색 옵션으로서 애기장대의 homologous gene의 유전자 기능을 활용한 gene description, Brassica research community에서 상용중인 *B. rapa* ID, 배추 표준 유전자에 대하여 관용적으로 사용하고 있는 gene symbol 그리고 배추 표준 유전자 상에 위치한 domain에 대한 pfam ID를 1차적으로 선정하였다(그림 6.A).

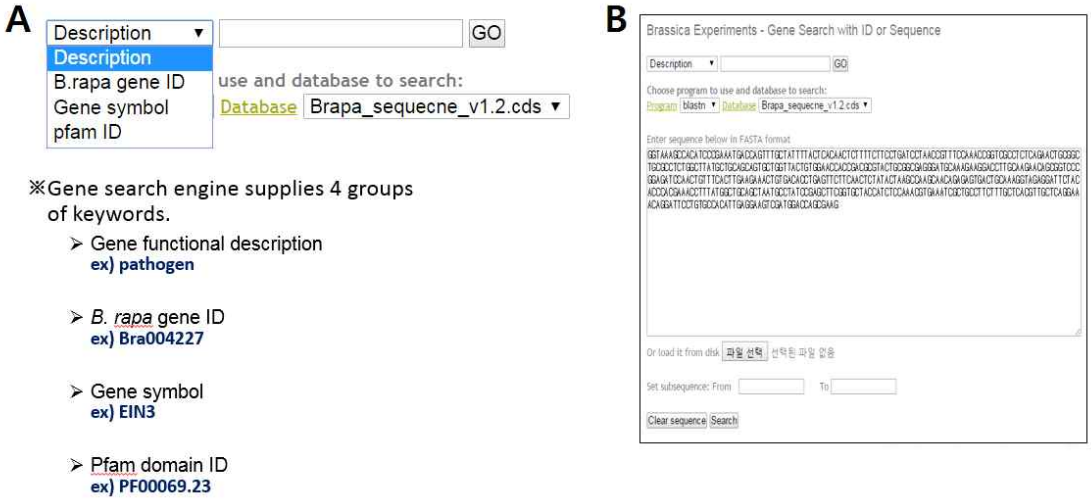
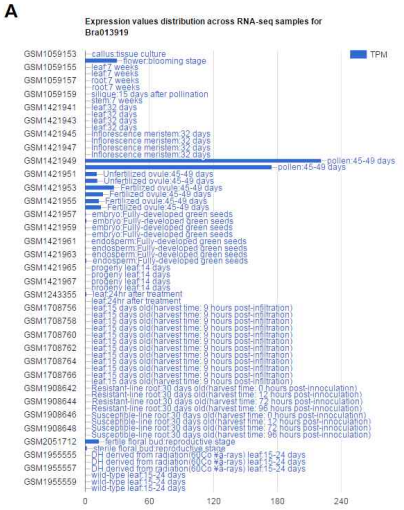


그림 6. A) BrTED의 배추 표준 유전자 검색 옵션과 예시 B) 사용자가 보유한 염기서열의 BLAST 처리를 통한 배추 표준 유전자 정보의 검색

사용자가 검색창 옆에 있는 옵션을 *B. rapa* gene ID로 설정하여 원하는 배추 표준 유전자 ID를 입력하면 해당 유전자가 BrTED에 입력된 모든 RNA-seq sample에서의 TPM normalization 처리가 된 발현량이 bar plot의 형태로 출력되어 해당 유전자가 어떠한 처리 및 조건에서 발현되는지를 확인할 수 있다. 또한 bar plot 하단의 table에서 조건 및 처리와 발현량에 대한 자세한 정보에 대하여 접근할 수 있다 (그림 7).

배추 표준 유전자의 검색 옵션을 *B. rapa* gene ID 이외의 keyword를 선택하여 query에 대한 결과를 출력할 경우 BrTED는 입력된 정보에 해당하는 모든 배추 표준 유전자에 대한 annotation을 순번, 배추 표준 유전자의 ID, 애기장대 유전자의 orthologous ID, gene symbol, gene description의 순으로 table 상에 출력한다(그림 8).

A



B

Rank	Experiment Code	Sample ID	Organism	Tissue	RNA extraction moment	Category	Treatment	Expression value(TPM)
1	GSE58895	GSM1421949	Brassica rapa (R-o-18)	pollen	45-49 days	Heat Stress Control	22	221.163
2	GSE58895	GSM1421950	Brassica rapa (R-o-18)	pollen	45-49 days	Heat Stress	42	174.695
3	GSE43245	GSM1059154	Brassica rapa (chifu)	flower	blooming stage	Transcript level estimation only	None	30.184
4	GSE58895	GSM1421953	Brassica rapa (R-o-18)	Fertilized ovule	45-49 days	Heat Stress Control	22	27.377
5	GSE58895	GSM1421954	Brassica rapa (R-o-18)	Fertilized ovule	45-49 days	Heat Stress Control	22	16.96
6	GSE58895	GSM1421956	Brassica rapa (R-o-18)	Fertilized ovule	45-49 days	Heat Stress	42	15.09
7	GSE58895	GSM1421955	Brassica rapa (R-o-18)	Fertilized ovule	45-49 days	Heat Stress	42	13.387
8	GSE77427	GSM2051712	Brassica rapa subsp. pekinensis	fertile floral bud	reproductive stage	Transcript level estimation only	None	13.263
9	GSE58895	GSM1421952	Brassica rapa (R-o-18)	Unfertilized ovule	45-49 days	Heat Stress	42	11.896
10	GSE58895	GSM1421951	Brassica rapa (R-o-18)	Unfertilized ovule	45-49 days	Heat Stress Control	22	11.165
11	GSE77427	GSM2051713	Brassica rapa subsp. pekinensis	sterile floral bud	reproductive stage	Transcript level estimation only	None	1.804
12	GSE51363	GSM1243355	Brassica rapa subsp. pekinensis(cultivar: cv. ZS05)	leaf	24hr after treatment	Control of chemical treatment	sprayed with: pure water	1.701
13	GSE58895	GSM1421958	Brassica rapa (R-o-18)	embryo	Fully-developed green seeds	Heat Stress Control	22	1.483

그림 7. A) 검색된 배추 표준 유전자 Bra013919의 수집 RNA-seq sample에 대한 expression profiling B) Bra013919의 expression profiling에 대한 detail

Brassica Experiments - Gene Search with ID or Sequence

Description ▼ GO

Result by searching keyword: photosynthesis

Select first sample

---- ▼ choose

No.	Identification	A.thaliana ortholog	Gene symbol	Description
1	Bra012057	AT5G66190	ATLFNR1, FNR1	AT5G66190 (E=8e-054) ATLFNR1, FNR1 | FNR1 (FERREDOXIN-NADP(+)-OXIDOREDUCTASE 1); NADPH dehydrogenase/ electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis/ electron transporter, transferring electrons within the noncyclic electron transp
2	Bra022019	AT5G46110	APE2, TPT	AT5G46110 (E=8e-194) APE2, TPT | APE2 (ACCLIMATION OF PHOTOSYNTHESIS TO ENVIRONMENT 2); antiporter/ triose-phosphate transmembrane transporter
3	Bra025037	AT5G46110	APE2, TPT	AT5G46110 (E=1e-191) APE2, TPT | APE2 (ACCLIMATION OF PHOTOSYNTHESIS TO ENVIRONMENT 2); antiporter/ triose-phosphate transmembrane transporter
4	Bra028418	AT5G38660	APE1	AT5G38660 (E=2e-113) APE1 | APE1 (ACCLIMATION OF PHOTOSYNTHESIS TO ENVIRONMENT)
5	Bra037182	AT5G66190	ATLFNR1, FNR1	AT5G66190 (E=1e-179) ATLFNR1, FNR1 | FNR1 (FERREDOXIN-NADP(+)-OXIDOREDUCTASE 1); NADPH dehydrogenase/ electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis/ electron transporter, transferring electrons within the noncyclic electron transp

그림 8. BrTED의 GENE SEARCH에서 Photosynthesis를 keyword로 검색된 관련 배추 표준 유전자 정보에 대한 결과

(나) BLAST 지원을 통한 사용자 보유 sequence의 homology 검색

사용자가 본인이 보유하고 있는 염기서열 혹은 아미노산 서열을 대상으로 BrTED에서 확인 가능한 배추 표준 유전자에 대한 관련된 정보를 열람하고자 할 때 GENE SEARCH의 BLAST 검색 기능을 이용할 수 있다. 현재 sequence를 input으로 한 검색에 대하여 BLASTN, BLASTP, BLASTX 가 지원되고 있으며 BLAST의 데이터베이스로서 *B. rapa* whole genome, coding gene sequence 그리고 protein sequence가 각각 1.2 version 과 1.5 version으로 활용이 가능하다(그림 6.B).

(3) DEG ANALYSIS

DEG ANALYSIS는 전세계의 연구진이 NCBI 및 EMBL에 등록한 배추의 전사체 발현에 대한 raw data를 재가공하여 처리 및 조건에 따른 배추 표준 유전자의 발현량 및 그에 대한 normalization 결과를 집약한 메뉴이다. 현재 BrTED의 핵심적인 data pool로서 수집된 10개의 서로 다른 실험 내에서 (그림 9) 하나의 전사체 데이터를 반영하는 2개의 샘플을 자유롭게 선택하고 이 서로 다른 두 조건에서 드러나는 DEG를 annotation과 함께 확인할 수 있다. 웹페이지 상에서 제시되는 옵션을 사용자가 직접 조정하여 DEG의 GO에 대한 grouping 및 KEGG pathway 상에서의 mapping과 같은 추가적인 분석이 가능하다.

Brassica Experiments - DEG Comparing

Please select one of available experiments. Experiment GO

No.	Experiment	Title	Platform	Sample number
1	GSE23409	Comparison of root and leaf samples from Brassica rapa	spotted oligonucleotide	6
2	E-MTAB-2386	Transcription profiling by array of 10 days old Brassica rapa ssp. chinensis seedlings treated with 2mM methyl jasmonate by spraying and harvesting 48 hours past treatment	spotted oligonucleotide	6
3	GSE47665	Comprehensive analysis of genic male sterility-related genes in Brassica rapa using Brassica rapa 300k microarray v2.0	in situ oligonucleotide	14
4	GSE43245	Transcriptome sequencing of Brassica rapa tissues	high-throughput sequencing	8
5	GSE58895	Elucidation of stress memory inheritance through epigenome alterations in Brassica rapa plants [RNA-seq]	high-throughput sequencing	28
6	GSE51363	Global Analysis of the Transcriptional Response of Chinese cabbage (Brassica rapa ssp. pekinensis) to Methyl Jasmonate Reveals JA Signaling on Enhancement of Secondary Metabolism Pathways	high-throughput sequencing	2
7	GSE69785	Transcriptome profiling of the defense response following elicitation by a bacterial pathogen or flg22	high-throughput sequencing	12
8	GSE74044	Transcriptome analysis of Brassica rapa near-isogenic lines carrying clubroot-resistant and -susceptible alleles in response to Plasmodiophora brassicae during early infection	high-throughput sequencing	8
9	GSE77427	Comparative transcriptome of the fertile and sterile buds of genic multiple-allele inherited male sterile AB line in Chinese cabbage	high-throughput sequencing	2
10	GSE75464	Physiological characterization and comparative transcriptome analysis of a developmentally retarded Chinese cabbage (Brassica campestris ssp. pekinensis) mutant	high-throughput sequencing	6

그림 9. BrTED에 수집된 배추의 전사체 발현에 대한 실험 및 이를 구성하는 샘플 현황

(가) 조건에 따른 DEG의 식별

사용자가 BrTED에서 제공되는 10개의 실험 중 하나를 선택하여 실험내의 서로 다른 조건의 두 샘플에 대한 DEG를 확인할 수 있다. DEG ANALYSIS의 첫 번째 분석 페이지에서는 두 샘플에서 나타나는 전사체의 발현량을 비교한다. 두 샘플에서 나타나는 DEG의 대푯값으로서 동일한 배추 표준 유전자에 대하여 log 2 fold change를 계산하였을 때 나타난 up/down regulated gene 각각에 대하여 절대값을 기준으로 최소 10개, 최대 200개의 DEG를 산출한다(그림 10. B,C).

두 샘플 간에서 나타나는 DEG에 대한 분석은 1차적으로 산출된 DEG에 대한 리스트를 php 플랫폼에서 BrTED 외부에 위치한 R script에 대한 input으로 구성하고 이를 R script에 입력하여 DEG 분석에 대한 연산을 수행한다. 연산 결과는 text 포맷으로 서버의 특정 경로 상에 저장되고 php 플랫폼이 이를 읽어 웹페이지 상의 포맷에 맞게 전환하여 사용자에게 제시하게 된다.

이를 위한 선택된 두 샘플의 전체적인 DEG 비교와 DEG 리스트의 작성은 분석 페이지 화면 상단의 선택 옵션에서 두 샘플에 대한 control/treatment 관계와 DEG의 threshold를 설정함으로써 수행할 수 있다(그림 10. A).

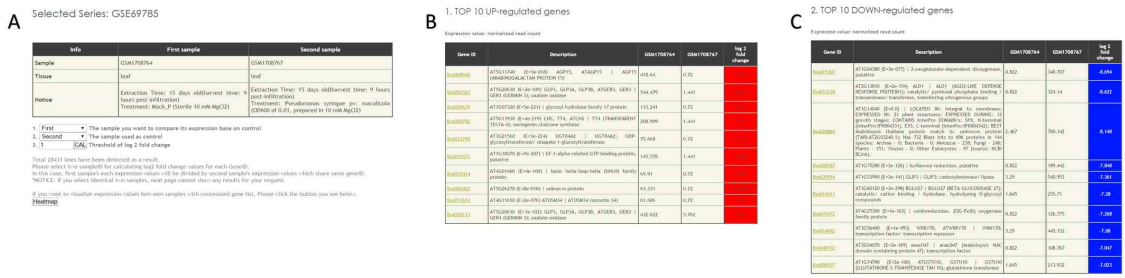


그림 10. A) DEG 분석을 위해 선택된 샘플의 detail B) 선택된 샘플 간의 top 10 up-regulated DEG C) 선택된 샘플 간의 top 10 down-regulated DEG

(나) R script에 의한 DEG의 분석

사용자에 의해 특정 실험에서 선택된 두 개의 샘플은 control/treatment 관계설정 및 log 2 fold change에 대한 threshold 조정 이후 추가 분석을 수행하면 threshold 조건을 충족시키는 전체 유전자의 수와 up/down regulated gene의 수를 분석 옵션 위에 표시한다. 옵션 하단의 table에는 BrTED에 기록되어 있는 배추 표준 유전자 ID의 순서대로 유전자 ID, 샘플별 발현량, 그리고 log2 fold change 순으로 웹페이지 상에 출력된다(그림 11). SORT 옵션을 통해 log 2 fold change를 오름차순(DESC) 혹은 내림차순(ASC)으로 정렬할 수 있으며 출력된 배추 표준 유전자 ID는 각 ID에 대한 annotation 페이지로 하이퍼링크를 통해 선택 유전자에 대한 정보를 열람할 수 있다.

Selected Series: GSE58895

Info	First sample	Second sample
Sample	GSM1421944	GSM1421942
Tissue	leaf	leaf
Notice	Extraction Time: 32 days Treatment: 42	Extraction Time: 32 days Treatment: 22

Total 643 lines have been detected as a result.
In this result, 374 genes are up-regulated, whereas 269 genes are down_regulated.
You can sort this table with log2 fold change values(Decrease or Increase order is available).
If you want to check B.rapa gene's information, Please click the GeneID you want to check in the first column of table.
DESC ▾ SORT

DEG_ANALYSIS

Gene ID	First sample GSM1421944	Second sample GSM1421942	log2 fold change
Bra000019	4.38114	0.831752	2.39708
Bra000093	6.5717	0.831752	2.98204
Bra000132	36.1444	301.926	-3.06235
Bra000338	33.9538	7.48577	2.18135
Bra000377	848.845	206.275	2.04094
Bra000382	29.5727	232.059	-2.97215
Bra000518	7.66699	1.6635	2.20443
Bra000610	16.4293	1.6635	3.30397
Bra000707	1.09528	10.8128	-3.30336
Bra000718	2.19057	9.14928	-2.06235
Bra000889	5.47642	275.31	-5.65168

그림 11. 선택된 두 샘플의 전사체 발현양상 전반의 비교를 통한 DEG 산출 결과의 일부

DEG_ANALYSIS 버튼을 선택하면 배추 표준 유전자 ID 순으로 나타난 DEG 리스트가 up/down regulated gene별로 분류되어 log2 fold change의 절대값을 기준으로 내림차순으로 정렬된다. 해당 결과는 두 개의 table에 나누어져 출력되며 분석 옵션으로 KEGG Enrichment test와 Gene Ontology Enrichment test를 제공한다(그림 12). 이와 같은 DEG 리스트에 대한 분석은 up-regulated gene 혹은 down-regulated gene 그룹별로 사용자의 선택에 의해 독립적으로 이루어지며 각 그룹별로 나타난 DEG 리스트는 리스트 내의 배추 표준 유전자에 상응하는 애기장대의 TAIR ID로 전환되게 된다. 전환된 리스트를 input으로 활용되고 이는 사용자가 선택한 분석 옵션을 관장하는 R script에 전달되어 유전자의 Enrichment test의 형식으로 분석이 이루어지게 된다. 분석 결과는 1차적으로 table 형태로 GO와 KEGG에 포함되는 특정 그룹에 대하여 enrich된 배추 표준 유전자들에 대한 리스트가 분류하게 되며 분류된 배추 표준 유전자 ID를 시각화시키는 기능을 제공한다(그림 13, 14).

1.Up regulation genes

Kegg Enrichment test | GO

Kegg Enrichment test

Gene Ontology Enrichment test

			Description	Log2FC
1	Bra006467	N/A	AT5G18600 (E=4e-052) | glutaredoxin family protein	-4.397
2	Bra040601	N/A	AT3G02800 (E=7e-079) | phosphatase/ phosphoprotein phosphatase/ protein tyrosine phosphatase	-4.098
3	Bra038367	N/A	AT1G68570 (E=2e-312) | proton-dependent oligopeptide transport (POT) family protein	-4.098
4	Bra031833	LBD37	AT5G67420 (E=4e-092) LBD37 | LBD37 (LOB DOMAIN-CONTAINING PROTEIN 37)	3.982
5	Bra012981	PDCB1	AT5G61130 (E=2e-069) PDCB1 | PDCB1 (PLASMODESMATA CALLOSE-BINDING PROTEIN 1); callose binding / polysaccharide binding	3.789
6	Bra003142	N/A	AT3G53530 (E=2e-094) | heavy-metal-associated domain-containing protein	3.719
7	Bra020696	N/A	AT5G48490 (E=2e-035) | protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	3.719
8	Bra021223	ATR2, AR2	AT4G30210 (E=2e-014) ATR2, AR2 | ATR2 (ARABIDOPSIS P450 REDUCTASE 2); NADPH-hemoprotein reductase	3.567
9	Bra040579	N/A	AT4G16442 (E=9e-056) | integral membrane family protein	3.567
10	Bra016286	N/A	AT1G26790 (E=3e-102) | Dof-type zinc finger domain-containing protein	3.567
11	Bra027397	ATCVINV1, ATBFRUCT1	AT3G13790 (E=0.0) ATCVINV1, ATBFRUCT1 | ATBFRUCT1; beta-fructofuranosidase/ hydrolase, hydrolyzing O-glycosyl compounds	3.567

2.Down regulation genes

Kegg Enrichment test | GO

RANK	Gene ID	Symbol	Description	Log2FC
1	Bra000889	ATCNGC12, CNGC12	AT2G46450 (E=4e-116) ATCNGC12, CNGC12 | ATCNGC12; cation channel/ cyclic nucleotide binding / ion channel	-5.652
2	Bra005455	N/A	AT2G34070 (E=2e-077) | unknown protein	-4.925
3	Bra013724	ATCHX17, CHX17	AT4G23700 (E=0.0) ATCHX17, CHX17 | ATCHX17 (CATION/H+ EXCHANGER 17); monovalent cation:proton antiporter/ sodium:hydrogen antiporter	-4.515
4	Bra025658	ATNAC2, ORE1, ANAC092, ATNAC6	AT5G39610 (E=4e-137) ATNAC2, ORE1, ANAC092, ATNAC6 | ATNAC6 (ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 6); protein heterodimerization/ protein homodimerization/ transcription factor	-4.188
5	Bra033229	N/A	AT1G02470 (E=6e-071) | FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: chloroplast; EXPRESSED IN: 10 plant structures; EXPRESSED DURING: 4 anthesis, C globular stage, petal differentiation and expansion stage; CONTAINS InterPro DOMAIN/s: Streptomyces cyclase/dehydrase (InterPro:IPR005031); BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT1G02475.1); Has 276 Blast hits to 276 proteins in 86 species: Archae - 0; Bacteria - 162; Metazoa - 0; Fungi - 0; Plants - 29; Viruses - 0; Other Eukaryotes - 85 (source: NCBI BLink).	-4.179
6	Bra027957	N/A	AT1G54890 (E=3e-182) | late embryogenesis abundant protein-related / LEA protein-related	-3.995
7	Bra034691	N/A	AT3G13310 (E=1e-062) | DNAJ heat shock N-terminal domain-containing protein	-3.925
8	Bra035136	atnudt3	AT1G79690 (E=0.0) atnudt3 | atnudt3 (Arabidopsis thaliana Nudix hydrolase homolog 3); hydrolase	-3.851

그림 12. Up/down-regulated gene 리스트와 Enrichment test 옵션

A

Info	First sample	Second sample
Sample	GSM1708766	GSM1708763
Tissue	leaf	leaf
Notice	Extraction Time: 15 days old(harvest time: 9 hours post-infiltration) Treatment: Pseudomonas syringae pv. maculicola (OD600 of 0.01, prepared in 10 mM MgCl2)	Extraction Time: 15 days old(harvest time: 9 hours post-infiltration) Treatment: Mock_P (Sterile 10 mM MgCl2)

Gene ID	Symbol	Description	EC number	Log2FC
Bra000064	WRKY33, ATWRKY33	AT2G38470 (E=4e-154) WRKY33, ATWRKY33 | WRKY33; transcription factor	N/A	3.859
Bra000985	N/A	AT3G01830 (E=1e-041) | calmodulin-related protein, putative	N/A	3.774
Bra001676	CNBT1, CNGC20, ATCNGC20	AT3G17700 (E=7e-210) CNBT1, CNGC20, ATCNGC20 | CNBT1 (CYCLIC NUCLEOTIDE-BINDING TRANSPORTER 1); calmodulin binding / cyclic nucleotide binding / ion channel	N/A	3.739
Bra004537	ATCNGC3, CNGC3, CNGC3.C	AT2G46430 (E=7e-311) ATCNGC3, CNGC3, CNGC3.C | ATCNGC3 (CYCLIC NUCLEOTIDE GATED CHANNEL 3); calmodulin binding / cyclic nucleotide binding / ion channel	N/A	3.712
Bra009653	CPK6, ATCDPK3	AT2G17290 (E=1e-297) CPK6, ATCDPK3 | CPK6 (CALCIUM-DEPENDENT PROTEIN KINASE 6); ATP binding / calcium ion binding / calmodulin-dependent protein kinase/ kinase/ protein kinase/ protein serine/threonine kinase	EC:2.7.11, EC:2.7.11.17	3.611
Bra011221	ATPRB1	AT2G14580 (E=2e-067) ATPRB1 | ATPRB1	N/A	3.511
Bra013231	RIN4	AT3G25070 (E=1e-079) RIN4 | RIN4 (RPM1 INTERACTING PROTEIN 4); protein binding	N/A	3.11

B

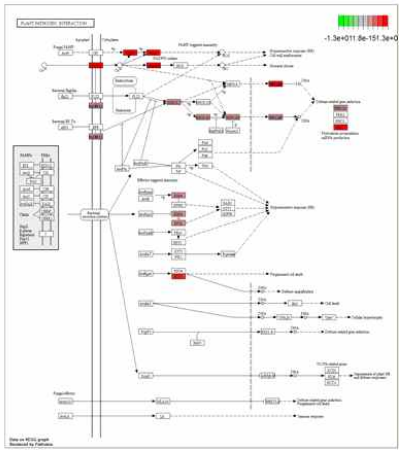


그림 13. A) *P. syringae* 감염에서 나타나는 up-regulated gene의 DEG에서 나타난 KEGG pathway내의 배추 표준 유전자의 정보 B) DEG의 KEGG pathway map 상의 mapping

A

If you want to visualize the result, Please click the button you see below this sentence.

Visualization

GO ID	Description	count	GeneRatio	geneID
GO:0019953	sexual reproduction	3	3/692	Bra008347 Bra009613 Bra036070
GO:0032504	multicellular organism reproduction	5	5/692	Bra010453 Bra016512 Bra036070 Bra037113 Bra037112
GO:0051321	meiotic cell cycle	4	4/692	Bra011509 Bra015436 Bra020835 Bra024436
GO:0006807	nitrogen compound metabolic process	129	129/692	Bra000330 Bra000366 Bra000393 Bra000557 Bra001377 Bra001752 Bra038219 Bra022436 Bra001815 Bra001921 Bra001972 Bra023643 Bra002112 Bra002747 Bra003239 Bra014692 Bra007243 Bra003666 Bra003828 Bra015964 Bra004388 Bra004473 Bra004540 Bra004736 Bra004767 Bra017117 Bra005104 Bra005775 Bra006178 Bra006195 Bra006197 Bra006210 Bra006481 Bra006749 Bra006864 Bra003229 Bra014692 Bra007243 Bra007826 Bra007869 Bra008131 Bra008435 Bra009193 Bra009885 Bra036563 Bra020546 Bra010419 Bra011299 Bra011422 Bra011461 Bra011509 Bra011529 Bra011545 Bra011880 Bra012269 Bra012780 Bra012999 Bra013426 Bra013584 Bra013714 Bra013802 Bra014417 Bra003239 Bra014692 Bra007243 Bra015048 Bra015306 Bra015307 Bra015435 Bra015481 Bra015842 Bra015935 Bra003828 Bra015964 Bra016252 Bra016389 Bra016432 Bra016512 Bra031035 Bra016553 Bra016833 Bra017117 Bra005104 Bra017264 Bra017408 Bra017495 Bra017623 Bra017800 Bra017871 Bra018725 Bra019052 Bra019346 Bra019706 Bra019776 Bra020179 Bra009885 Bra036563 Bra020546 Bra020574 Bra020713 Bra037506 Bra020835 Bra021047 Bra021048 Bra040159 Bra021047 Bra021048 Bra040159 Bra001752 Bra038219 Bra022436 Bra029107 Bra022605 Bra023598 Bra023643 Bra002112 Bra023651 Bra025400 Bra026280 Bra027057 Bra028202 Bra028918 Bra029107 Bra022605 Bra029553 Bra030246 Bra030430 Bra031649 Bra030748 Bra030989 Bra031035 Bra016553 Bra031043 Bra031042 Bra031043 Bra031042 Bra031588 Bra031649 Bra030748 Bra031903 Bra032601 Bra032665 Bra032762 Bra032845 Bra033431 Bra033436 Bra034413 Bra034705 Bra035320 Bra036327 Bra036360 Bra036550 Bra009885 Bra036563 Bra020546 Bra020713 Bra037506 Bra001752 Bra038219 Bra022436 Bra038494 Bra038509 Bra039265 Bra039755 Bra021047 Bra021048 Bra040159 Bra040557 Bra040902

B

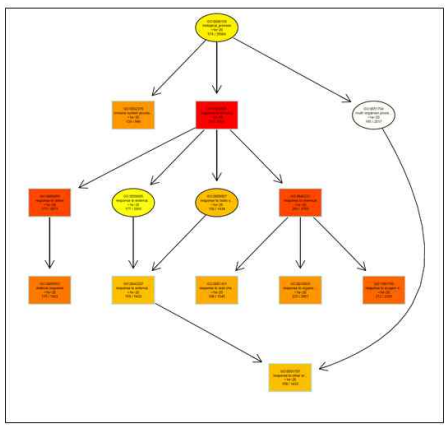


그림 14. A) *P. syringae* 감염 사례에 대한 DEG의 GO grouping B) 감염된 개체에서 나타나는 up-regulated gene list에 대한 GO network image의 시각화

(다) Heatmap을 통한 발현량의 시각화

DEG ANALYSIS 메뉴에서 제시된 DEG 분석 절차에서 사용자에 의해 설정된 threshold를 충족하는 배추 표준 유전자에 대한 정보만을 출력하는 한계를 극복하기 위한 절차를 별도로 구성하였다. 조건 및 처리에 따라 사용자에 의해 선택된 복수의 샘플에서 나타난 모든 유전자 발현량을 data pool로 활용하고 text area를 웹페이지 상에 구성하였다. 사용자가 발현량을 확인하고자 하는 배추 표준 유전자 ID에 대한 리스트를 text area상에 입력하게 되면 선택된 샘플들 상에서 해당 유전자 리스트에 대한 발현량을 출력하고 이를 Heatmap plot의 형식으로 출력하게 된다(그림 15).

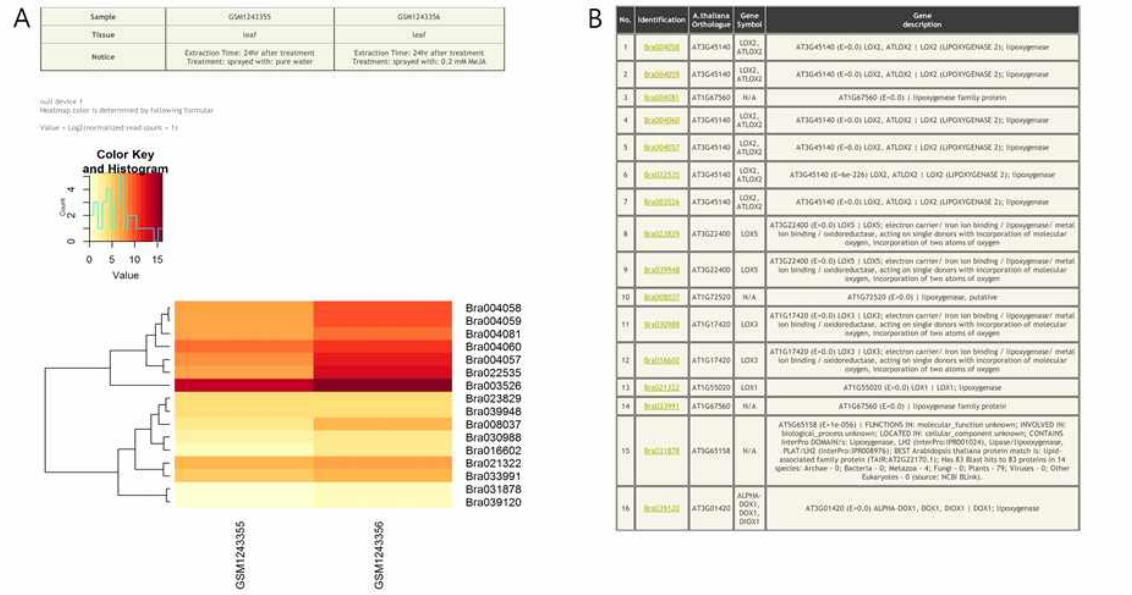


그림 15. A) Heatmap plot에 의한 배추 잎 조직의 meJA 처리 유무에 대한 DEG의 시각화 및 clustering B) Heatmap plot 구성에 사용된 배추 표준 유전자(LOX family)의 annotation table

(4) GROUP EXPRESSION

BrTED의 GROUP EXPRESSION 메뉴는 배추의 조건 및 처리에 대해 사용자가 지정한 옵션에 대하여 변동적으로 나타나는 DEG가 아닌 특정한 조건에 해당되는 유전자 리스트에 대한 발현량을 확인할 수 있는 기능을 제공한다.

메뉴내의 옵션의 조정을 통해 결과를 출력하면 조건에 해당하는 유전자들의 리스트와 그에 대한 간략한 annotation이 table 상에 나타난다. 그리고 table 위의 dynamic chained select box option을 통해 BrTED에서 제공하는 RNA-seq 및 Microarray 샘플에 접근하여 출력된 유전자들이 선택된 샘플에서의 발현량을 확인할 수 있다. 또한 옵션 상에 나타난 체크박스를 선

택한 후 다른 샘플을 선택하게 되면 1차적으로 선택한 샘플과 2차적으로 선택한 샘플의 발현량을 log2 fold change를 통해 비교하여 조건의 차이에 의해 특정 유전자의 집단의 발현량이 어떻게 변화하는지를 확인할 수 있다.

GROUP EXPRESSION에서 크게 세 가지의 유전자의 분류 기능이 제공되고 있다(그림 16).

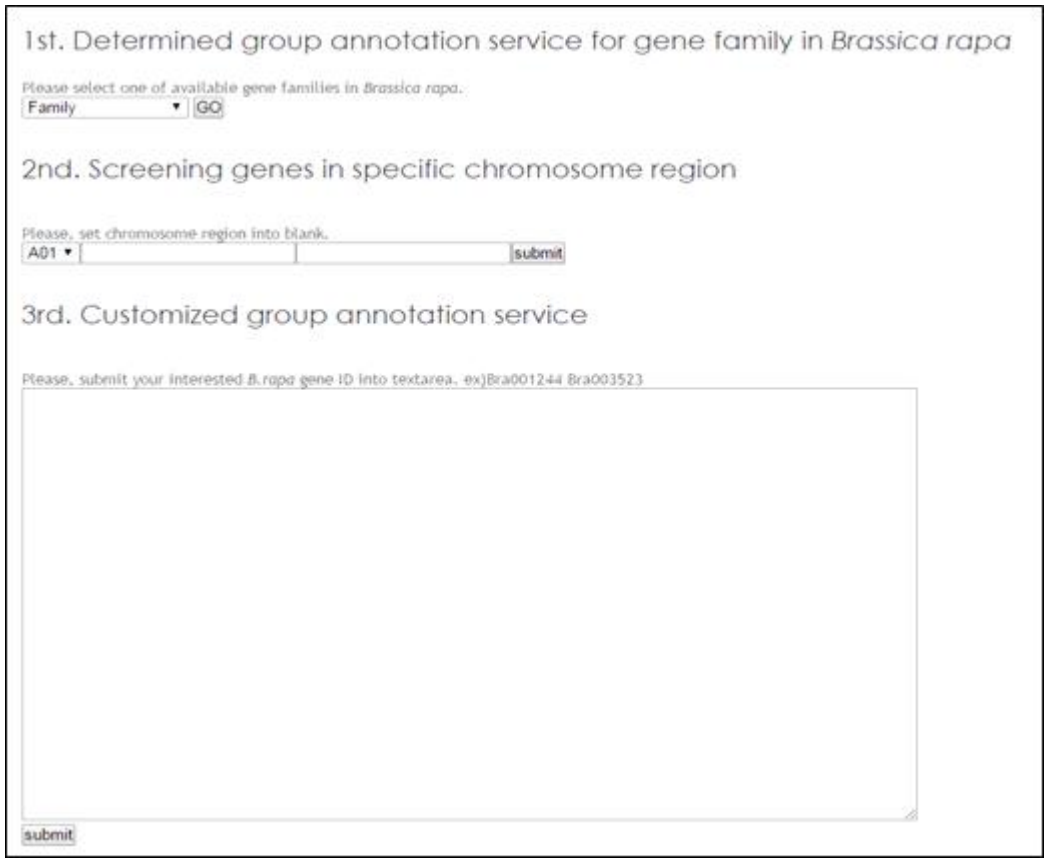


그림 16. GROUP EXPRESSION의 유전자 그룹 출력 옵션

(가) 배추의 gene family에 대한 조건별 발현량 산출

배추의 MADS BOX, VQ motif, Leucine Rich Repeat, NAC, MYB, 그리고 WRKY family에 속하는 유전자들의 리스트를 문헌조사 및 외부데이터베이스에서의 검색을 통해서 family별로 정리하였다. 사용자가 select box에서 특정 family의 이름을 선택하면 결과 페이지에서 선택된 family에 속하는 배추 표준 유전자 ID와 gene symbol, EC number, gene description이 출력된다. 이후 결과 페이지에서 제공하는 RNA-seq 샘플을 선택할 수 있는 select box에서 특정 샘플 ID를 선택하면 해당 샘플 내에서의 특정 gene family에 속하는 배추 표준 유전자의 발현량의 normalization된 값을 확인할 수 있다.

(나) 배추의 표준 유전체 상의 특정 영역의 유전자 출력

배추 표준 유전체 1.2 version에서 배추의 표준 유전자 각각에 대한 위치 정보를 활용하여 GROUP SEARCH 메뉴에서 배추 표준 유전자를 filtering 할 수 있는 기능을 구현하였다. 사용자가 select option에서 탐색을 원하는 chromosome 번호를 선택하고 선택된 chromosome 상의 물리적 위치를 시작 지점과 종결 지점을 설정하면 해당 영역 상에 있는 모든 배추 표준 유전자가 테이블 사에 출력되게 된다. 종결 지점을 입력하지 않은 경우 default option이 적용되어 시작 지점 이후 100,000bp 상의 영역을 자동을 탐색하여 결과를 출력한다.

(다) 사용자에 의해 입력된 배추 표준 유전자에 대한 annotation 출력

사용자가 특정 조건에 해당하는 배추 표준 유전자들의 조합과 그에 대한 BRAD 기준 유전자 ID를 확보한 경우, 이 유전자 ID를 GROUP SEARCH의 text area에 일렬로 입력하여 이에 대한 annotation을 출력할 수 있다.

(5) MIRNA

배추의 miRNA의 식별과 분류와 binding site를 예측한 문헌을 수집하고 각 문헌에서 제시된 miRNA간의 비교를 수행하였다. 수집된 모든 miRNA는 precursor sequence의 배추 표준 유전체 상의 물리적 위치를 기준으로 1차적으로 통합되었으며 precursor 상에서 mature miRNA가 발현되는 위치를 기준으로 2차 grouping을 진행하였다. 배추 표준 유전체 상에 존재하는 miRNA를 분류하고 물리적 위치를 기준으로 mature miRNA sequence의 unique position을 확인한 결과 총 1,055개의 position에 mature miRNA가 위치하며 이는 496개의 miRNA family로 분류가 가능한 것을 확인하였다.

그림 17. A) miRNA family인 miR164에 의한 검색 결과 B) 배추 표준 유전체 상의 특정 precursor sequence에 대한 검색 결과

miRNA에 대한 검색 옵션으로서 miRNA의 family와 miRNA가 구조내의 binding site를 통해 target으로 하는 배추 표준 유전자의 BRAD ID가 제공된다. miRNA family로 검색할 경우 keyword로 선택된 miRNA family의 precursor 서열을 독립적으로 coding하는 배추 표준 유전체 상의 각 위치 및 세부정보가 BrTED에서 부여한 ID 아래에 정렬되어 출력된다. 출력 화면에서 miRNA family의 하이퍼링크를 통해 miRBase(www.mirbase.org)의 관련 정보에 접근할 수 있다. 또한 reference 열의 하이퍼링크를 통해 각 mature miRNA의 정보를 생산한 기 보고된 문헌이 위치한 원문 페이지에 접근할 수 있다 (그림 17.A).

group 열의 BrTED에서 부여한 miRNA precursor의 group ID의 하이퍼링크에 접속하면 해당 group의 배추 표준 유전체 상에서의 위치를 확인할 수 있다. 또한 수집한 문헌에서의 mature miRNA의 관련 정보와 target gene에 대한 정보를 열람할 수 있다(그림 17.B).

(6) KEGG

KEGG 메뉴에서는 현재 알려진 41,020 개의 배추 표준 유전자에 최대한의 KEGG pathway annotation을 부여하였다(Kanehisa et al, 2017). 이를 통해 각 유전자가 어떠한 생체내의 pathway에 포함되어 활동하는 것에 대한 정보를 검색할 수 있도록 메뉴를 설계하였다. 외부 데이터베이스에서의 검색이나 R package 등을 사용한 BRAD에서 사용되는 배추 표준 유전자 ID로부터의 KEGG 정보 추출이 어려웠기 때문에 배추에 비해 상대적으로 pathway에 대한 정보를 쉽게 연결시킬 수 있는 애기장대의 TAIR ID를 매개로 배추의 표준 유전자 ID에 KEGG pathway ID를 연결하는 작업을 수행하였다.

KEGG annotation 이후, 애기장대에 대한 KEGG pathway ID 및 description, 그리고 배추 표준 유전자 ID를 keyword로 하여 KEGG 메뉴의 검색 옵션을 구성하였다. KEGG pathway 관련 keyword의 검색을 통해 검색 결과에서 선택한 pathway에 대한 ID와 정보 그리고 해당 pathway 내에 포함되는 배추 표준 유전자의 리스트가 table 상에 나타난다(그림 18).

검색 결과가 나타난 table 상의 첫 번째 열의 KEGG pathway ID를 선택하면 해당 pathway의 세부 정보를 출력하는 BrTED 내의 페이지를 열람할 수 있다. 이 pathway 정보 페이지는 이용자에 의하여 요청된 KEGG pathway ID에 대한 간략한 설명과 pathway map, 그리고 pathway에 포함되는 모든 배추의 표준 유전자 ID와 그에 대한 annotation을 포함한다. 페이지 상에서 이용자가 KEGG pathway ID를 클릭하면 하이퍼링크를 통해 KEGG에서 관련 정보를 확인 할 수 있으며, 특정 조건하에서 나타나는 KEGG pathway에 관련된 유전자들의 발현량을 전사체 샘플을 선택하여 확인 및 비교가 가능하다.

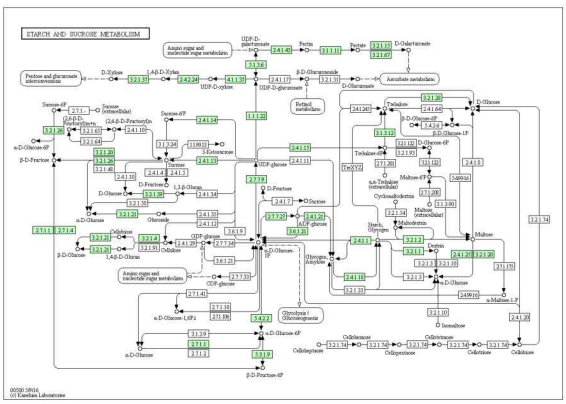
KEGG pathway ID list(Based on *A.thaliana*)

The results related your keyword: ath00500

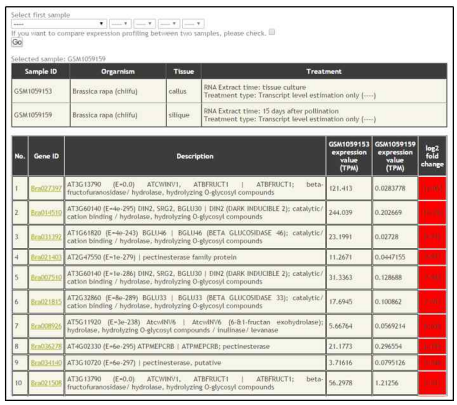
No.	Pathway ID	Name	Related genes in *B.rapa*
1	ath00500	Starch and sucrose metabolism < 344 >genes	Bra026644 Bra038548 Bra014992 Bra015239 Bra017955 Bra039576 Bra026500 Bra030291 Bra030289 Bra031204 Bra030256 Bra011928 Bra035672 Bra021650 Bra018248 Bra022839 Bra005550 Bra005551 Bra021815 Bra017257 Bra023053 Bra005269 Bra000081 Bra005091 Bra017110 Bra004590 Bra004834 Bra004835 Bra037651 Bra004837 Bra037650 Bra004839 Bra000382 Bra040320 Bra000385 Bra004535 Bra000780 Bra004447 Bra021403 Bra000457 Bra021438 Bra001157 Bra039421 Bra040532 Bra039419 Bra001320 Bra029755 Bra029756 Bra034060 Bra035508 Bra034140 Bra021486 Bra021487 Bra039138 Bra001040 Bra021466 Bra027421 Bra027398 Bra021508 Bra027397 Bra001547 Bra021548 Bra021549 Bra027350 Bra001605 Bra027210 Bra027211 Bra036447 Bra001025 Bra039149 Bra032004 Bra021291 Bra021292 Bra022440 Bra035821 Bra035782 Bra023886 Bra026178 Bra001923 Bra014987 Bra028348 Bra001937 Bra015025 Bra013229 Bra015104 Bra025376 Bra033083 Bra036219 Bra025392 Bra033098 Bra034960 Bra019455 Bra034983 Bra018180 Bra018177 Bra018171 Bra018172 Bra033795 Bra018152 Bra033792 Bra007007 Bra039799 Bra003286 Bra007332 Bra014620 Bra003357 Bra007420 Bra003378 Bra007452 Bra014540 Bra007483 Bra004833 Bra004840 Bra007507 Bra014511 Bra037647 Bra003401 Bra007508 Bra007509 Bra007510 Bra014510 Bra007573 Bra014453 Bra021571 Bra006217 Bra003511 Bra007705 Bra014390 Bra033195 Bra000704 Bra029428 Bra029444 Bra033116 Bra012676 Bra026230 Bra012642 Bra040180 Bra037344 Bra036278 Bra039420 Bra036279 Bra036282 Bra037315 Bra013557 Bra038755 Bra038756 Bra000461 Bra006755 Bra006756 Bra013234 Bra015096 Bra020858 Bra021441 Bra030064 Bra013628 Bra019370 Bra013756 Bra019249 Bra013819 Bra019214 Bra036881 Bra019043 Bra007698 Bra007699 Bra014395 Bra015094 Bra034697 Bra031105 Bra036699 Bra039025 Bra008091 Bra011194 Bra013531 Bra024075 Bra011249 Bra024022 Bra011428 Bra034553 Bra010616 Bra011789 Bra033560 Bra011882 Bra033591 Bra033592 Bra011881 Bra033604 Bra010658 Bra011835 Bra033634 Bra005825 Bra009417 Bra005833 Bra009406 Bra004202 Bra009398 Bra006047 Bra009057 Bra006090 Bra009830 Bra006111 Bra006129 Bra006130 Bra008926 Bra008748 Bra006301 Bra008696 Bra023519 Bra006303 Bra008695 Bra023520 Bra030352 Bra008572 Bra023600 Bra006395 Bra009521 Bra002221 Bra023713 Bra002257 Bra020059 Bra002289 Bra020096 Bra002332 Bra006578 Bra006587 Bra002346 Bra008398 Bra020139 Bra009747 Bra004838 Bra037648 Bra009938 Bra036610 Bra035347 Bra039501 Bra039502 Bra025395 Bra026264 Bra028234 Bra028430 Bra027452 Bra004836 Bra019555 Bra027873 Bra039970 Bra014349 Bra024910 Bra015135 Bra037495 Bra020658 Bra037432 Bra022549 Bra028254 Bra022552 Bra028256 Bra029170 Bra028257 Bra028278 Bra028279 Bra029158 Bra002978 Bra002963 Bra008618 Bra022739 Bra029010 Bra002891 Bra035579 Bra002673 Bra020389 Bra006722 Bra020325 Bra024311 Bra031912 Bra037782 Bra037799 Bra024359 Bra031879 Bra015444 Bra030634 Bra015468 Bra032442 Bra015497 Bra030651 Bra015529 Bra031537 Bra013692 Bra016838 Bra020848 Bra031771 Bra016812 Bra019749 Bra026984 Bra030511 Bra033348 Bra026010 Bra026011 Bra016552 Bra025667 Bra031036 Bra016357 Bra024568 Bra016328 Bra016276 Bra024709 Bra024708 Bra015364 Bra010872 Bra010873 Bra030510 Bra032585 Bra033347 Bra033352 Bra036338 Bra033344 Bra032841 Bra032842 Bra034400 Bra004433 Bra015599 Bra014254 Bra018850 Bra030490 Bra040563 Bra017888 Bra031526 Bra028383 Bra028384 Bra031392 Bra027030 Bra036652 Bra036653 Bra039757 Bra004168 Bra004054 Bra009265 Bra028699 Bra007906 Bra003845 Bra015995 Bra008230 Bra015784 Bra008344 Bra015631 Bra008366 Bra035049

그림 18. KEGG pathway ID ath00500에 관련된 배추 표준 유전자의 리스트

A



B



Sample ID	Organism	Tissue	Treatment
GSM1059153	Brassica rapa (chiifu)	callus	RNA Extract time: tissue culture Treatment type: Transcript level estimation only (----)
GSM1059159	Brassica rapa (chiifu)	silique	RNA Extract time: 15 days after pollination Treatment type: Transcript level estimation only (----)

No.	Gene ID	Description	GSM1059153 expression value (TPM)	GSM1059159 expression value (TPM)	log2 fold change
1	Bra027297	AT3G13790 (E=0.0) ATCWINV1, ATBFRUCT1 | ATBFRUCT1; beta-fructofuranosidase/ hydrolase, hydrolyzing O-glycosyl compounds	121.413	0.0283778	-12.1
2	Bra014599	AT3G60140 (E=4e-295) DIN2, SRG2, BGLU30 | DIN2 (DARK INDUCIBLE 2); catalytic/ cation binding / hydrolase, hydrolyzing O-glycosyl compounds	244.039	0.202669	-10.2
3	Bra031394	AT1G61820 (E=4e-243) BGLU46 | BGLU46 (BETA GLUCOSIDASE 46); catalytic/ cation binding / hydrolase, hydrolyzing O-glycosyl compounds	23.1991	0.02728	-9.7
4	Bra021403	AT2G47550 (E=1e-279) | pectinesterase family protein	11.2671	0.0447155	-8.0
5	Bra007516	AT3G60140 (E=1e-286) DIN2, SRG2, BGLU30 | DIN2 (DARK INDUCIBLE 2); catalytic/ cation binding / hydrolase, hydrolyzing O-glycosyl compounds	31.3363	0.128688	-7.9
6	Bra021815	AT2G32860 (E=8e-289) BGLU33 | BGLU33 (BETA GLUCOSIDASE 33); catalytic/ cation binding / hydrolase, hydrolyzing O-glycosyl compounds	17.6945	0.100862	-7.5
7	Bra008924	AT5G11920 (E=3e-238) AtcwINV6 | AtcwINV6 (6-B-1-fructan exohydrolase); hydrolase, hydrolyzing O-glycosyl compounds / inulinase/ levanase	5.66764	0.0569214	-6.6
8	Bra026279	AT4G02330 (E=6e-295) ATPMEPCRB | ATPMEPCRB; pectinesterase	21.1773	0.296554	-6.2
9	Bra034140	AT3G10720 (E=6e-297) | pectinesterase, putative	3.71616	0.0795126	-5.5
10	Bra021508	AT3G13790 (E=0.0) ATCWINV1, ATBFRUCT1 | ATBFRUCT1; beta-fructofuranosidase/ hydrolase, hydrolyzing O-glycosyl compounds	56.2978	1.21256	-5.5

그림 19. ath00500에 대한 정보 출력 결과, A) ath00500의 pathway 이미지, B) 서로 다른 두 조직(GSM1059153과 GSM1059159)에서 발현하는 ath00500의 관련 유전자들의 발현량 비교 결과의 일부

제5절 수박의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영

1. 수박의 육종 특화 데이터베이스 구축을 위한 기반 정보 수집

가. 데이터베이스 구축 기반 정보 수집을 위한 외부 데이터베이스의 탐색

(1) 수박의 표준 유전체 정보의 수집

4차 년도에는 수박의 육종 특화 데이터베이스 구성을 위해 관련된 모든 작업에 지속적으로 활용하기 위한 표준 유전체 정보를 우선적으로 확보하기 위해 Cucurbit Genomics Database(www.icugi.org/cgi-bin/ICuGI/index.cgi)의 Download 영역에서 2013년, 수박의 표준 유전체로서 공개된 계통명 97103의 유전체 정보를 수집하였다(Guo et al, 2013).

수박의 전사체 및 계통별 변이 정보의 수집과 정보의 재생산을 위해 NCBI의 Sequence Read Archive (SRA: https://www.ncbi.nlm.nih.gov/sra) 중심으로 데이터베이스의 기본 구성에 필요한 정보를 확보하였다. 본 보고서에서 기술할 모든 생물정보학적 작업은 본 실험실에서 채소 육종 특화 데이터베이스(168.188.15.201/vege)가 탑재되어 있는 서버에서 이루어졌고 지난 과제 수행 기간 동안 구축한 분석용으로 할당된 50TB의 저장공간, 1TB memory 그리고 32 CPU core를 system resource로 활용하여 수행되었다.

본 4차 년도에 구성된 수박의 육종 특화 데이터베이스는 현재, 개발 과정상의 보안에 의해 접속이 승인된 이용자에 한하여 온라인으로 접속 및 이용이 가능하다(URL: 168.188.15.201/vwatermelon).

표 1. 수박 유전체 관련 데이터베이스 구축을 위한 기반 정보 제공 데이터베이스

Database	Retrieved information	Address
NCBI GEO	Protein redundant sequence Raw reads for RNA-seq Raw reads for DNA-seq. NCBI ID related biological info	http://www.ncbi.nlm.nih.gov/geo/
Cucurbit Genomics Database	Genome sequence coding gene sequence Protein sequence Gene location annotation file	http://www.icugi.org/cgi-bin/ICuGI/genome/search.cgi
Pfam	protein domain annotation service	http://pfam.xfam.org/
NCBI FTP server	GO ID gene description Plant protein sequence fasta file KOG information	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2info ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plant/ ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/big_endian/Kog_Be.tar.gz
Gene Ontology consortium	EC information	http://www.geneontology.org/external2go/ec2go
Uniprot	Additional annotation information	http://www.uniprot.org/uploadlists/

(2) 수박 변이 정보의 수집

현재 수박 유전체에 대한 기존 연구로부터 공개된 계통별 변이 데이터를 얻기 위하여 수박의 Whole genome sequencing project (Guo et al, 2013)에서 계통명 97103을 기준으로 표준 유전체를 구성하면서 이를 reference로 하여 re-sequencing가 수행된 20계통의 genome re-sequencing 파일은 NCBI의 SRA에서 SRA052158로 검색 및 열람이 가능하다. 이를 분석용 서버로 다운로드한 이후, GATK를 중심으로 구축한 파이프라인을 통해 가공하여(그림 1) 수집된 계통들의 reference genome인 97103에 대한 InDel 및 SNP에 대한 변이 정보를 생산하고자 하였으며 수집된 계통과 그 sequence 정보는 표 2와 같다.

표 2. 변이 정보 생산을 위해 SRA로부터 수집한 20개 계통의 sequence 정보

No.	SRX ID	Accession name	SRA ID	library type	read length
1	SRX146279	JX-2	SRR494422	PAIRED	45
2	SRX146281	JXF	SRR494424	PAIRED	45
3	SRX146282	RZ-901	SRR494425	PAIRED	45
			SRR494426	PAIRED	45
4	SRX146283	XHBFGM	SRR494427	PAIRED	45
5	SRX146284	Black Diamond	SRR494428	PAIRED	45
6	SRX146285	Calhoun Gray	SRR494429	PAIRED	45
7	SRX146286	Sugarlee	SRR494430	PAIRED	45
8	SRX146287	Sy-904304	SRR494431	PAIRED	75
9	SRX146288	RZ-900	SRR494432	PAIRED	45
			SRR494433	PAIRED	45
10	SRX146289	PI482271	SRR494434	PAIRED	45
11	SRX146290	PI189317	SRR494441	PAIRED	90
12	SRX146291	PI500301	SRR494444	PAIRED	90
13	SRX146292	PI595203	SRR494439	PAIRED	45
			SRR494440	PAIRED	45
14	SRX146293	PI249010	SRR494443	PAIRED	90
15	SRX146294	PI248178	SRR494446	PAIRED	90
16	SRX146295	PI482276	SRR494437	PAIRED	45
			SRR494438	PAIRED	45
17	SRX146296	PI482303	SRR494442	PAIRED	45
18	SRX146297	PI296341-FR	SRR494435	PAIRED	45
			SRR494436	PAIRED	45
19	SRX146298	PI482326	SRR494445	PAIRED	45
20	SRX146280	JLM	SRR494423	SINGLE	88

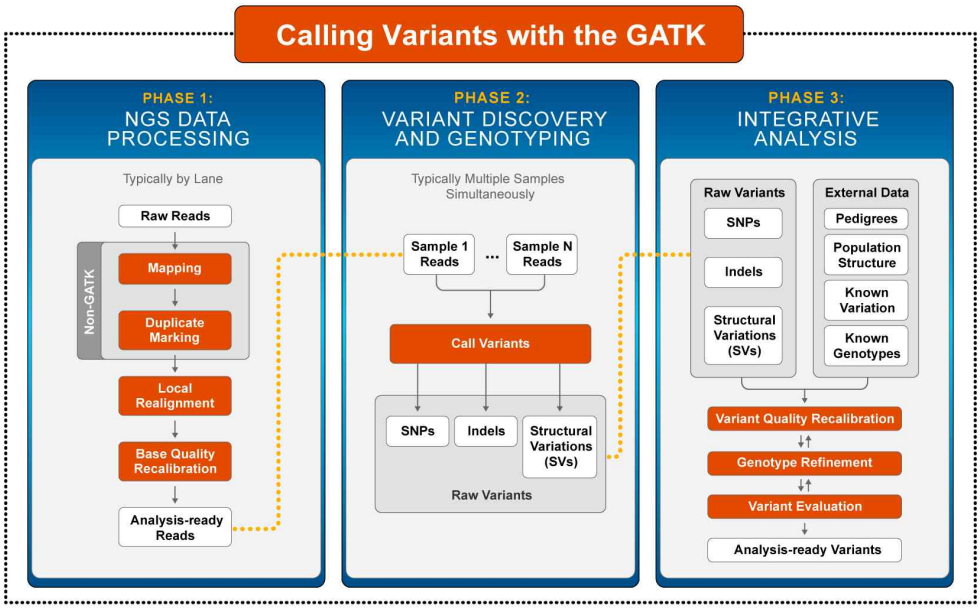


그림 1. GATK 파이프라인 개요

(3) 수박 전사체 정보의 수집

수박의 학명인 *Citrullus lanatus*를 키워드로 하여 NCBI Sequence Read Acrchive (http://www.ncbi.nlm.nih.gov/sra)에서 관련된 transcriptome 데이터의 목록 및 경로를 확인하였다. 이중에서 high-throughput RNA-seq 데이터를 리눅스 서버에서 wget 명령어 (wget -c /NCBI ftp 서버상의 파일경로/)를 사용하여 개별적으로 다운로드하고 후속 작업이 이루어질 특정 경로에 저장하였으며 확보한 26개의 데이터의 목록 및 기본정보는 표 3과 같다. 크게 웅성불임성 (GSE69703)과 (Li et al, 2015) 표준 계통인 97103 (SRP012849) 및 야생종 (SRP051354)의 과육(flesh) 및 과피(rind)에서의 화기의 수정이 발생한 이후의 시기별 유전자 발현량에 대한 RNA-seq 라이브러리를 구축한 작업 (Guo et al, 2015)으로 분류할 수 있었다.

표 3. 수집한 수박의 표현형 별 transcriptome의 RNA-seq 정보

*DAP: Day after pollination

Series	Sample ID	Line name	Tissue	Read type	status
GSE69703	GSM1692542 GSM1692543	DAH3615	floral bud flower	PAIRED	fertile
	GSM1692544 GSM1692545	DAH3615-MS	floral bud flower	PAIRED	sterile
SRP012849	SRR494406 SRR494407	97103	fruit flesh	SINGLE	10 DAP
	SRR494408 SRR494409				18 DAP
	SRR494410 SRR494411				26 DAP
	SRR494414 SRR494415				34 DAP
	SRR494416 SRR494417	97103	fruit rind	SINGLE	10 DAP
	SRR494418 SRR494419				18 DAP
	SRR494420 SRR494421				26 DAP
	SRR494412 SRR494413				34 DAP
SRP051354	SRR1724899	PI296341-FR	fruit flesh	SINGLE	10 DAP
	SRR1724900				18 DAP
	SRR1724901				26 DAP
	SRR1724902				34 DAP
	SRR1724903				42 DAP
	SRR1724943				50 DAP

2. 수박 유전체의 수집된 정보 재해석을 통한 생물정보의 생산

가. 수박 유전체에서 prediction된 23,440개 유전자의 Annotation

Cucurbit Genomics Database로부터 확보한 1.0 version 수박 표준 유전체는 전체 genome size가 425MB 수준이며 Whole genome sequencing project의 결과 330MB가 물리지도로서 현재 이용 가능하다. 또한 수박 표준 유전체는 총 11개의 개별적 chromosome과 하나의 scaffold (Chr0)로 구성되며 총 23,440개의 유전자가 예측이 된 것으로 알려져 있다. 수박 유전체의 모든 유전자의 염색체 상의 물리적 위치를 기재한 gff파일과 protein 및 nucleotide sequence fasta 파일을 통해 이에 해당하는 모든 수박 유전자의 기본정보를 얻을 수 있었으나 수박의 육종 특화 데이터베이스에서 annotation으로서 사용할 그 외의 GO (Gene ontology) ID, EC (Enzyme commission) number, gene description, KEGG orthologue, Pfam domain 정보등과 같은 부가적 정보들의 생산은 표 4에 나타난 프로그램 및 패키지를 통해 이루어졌고 수박 유전체의 annotation의 전반적인 작업 절차는 그림 1과 같다.

표 4. 예측된 23,440개 수박 유전자의 annotation에 사용된 program 및 tools

Program/tools	Purpose
BLASTP	NCBI에 등재된 protein sequence에 대한 homology search를 통해 수박 유전자 ID에 NCBI의 GI number 부여 별도의 KOG (euKaryotic Orthologous Groups) DB 구성을 통해 수박 유전자 ID에 대한 KOG 정보 부여
In home perl script	수박 유전자 ID에 부여된 GI number를 기준으로 GO ID, gene description,
Pfam batch search	주어진 수박 유전자 상의 domain을 식별
Uniprot Retrieve/ID mapping	수박 유전자 ID에 부여된 GI number를 기준으로 gene symbol, pubmed, panther ID 정보를 부여
KEGGREST	R 환경에서 수박 유전자 ID에 부여된 GI number를 기준으로 KEGG ID 부여

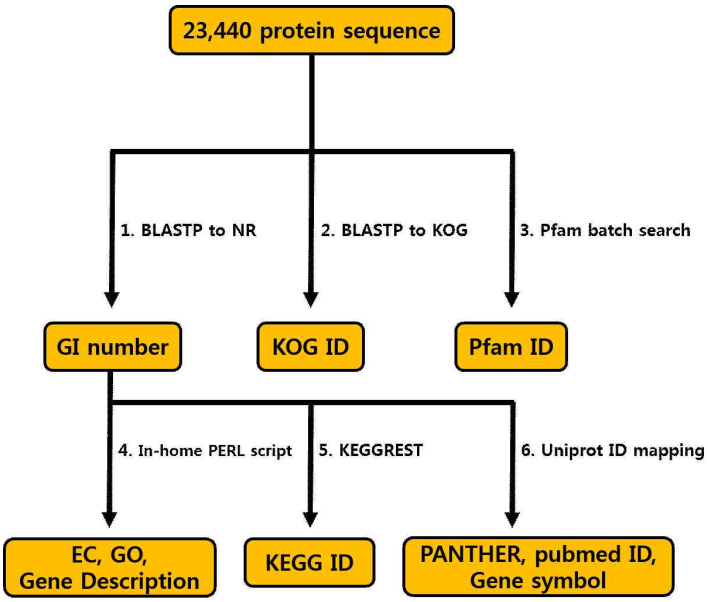


그림 1. 수박 유전자 annotation의 workflow

(1) BLASTP를 통한 NCBI GI number와 KOG ID의 부여

확보한 23,440 개의 수박 유전자 ID는 염기/단백질 서열 정보와 유전체 상에서의 위치만을 정보로서 갖기 때문에 모든 수박의 단백질 서열들을 NCBI에서 수집한 Non redundant (NR) database (식물 유래 데이터만 포함)와 KOG database에 대하여 BLASTP 처리하여 주어진 서열에 NCBI GI number와 KOG ID에 대한 정보를 얻을 수 있었다.

(2) Hash script를 통한 annotation (EC, GO, description)

Gene annotation에 반영할 기본 구성 요소의 수집은 NCBI의 ftp 서버 및 Gene ontology consortium의 웹페이지의 다운로드 페이지 상에 공개된 자료를 통해 수집이 가능하였으며 이를 리눅스 서버상의 vi editor로 열람 가능한 파일의 형태로 gene2go (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go: NCBI ID, Gene ontology number 와 그의

description 연관 관계 그리고 Biological process, Cellular component, Molecular function 여부), gene2info(ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2info: NCBI ID 및 gene symbol 과 gene description 연관 정보 기재), ec2go (http://www.geneontology.org/external2go/ec2go: EC number와 GO ID 연관 정보 포함)와 같은 파일명으로 저장하였다.

상술한 정보들은 각 파일내의 데이터에 해당하는 GI number를 갖고 있기 때문에 이 연관관계를 이용하여 BLASTP를 통해 수박 유전자에 대응된 GI number에 EC, GO, gene description을 연결시킬 수 있다. 이를 구성하기 위해 자체적으로 perl script를 개발하였으며 이는 PERL 언어의 Hash 기능을 통해 공통된 유전자 ID에 부여된 GI number를 기준으로 이에 대응되는 EC, GO, gene description에 대한 정보를 gene2go, gene2info, ec2go에서 불러와 주어진 유전자 ID에 부여하는 방식으로 설계되어 주어진 기능을 수행한다.

(3) Uniprot customized annotation의 추가

BLASTP에 의해 GI number가 부여된 수박 유전자 ID들을 대상으로 얻어진 GI ID를 Uniprot의 ID mapping service(www.uniprot.org/uploadlists/)에 입력하여 (그림 2) 사용자에 의해 지정된 annotation을 얻고 이를 text file의 형태로 다운로드 받을 수 있다. GI number에 의해 Uniprot database로부터 추출할 annotation은 Organism, PANTHER ID, Pubmed ID, Gene symbol로 설정하였으며 output을 다운로드 받은 이후, 이를 리눅스 분석 서버로 이관하여 SED 명령어를 통해 organisms정보가 식물계에 포함되지 않는 행을 제거하였다.

그림 2. Uniprot의 Retrieve/ID mapping service 페이지

(4) KEGG Orthology의 추가

수박 유전자 23,440개의 NR에 대한 BLASTP를 통해 얻은 x개의 GI number를 text file로 만들고 이를 R package인 KEGGREST에 input으로 불러왔다. 그리고 R 환경내에서 for loop문을 구성하여 input 파일내의 GI number를 순차적으로 package내에 입력하여 그에 해당하는 KEGG ID를 package 내의 library로부터 추출하여 GI number와 KEGG ID 사이의 연결 관계를 output으로 생산하였다(그림 3).

```
#!/usr/bin/Rscript

setwd(getwd())

library(KEGGREST)
print("package has been loaded.")
path <- getwd()
file_ID <- paste("uniq_gi_number_list",".txt",sep="")
gi_list <- read.table(file_ID,header=FALSE)
total_num <- as.numeric(nrow(gi_list))

for(i in 1:total_num){

target <- gi_list[i,1]
print(target)
info <- keggFind("genes",target)

outputname <- paste(target,".txt",sep="")

write.table(info, file=outputname, sep="\t", quote=FALSE, col.names=FALSE, row.names=FALSE)
}
```

그림 3. KEGG 정보 추출을 목적으로 작성된 R script

(5) Pfam scanning을 통한 수박 유전자상의 domain 식별

23,440개 유전자 ID의 단백질 서열상의 domain을 식별하기 위하여 단백질 서열 fasta 파일을 sed 명령어를 통해 8개의 fasta 파일로 나누어 각 파일을 순차적으로 Pfam(pfam.xfam.org/search)의 batch search 페이지에 입력하였다 (그림 4).

입력한 데이터의 결과는 입력한 사용자의 메일 계정으로 회신되고 이로부터 얻은 8개의 결과 파일을 텍스트 형식으로 저장 후, 리눅스 분석 서버로 이관하여 cat 명령어로 하나로 취합하였다. vi editor에서 불필요한 행 및 중복되는 행을 삭제함으로써 수박의 각 유전자 상에 존재하는 도메인들의 종류와 Pfam accession number 정보를 얻을 수 있었다(Finn et al, 2016). 또한 상술한 정보에서 선택적 열 추출과 정렬 및 중복 정보를 가진 행의 제거를 통해 수박의 각 유전자 상에 존재하는 도메인이 3,811개의 고유 Pfam accession으로 구분되는 것을 확인하였다.

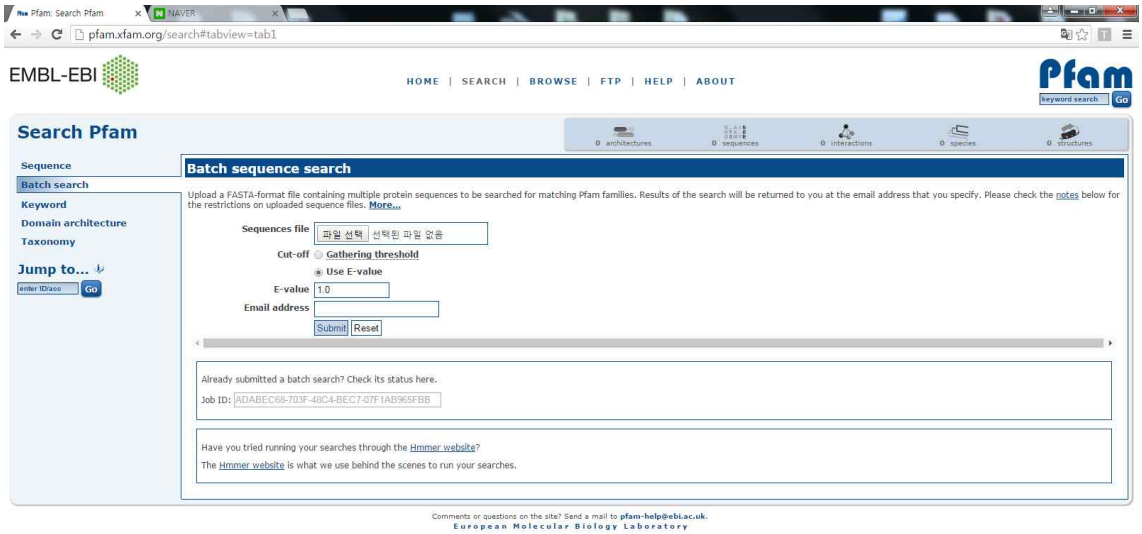


그림 4. Pfam Batch search의 웹페이지

나. 수박 20개 계통의 re-sequencing 데이터의 분석을 통한 변이 정보의 재생산

(1) Re-sequencing 데이터의 Quality check 와 trimming

Watermelon genome sequencing project에 의해 NCBI SRA에 업로드된 20 계통에 대한 read 데이터를 fastqc 처리하여 각 계통의 genomic read의 quality를 확인한 결과 모든 계통에 대하여 정제된 read 데이터를 공개한 것으로 판단되어 별도의 trimming 과정을 수행하지 않고 각 계통별 read 데이터를 수박 표준 유전체에 alignment 시키는 작업을 진행하였다.

(2) Burrows-Wheeler Aligner(BWA)에 의한 genomic read의 reference로의 mapping 과 Variant calling

각 계통에 대한 read 데이터는 BWA를 통해 (Li and Durbin, 2009) reference genome에 mapping 되었으며 mapping 결과로 sequence alignment map(SAM) 형식의 파일을 얻었다. 이는 picardtool에 의해 후속 연산에 불필요한 duplicated read들을 mark하고 이를 data pool에서 제거한다. 남은 read 데이터는 GATK에 의해 계통별 read 데이터를 수박 표준 유전체 서열에 대조하여 SNP와 InDel에 대한 데이터를 생산되어 bcftool에 의해 variant calling format(VCF) 형식으로 전환된다. 얻어진 VCF 파일은 하나의 파일 내에 특정 계통이 보이는 SNP와 InDel에 대한 표준 유전체 상의 위치를 전부 갖기 때문에 이를 vcftool을 사용하여 SNP에 대한 VCF 와 InDel에 대한 VCF로 분리하였다.

(3) SNP matrix의 구성과 변이의 평가(evaluation)

BWA와 GATK 파이프라인을 거친 20계통의 genomic DNA로부터 얻은 변이 데이터는 GATK package에서 제공하는 CombineVariants 기능을 통해 하나의 VCF 파일로 통합되었다. 이를 대상으로 표준 유전체의 정보를 활용하여 SNP 위치에 대한 annotation을 수행하는 SNPeff를 통해 수행되었다(Cingolani et al, 2012).

SNPeff에 의한 수박 20계통의 변이 분석 결과는 html 형식으로 출력되고 이를 리눅스 분석 서버에서 윈도우스 운영체제로 전송 후 파일 내의 정보를 시각적으로 확인할 수 있다. 그리고 텍스트 형식으로 제공되는 파일을 통해 SNP가 발생한 유전자와 그 구조상의 위치, 유전자 구조상에 SNP 발생에 의한 codon이 지정하는 아미노산의 변경 효과를 확인할 수 있었다.

Number of effects by type and region

Type

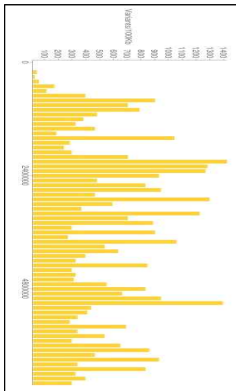
Type (alphabetical order)	Count	Percent
downstream_gene_variant	1,525,203	20.139%
initiator_codon_variant	18	0%
intergenic_region	3,811,317	50.325%
intron_variant	429,895	5.676%
missense_variant	64,640	0.854%
splice_acceptor_variant	165	0.002%
splice_donor_variant	143	0.002%
splice_region_variant	10,027	0.132%
start_lost	232	0.003%
stop_gained	944	0.012%
stop_lost	352	0.005%
stop_retained_variant	164	0.002%
synonymous_variant	72,917	0.963%
upstream_gene_variant	1,657,347	21.884%

Region

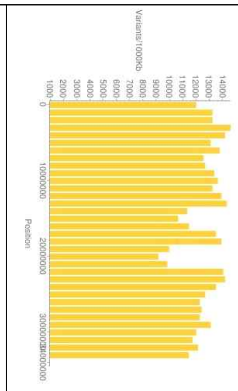
Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,525,203	20.167%
EXON	138,064	1.826%
INTERGENIC	3,811,317	50.394%
INTRON	421,987	5.58%
SPLICE_SITE_ACCEPTOR	165	0.002%
SPLICE_SITE_DONOR	143	0.002%
SPLICE_SITE_REGION	8,803	0.116%
UPSTREAM	1,657,347	21.914%

그림 5. SNPeff를 활용한 수박 20계통으로부터 얻은 SNP 변이에 대한 평가

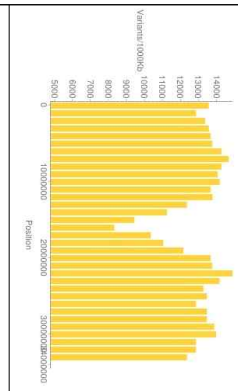
표 5. 수집한 수박 20 계통의 SNP 변이 데이터에서 얻은 유전체 상의 SNP 분포



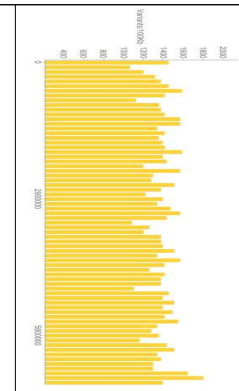
Chr0	
Length	24,621,398bp
SNPs	148,214
SNP rate	166bp/snp



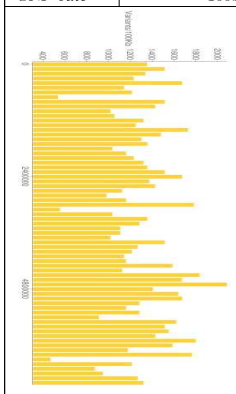
Chr1	
Length	34,083,085bp
SNPs	432,284
SNP rate	78bp/snp



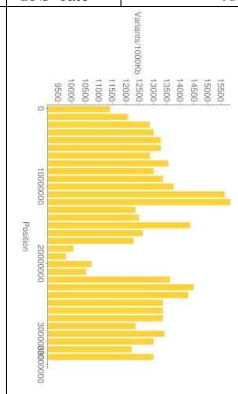
Chr2	
Length	34,414,252bp
SNPs	449,756
SNP rate	76bp/snp



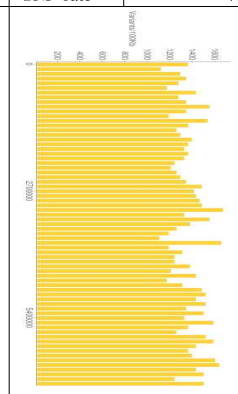
Chr3	
Length	28,939,167bp
SNPs	378,957
SNP rate	76bp/snp



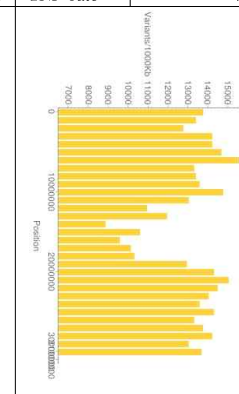
Chr4	
Length	24,315,960bp
SNPs	314,534
SNP rate	77bp/snp



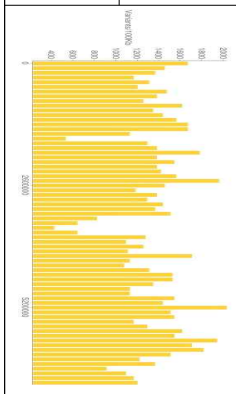
Chr5	
Length	33,714,806bp
SNPs	435,472
SNP rate	77bp/snp



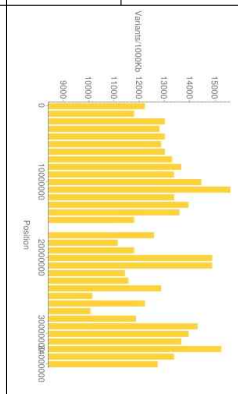
Chr6	
Length	27,018,480bp
SNPs	322,876
SNP rate	83bp/snp



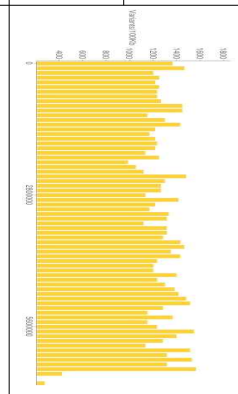
Chr7	
Length	31,477,646bp
SNPs	413,031
SNP rate	76bp/snp



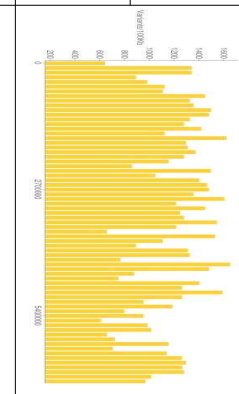
Chr8	
Length	26,149,438bp
SNPs	332,301
SNP rate	78bp/snp



Chr9	
Length	34,986,854bp
SNPs	449,950
SNP rate	77bp/snp



Chr10	
Length	28,419,553bp
SNPs	359,298
SNP rate	79bp/snp



Chr11	
Length	27,106,780bp
SNPs	344,286
SNP rate	78 bp/snp

다. 수박 RNA-seq 데이터를 활용한 조직 및 시기별 전사체 데이터의 생산

(1) 수집한 수박 RNA-seq 데이터의 quality control

수집한 모든 수박 전사체 발현에 관련된 실험(GSE69073, SRA052198, SRP051354)에 포함되는 총 26개의 RNA-seq 결과에 대하여 개별적으로 fastqc로 read의 quality를 다양한 측면에서 확인하고 모든 경우에 대하여 기본적으로 phred scrore가 30 미만으로 나타난 염기서열은 fastx-toolkit의 fastq-trimmer로 quality trimming을 수행하였다.

그러나 수박 표준 계통인 97103의 과피 및 과육에 대한 전사체 샘플들(SRA052198)의 경우 모든 샘플들로부터 전반적인 overrepresented sequence(O.S)에 의한 오염이 나타났다. 일반적으로 O.S는 RNA-seq 라이브러리의 제작과정에서 상용되는 illumina의 어댑터 sequence로 이들에 대한 정보가 fastqc에 입력되어 있어 O.S에 대한 출처를 fastqc 이후 확인할 수 있으나 이 경우 모든 출처가 No hit로 나타났다. O.S에 대한 출처를 찾기 위해 모든 fastqc에서 나타난 O.S를 리눅스 분석 서버상에 한 파일내에 통합하고 이를 sort하고 unique한 sequence를 filtering하여 NCBI BLAST(blast.ncbi.nlm.gov/Blast.cgi)에서 NR을 데이터베이스로 하여 BLASTN을 수행하였다.

BLAST 결과 입력된 O.S는 zucchini virus의 구성 sequence로 나타났다. 얻어진 모든 O.S를 fasta 파일 포맷으로 데이터베이스화시키고 이를 Trimmomatic을 이용하여 O.S에 대한 trimming을 별도로 수행하여 QT를 완료 후 fastqc에서 O.S가 샘플별로 전체 read data pool에서 차지하는 비중이 현격이 줄어든 것을 확인 후 이를 tophat을 통한 reference mapping에 활용하였다.

(2) 수박 유전체 상의 intron length 분포에 대한 평가

Quality check가 완료된 read를 reference genome 상에 alignement를 수행할 경우 실제 alignment를 수행하는 tophat의 인트론 사이즈에 대한 paramerter를 적정한 값을 최대값으로 설정하여 작업을 수행하지 않으면 설정한 intron size의 최대값을 넘어서는 intron size를 갖는 read sequence들이 reference genome상에 align되지 못하고 그 데이터를 unmapped 된 상태로 남게 되기 때문에 가용데이터(read count)손실이 최대화된다. 따라서 genome 상의 intron size의 분포를 가시화하여 이를 확인하고 tophat에서 최대 intron size로 허용할 최적값을 채택함으로써 작업 속도 및 aligned sequence 데이터 산출의 효율화를 기대할 수 있다. watermelon_v1.gff의 파일에서 mRNA를 포함한 행을 sed 명령어를 통해 제거하고 새로운 파일을 작성한 후 perl script를 통해 위아래 행의 유전자 ID가 같은 경우 두 행이 보유한 첫 번째 exon이 끝나는 위치와 두 번째 엑손인 시작하는 엑손의 위치를 차한 값을 intron size로서 인출하고 이 값이 입력받은 한계치 이상인 경우를 세어 그 총합을 입력한 한계치 이상의 인트론이 수박의 전체 유전체 상에 있는 수로 나타내어 한계치 x 이상의 *C. lanatus* 유전체 상의 분포를 확인하고자 하였으며 그 결과는 그림00과 같다. tophat 작업에서는 intron max size를 3000으로 설정하였다.

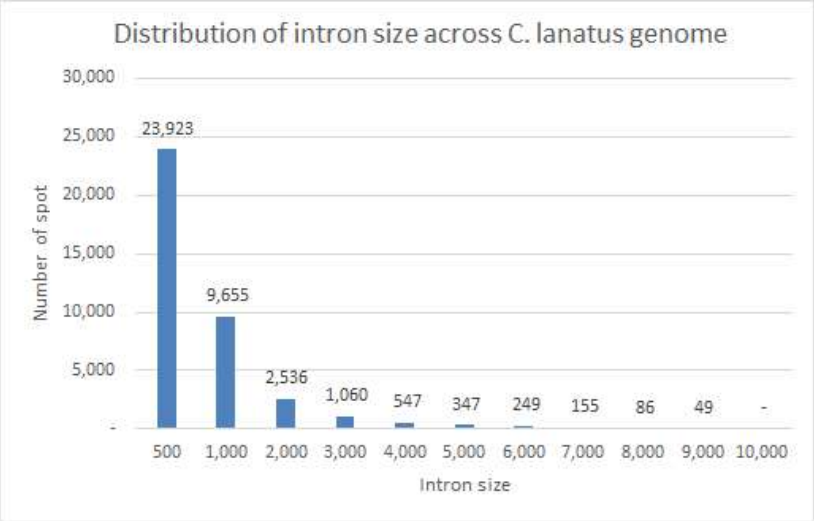


그림 6. 수박 유전체의 intron 분포의 시각화

(y축의 값은 x축에서 나타난 수치 이상의 intron size를 갖는 수박의 reference 상의 위치의 개수를 의미함.)

(3) tophat을 통한 read 데이터의 reference mapping

수박 표준 계통 97103의 1.0 version에 해당하는 whole genome sequence와 유전체 상의 수박의 예측된 유전자의 물리적 위치를 나타낸 gtf 파일을 tophat에 입력하고 각 RNA-seq 샘플별로 QC와 QT가 완료된 read 데이터에 대한 reference mapping을 진행하였다. 수박 zucchini virus에 의한 O.S를 높은 강도로 제거한 SRA052198의 경우 그 오염의 수준이 샘플별로 큰 차이를 보였다(그림 7. 표 6).

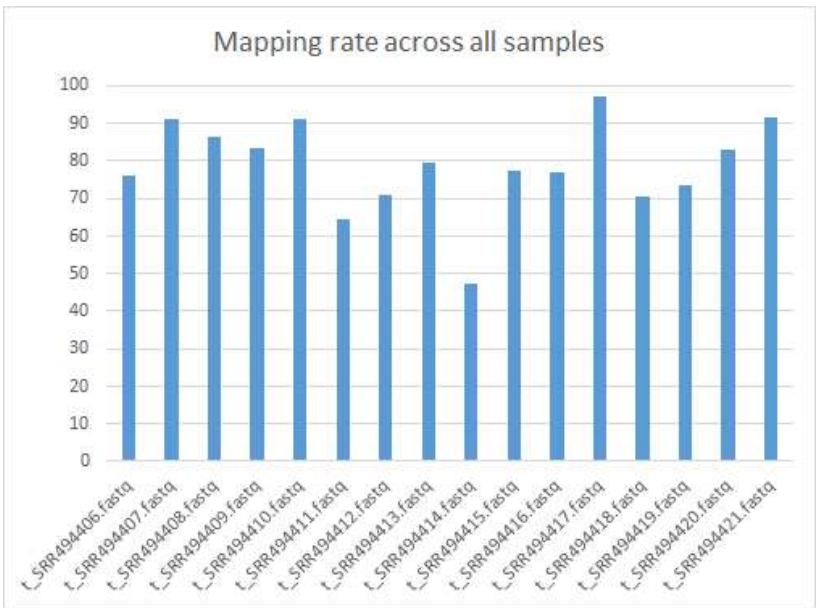


그림 7. SRA052198의 샘플들의 reference에 대한 mapping rate

표 6. SRA052198의 trimming 전후 read의 수 비교와 QT이후의 reference에 대한 mapping rate

File name	before trimming	trimmed input reads	mapped reads	unmapped reads	mapping rate(%)
t_SRR494406.fastq	10131218	7877824	5984341	1893483	76
t_SRR494407.fastq	10752201	8772609	7989279	783330	91.1
t_SRR494408.fastq	18914328	14858686	12846438	2012248	86.5
t_SRR494409.fastq	12077551	9390995	7847970	1543025	83.6
t_SRR494410.fastq	13588792	11374261	10385148	989113	91.3
t_SRR494411.fastq	12345625	9304807	6013959	3290848	64.6
t_SRR494412.fastq	9778317	7206857	5122472	2084385	71.1
t_SRR494413.fastq	11572988	8788377	6992884	1795493	79.6
t_SRR494414.fastq	10306366	6420885	3051152	3369733	47.5
t_SRR494415.fastq	8148638	6088521	4723697	1364824	77.6
t_SRR494416.fastq	8647278	6546653	5041324	1505329	77
t_SRR494417.fastq	8488793	7209221	7009907	199314	97.2
t_SRR494418.fastq	10154906	7489916	5290924	2198992	70.6
t_SRR494419.fastq	12837948	9563049	7049248	2513801	73.7
t_SRR494420.fastq	10861345	8446482	7012177	1434305	83
t_SRR494421.fastq	11755860	9593041	8784694	808347	91.6

(4) 수박 표준 유전자의 조건별 발현량 데이터의 산출과 normalization

tophat에 의해 산출된 실험별 RNA-seq의 mapping 결과를 구분하여 DESeq의 manual에서 제공하는 파이프라인을 이용하여 bam 파일을 R 환경으로 load하였다. 수박 표준 유전체의 유전자의 exon 위치정보를 활용하여 각 유전자의 coding 영역에 mapping 된 read의 수를 세어 각 bam 파일 별로 유전자별 발현량을 산출한 이후 동일 유전자의 발현량에 대한 샘플간 비교를 위해 TPM normalization을 모든 샘플에 대해 수행하였다.

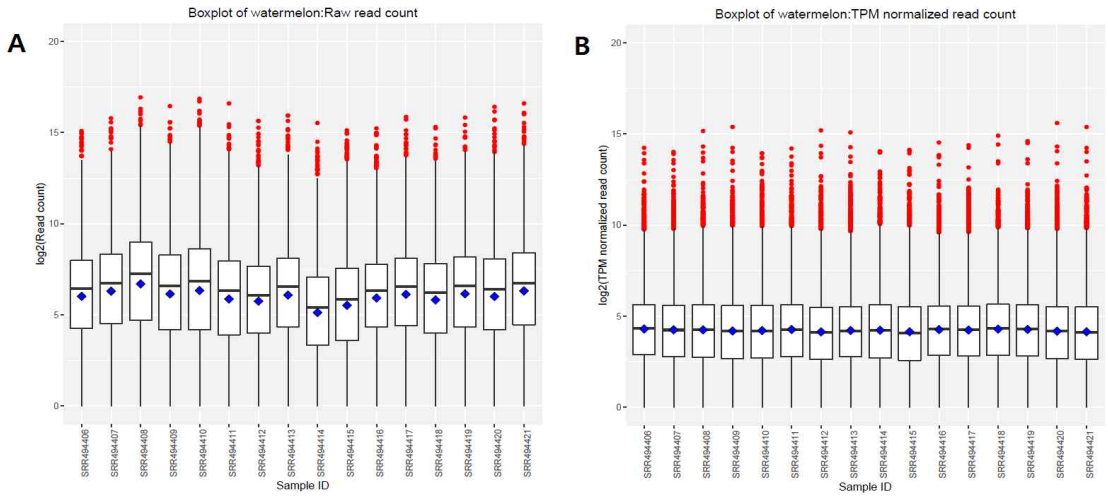


그림 8. SRA052198의 샘플별 발현량에 대한 boxplot A) Read count B) TPM normalization

3. 수박 분자 육종 활성화를 위한 특화 데이터베이스 구축

가. 수박의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영

현재 웹상에 공개된 수박의 표준 유전체 정보를 확보하고 이를 다양한 외부 생물정보 데이터베이스의 단백질 정보와 연결지어 최신 정보를 반영한 수박 유전체 상에 존재하는 23,440개에 대한 annotation을 수행하였다. 또한 이를 활용하여 다양한 계통의 re-sequencing 데이터로부터 SNP/InDel 변이 데이터를 재생산하고 과피, 과육 그리고 화기로부터 구성한 전사체 데이터로부터 수박 표준 유전자에 대한 발현량을 산출하였다. 4년차 과제의 수행과정에서 얻어진 모든 수박에 관련된 생물정보를 취합하였고 이는 수박의 종자개발을 위한 육종 특화 통합 DB의 기초 자료로서 활용되었으며 지난 3년차까지의 과정에서 구성한 분석 및 시각화 플랫폼을 적용하여 입력된 정보를 이용자 친화적인 인터페이스를 통해 열람할 수 있도록 하였다.

(1) Genome browser 기반의 유전체 정보의 시각화

현재, 수박의 육종특화 데이터베이스에서는 확보한 수박 표준 유전체의 whole genome sequence와 유전체 상의 유전자의 위치 정보를 이용하여 Genome browser를 구현하였다. Genome browser를 조작하여 이용자가 원하는 수박 유전체의 특정 영역을 탐색할 수 있으며 이 영역 상의 유전자를 선택하여 annotation 작업으로부터 얻은 sequence, physical position, GI number, Gene description, GO number, KOG ID, 그리고 KEGG ID 정보를 기본 유전자 정보로서 열람할 수 있다. 또한 해당 유전자의 조직 및 시기별 발현량을 normalized read count data로부터 확인할 수 있으며 수집한 20 계통이 선택된 유전자에서 보이는 변이를 시각화하여 화면 하단의 SNP matrix와 함께 제공되며 수집된 계통의 일부만을 선택하여 re-align이 가능하다.

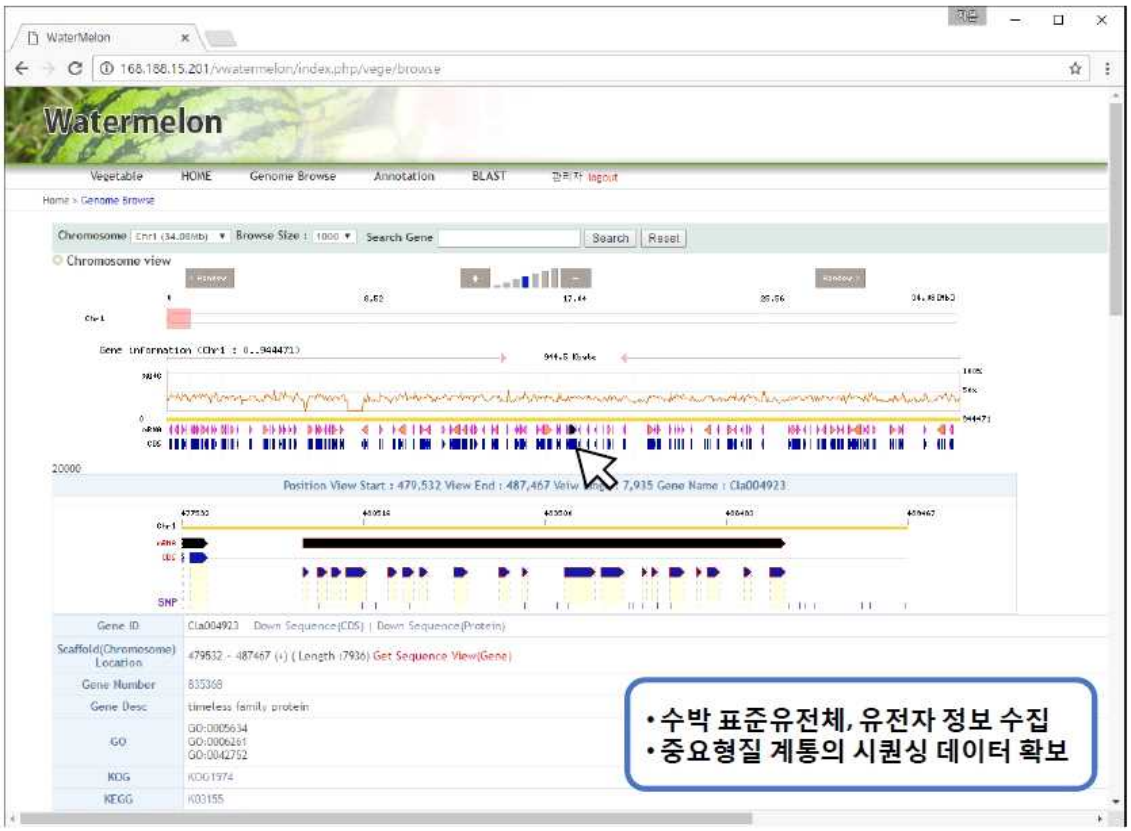


그림 9. 수박 표준 유전체 정보에 기반 Genome browser

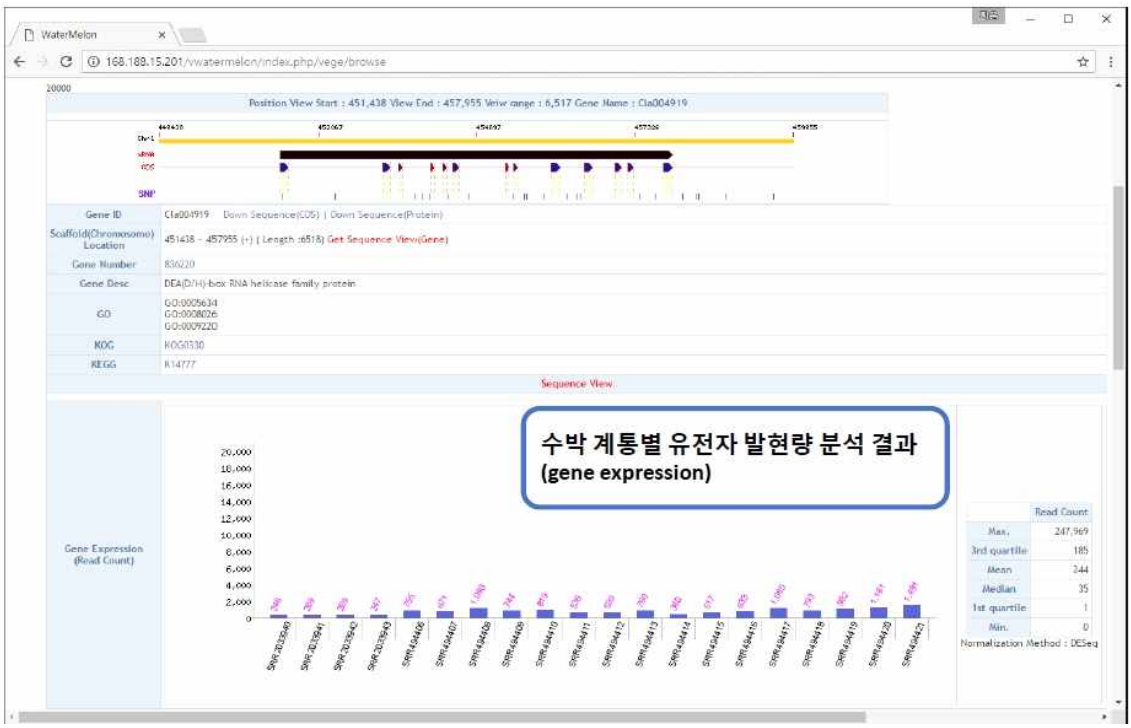


그림 10. 선택 유전자의 조건별 expression profiling

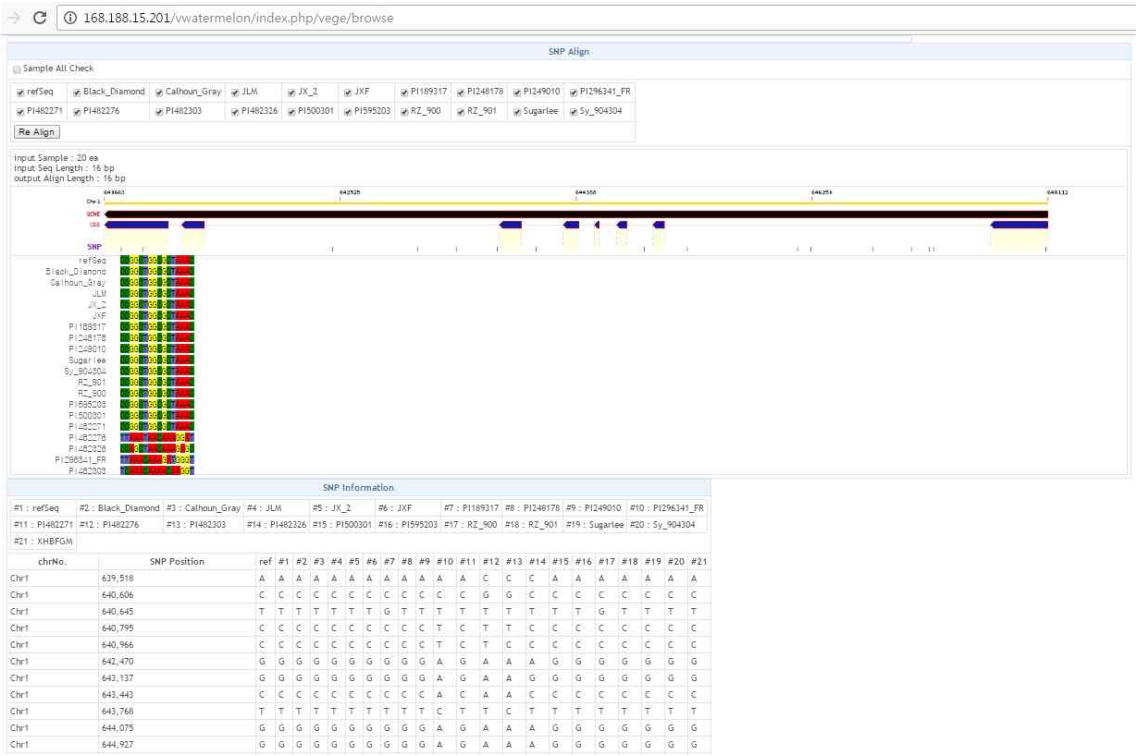


그림 11. 선택된 유전자의 계통별 변이의 시각화와 SNP matrix

(1) 23,440개 수박 유전자의 annotation 결과 출력 페이지

확보한 수박 표준 유전체 상에서 예측된 23,440개 수박 유전자에 대한 annotation 결과는 표의 형태로 정리되어 공개되어 있으며 9개의 keyword를 통해 이용자가 원하는 그룹의 유전자 및 특정 유전자의 검색이 가능하다. Gene number, PFAM, KOG 열에 해당하는 annotation은 외부 데이터베이스의 해당 정보와 하이퍼링크로 연결되어 있어 annotation 테이블에서 외부 데이터베이스에서 다루는 정보를 얻을 수 있다.

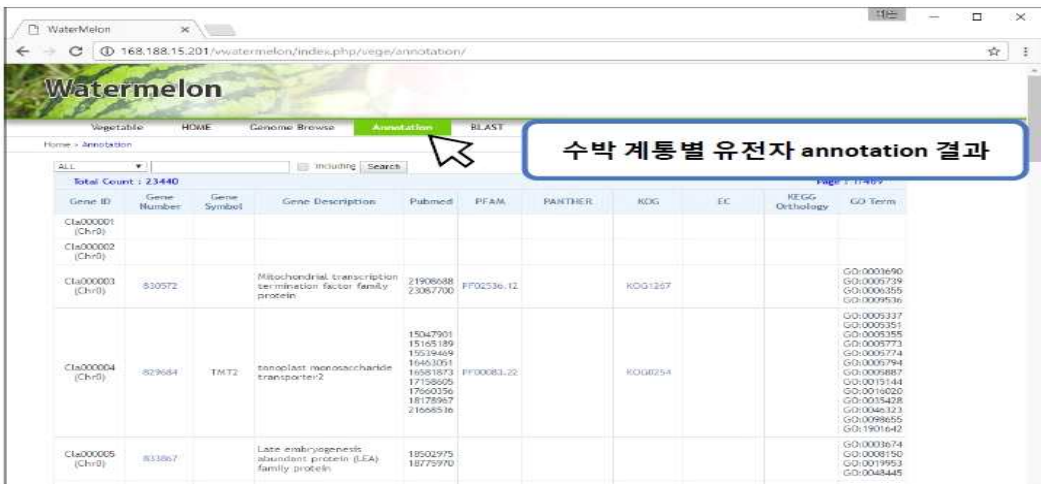


그림 12. 수박 표준 유전체의 annotation 테이블

(2) 입력 서열의 homologous gene search를 위한 web-BLAST

BLAST 프로그램을 웹페이지 상에 포함시켜 이용자가 보유한 염기서열 혹은 단백질 서열을 query로 활용하여 수집한 수박의 표준 유전체 정보를 BLAST에 대한 데이터베이스로 활용하여, web-BLAST는 사용자가 보유한 염기서열의 유전체 상의 위치 식별 및 homologous gene의 탐색을 지원한다.

BLAST의 데이터베이스로서 수박 표준 유전체의 whole genome sequence, coding gene sequence, protein sequence의 1.0 version이 활용되었으며 웹페이지에서의 BLAST에 대한 심화 기능 분석을 위해 BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX가 지원되었다. 그리고 sequence 입력란 하단의 일련의 BLAST 관련 parameter를 설정할 수 있게 함으로써 사용자에 의한 염기서열 분석의 자유도를 극대화할 수 있도록 하였다.

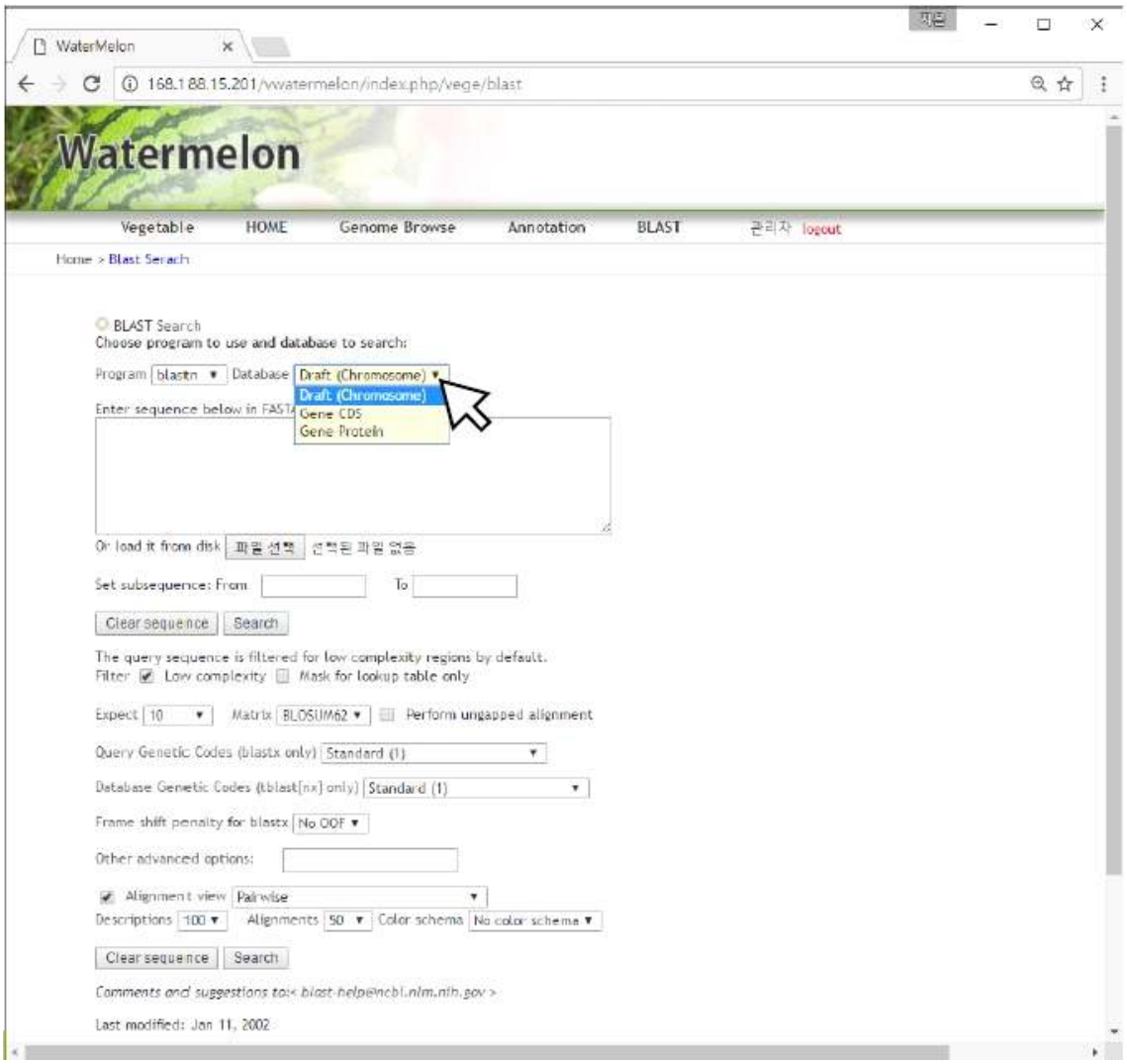


그림 13. 수박 표준 유전체에 대한 입력서열의 서열 유사도 검사를 위한 web-BLAST

제 4 장 목표 달성도 및 관련 분야에 대한 기여도

제1절 목표 달성도

1. 연차별 연구 목표 및 내용

가. 1차 년도

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
1차년도 (2013)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**배추의 데이터베이스 구축을 위한 기반 정보 수집**	100	배추 표준 유전체, 유전자, 분자마커를 수집하여 데이터베이스를 구축할 수 있도록 재가공
			100	기존의 연구를 통해 확보한 유전체 정보를 수집하여 데이터베이스 구축을 위한 가공
			100	프로젝트 수행을 위한 육종 소재의 정보 수집
			100	배추의 스트레스 저항성 및 내병성 형질 관련 조사를 통한 관련 유전자 및 분자마커 정보의 DB화
		배추 유전체 정보의 재가공/재해석	100	배추 표준 유전체의 지놈 브라우저 구현
			100	유전체 정보와 유용 형질 관련 유전자 및 분자마커 정보 연동
			100	재분석을 통한 교배친 및 우수자원 100계통의 대량 SNP 발굴
		배추 분자육종 활성화를 위한 특화된 데이터베이스 구축	100	배추 분자육종 활성화를 위한 특화된 데이터베이스를 구축하기 위한 구성요소, 제공 내역, 사용자 편의성을 고려하여 유전체 기반의 데이터베이스 디자인
			100	배추 유용 형질 정보 DB 구축

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
1차년도 (2013)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**이용자가 직접 분석할 수 있는 사용자 중심의 데이터베이스 구축**	100	배추 지놈 브라우징 시스템 구축
			100	배추 유용 형질 관련 유전자 정보 검색 및 유전자 서열 추출 시스템 구현
			100	유전자 서열 유사도 검사를 위한 BLAST 시스템 구축
		DB 및 시스템 구축을 위한 제반환경 구현	100	분석용 서버 구축 (CPU: 64 core, RAM: 1TB, DISK: 40TB)
			100	DB용 서버 구축 (CPU: 16 core, RAN: 128GB, DISK: 20TB)
			100	백업용 서버 구축 (CPU: 8 core, RAM: 64GB, DISK: 200TB)

나. 2차 년도

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
2차년도 (2014)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**배추의 데이터베이스 구축을 위한 기반 정보 수집**	100	배추의 목표 형질을 연구할 수 있는 육종 소재 및 형질 조사 내용 수집
			100	배추의 내병성 형질 관련 조사를 통한 관련 유전자 및 분자 마커 정보 수집
			100	환경 스트레스 형질 관련 조사를 통한 관련 유전자 및 분자마커 정보 수집

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
2차년도 (2014)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**배추의 DB를 구축하기 위한 중요 형질 유전체 정보 생산**	100	배추 RIL의 resequencing 데이터 생산
			100	유용 형질 관련 배추의 transcriptome 데이터 분석
		배추 유전체 정보의 재가공/재해석	100	배추 RIL의 resequencing 정보 분석을 통한 대량 SNPs 발굴
			100	유용 형질 관련 배추의 transcriptome 데이터 분석
			100	유용 유전자원의 발현량 정보 조사
			100	배추 유전체의 유용 형질 관련 유전자 구조 정보 도식화
			100	transcriptome 정보를 이용한 형질 관련 유전자 발굴
			100	유전체 정보를 이용한 형질 관련 배추 transcriptome의 발현량 정보 도식화
		배추 분자육종 활성화를 위한 특화된 데이터베이스 구축	100	배추 RIL의 genetic map DB 구축
			100	배추 RIL 집단을 이용한 SNP, In/Del DB 구축
			90	목표형질 선발용 분자마커(MAS) 데이터베이스 구축
			100	Genome-wide SNPs를 이용한 여교잡 선발용 분자마커 (MAB) DB 구축
		이용자가 직접 분석할 수 있는 사용자 중심의 데이터베이스 구축	100	형질 관련 유전자 기능 유추를 위한 배추과 유전자 homology 정보 검색 DB 구축
			100	형질 관련 마커 예측을 위한 eQTL 분석 시스템 구축
			100	대량 SNPs 발굴 정확도 측정을 위한 mapping viewer 구축

다. 3차 년도

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
3차년도 (2015)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**배추의 데이터베이스 구축을 위한 기반 정보 수집**	100	배추의 형질별 육종 소재에 대한 정보 DB화
			100	문헌 조사를 통한 형질 관련 유전자 및 분자마커 정보의 DB화
		배추의 DB를 구축하기 위한 중요 형질 유전체 정보 생산	100	배추 내병성 계통 resequencing 데이터 생산
			100	무, 배추 유용 형질 Transcriptome 데이터 생산
			100	배추 유용 형질 표현형 조사
		배추 유전체 정보의 재가공/재해석	100	배추 RIL의 resequencing 데이터 분석
			90	LD 분석을 통한 배추 대량 SNPs 마커의 신뢰도 평가
			100	대량 마커를 이용한 품종간 유전체 비교(GWAS) 연구
			100	Transcriptome 정보를 이용한 형질 관련 유전자 발굴
			100	유전체 정보를 이용한 형질 관련 배추 transctiptome의 발현량 정보 도식화
			100	배추 RIL의 genetic map의 정확도 평가
		배추 분자육종 활성화를 위한 특화된 데이터베이스 구축	100	배추 RIL의 genetic map DB 구현
			100	목표형질 선발용 분자마커(MAS) 데이터베이스 구축
			100	Linkage Disequilibrium(LD)을 이용한 배추 RIL의 유전체 재조합 정보 분석 및 Haplotype 정보 도식화 DB 구축
			100	Transcriptome 정보를 이용한 형질 관련 유전자 DB화
			100	유전체 정보를 이용한 형질 관련 배추 transcriptome의 발현량 정보 DB화

구분 (연도)	세부프로젝트명	세부연구목표	달성도 (%)	연구개발 수행내용
3차년도 (2015)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**이용자가 직접 분석할 수 있는 사용자 중심의 데이터베이스 구축**	100	형질 관련 유전자 기능 유추를 위한 배추과 유전자 homology 정보 검색 DB 구축
			100	유전자 정보 도식화를 위한 유전자의 conserved domain 정보 검색 DB 구축
			100	육종 가속화를 위한 형질연관 분자마커 선별(MAS, MAB) 시스템 구축
			100	형질 관련 마커 예측을 위한 eQTL 분석 시스템 구축
			100	대량 유전자 비교 synteny 분석을 통한 종간 유사도 측정
			90	LD, Haplotype을 이용한 육종시 필요 종자수 계산 시스템 구축
			100	대량 마커를 이용한 품종간 유전체 비교 (GWAS) 시스템 구축

라. 4차 년도

구분 (연도)	세부프로젝트명	세부연구내용	달성도 (%)	연구범위
4차년도 (2016)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**데이터베이스 구축을 위한 기반 정보 수집**	100	·각 프로젝트에서 연구된 수박 표준 유전체, 비교 유전체 정보 수집 및 데이터베이스화를 위한 정보 재가공 ·각 프로젝트에서 연구된 수박 유용 형질 정보 조사 ·각 프로젝트에서 연구된 형질 관련 유전자 및 분자마커 정보 수집 및 데이터베이스화를 위한 정보 재가공

구분 (연도)	세부프로젝트명	세부연구내용	달성도 (%)	연구범위
4차년도 (2016)	**채소작물의 종자개발을 위한 육종 특화 통합 DB 구축 및 운영** • 5대 채소 작물의 분자육종을 지원할 수 있는 특화된 데이터베이스의 기반 구축 • 배추를 모델로 웹 데이터베이스 디자인 및 지놈 브라우저와 genome-wide SNP의 DB화	**DB를 구축하기 위한 중요 형질 유전체 정보 생산**	100	·유용 형질 부/모본 유전체 resequencing 데이터 생산 ·유용 형질 관련 transcriptome 데이터 생산 혹은 수집
		유전체 정보의 재가공/재해석	100	·박과작물 표준 유전체의 지놈 브라우저 구현 ·유전체 정보와 유용 형질 관련 유전자 및 분자마커 정보 연동 ·유용 유전자원의 발현량 정보 조사 ·유전체의 유용 형질 관련 유전자 구조 정보 도식화 ·Re-sequencing 정보를 이용한 대량 SNPs 마커 발굴
		분자육종 활성화를 위한 특화된 데이터베이스 구축	85	·수박 유전체 DB 구축 ·수박 유용 형질 정보 DB 구축 ·목표형질 선발용 분자마커(MAS) 데이터베이스 구축 ·Linkage Disequilibrium(LD)을 이용한 RIL의 유전체 재조합 정보 분석 및 Haplotype 정보 도식화 ·Transcriptome 정보를 이용한 형질관련 유전자 발현 DB 구축
		이용자가 직접 분석할 수 있는 사용자 중심의 데이터베이스 구축	90	·수박 지놈 브라우징 시스템 구축 ·유용 형질 관련 유전자 정보 ·검색 및 유전자 서열 추출 시스템 구현 ·형질 관련 마커 예측을 위한 ·eQTL 분석 시스템 구축 ·분자마커(SNPs, CAPS) 디자인 및 마커선발 정확도 검증을 위한 in silico PCR 시스템 구축
		DB 및 시스템 구축을 위한 제반환경 강화	100	백업용 서버 구축(DISK: 50TB)
		타작물의 웹데이터베이스(Web DB) 자료 업데이트	80	·기존의 구축된 타 작물의 기반정보 수집 및 ·데이터베이스화를 위한 정보 재가공 ·타 작물의 구축된 Web DB의 추가 자료를 계속하여 업데이트

제2절 관련분야의 기여도

1. 기술적 측면

○ 육종가의 경험적인 역량으로 선발과 교배가 이루어져 왔던 전통육종에 NGS (차세대 염기서열분석)를 통해 발굴된 대량 분자마커를 활용하여 현대 육종의 진화를 도모함.

○ 분자육종에 전문화된 데이터베이스를 구축하여 작물 육종에서 필요로 하는 형질에 관련된 유전자의 분석된 정보를 제공함으로써, 분자마커 개발에 소요되는 시간을 단축하여 효율적이고 차원 높은 분자육종 시스템을 구축하는데 기여할 수 있음.

○ 육종 기술에 대한 사용자 중심의 데이터베이스 및 분석 지원 시스템을 통하여 검증된 분자마커의 정보를 국내 육종공동체로의 보급하여 신품종 육종과정에서 체계적으로 활용이 가능함.

○ 5대 작물에 대한 분자마커 육종지원시스템 개발을 통해 분자마커를 이용하는 육종기술의 체계화 및 타 작물에 대한 유사 시스템 기술지원이 가능함.

○ 작물 육종 활동에 대한 GWAS 분석 방식의 접목은 다양한 표현형 중 특정 형질에 대한 신규 분자마커의 개발을 집단 작성 및 전개에 소요되는 시간과 비용을 크게 절감할 수 있음.

2. 경제적·산업적 측면

○ 복잡 형질의 유전 현상을 시스템적으로 이해하고 예측 가능한 육종 계획을 세움으로 육종 기간을 단축하고 맞춤육종(breeding-by-design)을 구현할 수 있는 기반을 제공함.

○ 공통된 작물을 육종하는 여러 육종가에게 동일한 분자마커의 사용을 유도함으로써 육종 연한의 감소를 이룩함.

○ 다양한 작물 종을 대상으로 연구하는 분자육종가에게 공통의 분자마커 사용을 촉진함으로써, 신품종 육성 가속화로 인한 육종 기간 단축과 분자마커 개발 및 파생 기술을 이용한 산업화를 통하여 다가오는 바이오경제의 경쟁력을 확보할 수 있을 뿐만 아니라 관련 민간기업의 활성화에 기여할 수 있음.

○ GWAS 분석 시스템의 구축은 관심 형질에 대한 candidate gene 상의 유전자 SNP를 신속하게 분자마커화 할 수 있기 때문에 이를 활용한 육종과 신품종의 생산 및 시장 진출을 유도하는 기반 연구의 초석을 마련함.

○ MAB와 MAS를 통해 신품종 육성에 대한 최적의 교배 조합과 필요 개체수의 예측으로 육종 효율성을 극대화시켜 우리 종자의 세계시장에서의 경쟁력을 확보하고 종자 수출량의 증대를 기대.

○ 시장 동향 파악을 통해 빠르게 변화하는 세계 종자시장의 수요에 대한 맞춤형 품종 개발의 효율성을 극대화 할 것으로 기대.

제 5 장 연구개발 성과 및 성과활용 계획

제1절 연구개발 성과

1. 연차별 연구성과 목표 및 달성

(단위: 건 수)

구분		논문		특허		인력 양성	데이터베이스 구축 사례	분자 마커	기타
		SCI	비SCI	출원	등록				
1차 년도	목표								
	달성			1			1		
2차 년도	목표	1							
	달성	1		2	2			1	GMO 마커검정
3차 년도	목표		1						
	달성	7		4	1	2		6	GMO 마커검정
4차 년도	목표	1					1		
	달성	3			3	3	1		
계	목표	2	1						
	달성	11	0	7	6	5		7	2

2. 논문게재 성과

○ 논문게재 SCI 11건

순번	발간연도	논문명	주저자	학술지명	Vol (No)	구분
1	2014	Genetic Detection of Clubroot Resistance Loci in a New Population of *Brassica rapa*	Wenxing Pang, Shan Liang, Zhongyun Piao	Horticulture, Environment, and Biotechnology	55(6)	SCI
2	2015	Quantitative trait loci mapping of partial resistance to Diamondback moth in cabbage (*Brassica oleracea* L)	Nirala Ramchiary, Yong Pyo Lim	Theoretical and Applied Genetics	128(6)	SCI
3	2015	Mapping QTLs of resistance to head splitting in cabbage(*Brassica oleracea* L. var. *capitata* L.)	Wenxing Pang, Xiaonan Li, Yong Pyo Lim	Molecular Breeding	35(5)	SCI
4	2015	The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases.	Arne Schwelm, Yong Pyo Lim, Jutta Ludwig-Müller, Christina Dixelius	Scientific Reports	10 (1038)	SCI
5	2015	Construction of chromosome segment substitution lines enables QTL mapping for flowering and morphological traits in *Brassica rapa*.	Xiaonan Li, Yong Pyo Lim, Zhongyun Piao	frontiers in plant science	6	SCI
6	2015	The 2015 KSM-ICWG-GSP Joint Clubroot Symposium Meeting Report.	Vignesh Dhandapani, Yong Pyo Lim	Journal of Plant Growth Regulation	34(2)	SCI
7	2015	Anatomic Characteristics Associated with Head Splitting in Cabbage (*Brassica oleracea* var. *capitata* L.).	Wenxing Pang, Yoon-Young Kim, Yong Pyo Lim	PLOS ONE	10(11)	SCI
8	2015	Genomic and Post-Translational Modification Analysis of Leucine-Rich-Repeat Receptor-Like Kinases in *Brassica rapa*.	Jana Jeevan Rameneni, Yeon Lee, Man-Ho Oh, Yong Pyo Lim	PLOS ONE	10(11)	SCI
9	2016	Quantitative Trait Loci for Morphological Traits and their Association with Functional Genes in *Raphanus sativus*	Xiaona Yu, Su Ryun Choi, Yong Pyo Lim	frontiers in plant science	7	SCI
10	2016	Genome-Wide Analysis and Characterization of Aux/IAA Family Genes in *Brassica rapa*	Parameswari Paul, Vignesh Dhandapani, Yong Pyo Lim	PLOS ONE	11(4)	SCI
11	2016	Genome wide identification and functional prediction of long non-coding RNAs in *Brassica rapa*	Parameswari Paul, Yong Pyo Lim	Genes & Genomics	38(6)	SCI

3. 특허 성과

○ 특허실적 13건 (특허출원 7건, 특허등록 6건)

순번	구분	출원여부	년도	특허명	출원인	발명인	출원(등록)번호
1	특허	출원	1	칼슘 함량이 증가된 배추 품종 및 이의 육종방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-2013-0153649
2	특허	출원	2	배추좀나방 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-2014-0062420
3	특허	출원	2	열구 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-2014-0062417
4	특허	등록	2	배추좀나방 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-1474910
5	특허	등록	2	열구 저항성 양배추 품종을 선별하기 위한 프라이머 세트, 방법 및 키트	충남대학교	임용표, 나종현, 최수련, 방문성	10-1474914
6	특허	출원	3	열근 무 품종을 구분하기 위한 특이 마커 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-2015-0015252
7	특허	등록	3	칼슘 함량이 증가된 배추 품종 및 이의 육종방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1557420
8	특허	출원	3	배추 글루코시놀레이트 함량 조절인자 기반의 SNP 마커 및 이의 용도	충남대학교	임용표, 이소남, 비그니쉬 단다파니, 최수련, 강동현, 정소영, 박선규	10-2015-0190309
9	특허	출원	3	초장이 짧은 야생무로부터 초장이 긴 개량무 품종을 특이적으로 구분하기 위한 프라이머 세트 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-2015-0188806
10	특허	출원	3	해조류 추출물을 이용한 위타니아 솜니페라 유래의 스테로이드계 락톤의 생산 방법	충남대학교	임용표, 시와난단 가내산	10-2015-0142520
11	특허	등록	4	글루코브라씨신 함량이 증가된 배추 품종 및 이의 육종 방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1602022
12	특허	등록	4	글루코나스튜틴 함량이 증가된 배추 품종 및 이의 육종 방법	충남대학교	임용표, 박종태, 최수련, 조만현, 함인기, 김태일, 이은모	10-1602536
13	특허	등록	4	열근 무 품종을 구분하기 위한 특이 마커 및 이의 용도	충남대학교	임용표, 최수련, 우효나, 이수희, 선해정	10-1627197

4. 인력활용/양성 성과

순번	양성일자	연구기관	학위	성별	국적	비고
1	2015.1.12	충남대학교	석사	남	대한민국	
2	2015.12.10	충남대학교	석사	남	대한민국	
3	2016.2.25	충남대학교	박사	여	인도	
4	2016.2.25	충남대학교	석사	남	대한민국	
5	2016.8.25	충남대학교	석사	남	대한민국	

5. 데이터베이스 구축 성과

순번	구축년도	연구기관	데이터베이스명	URL	비고
1	2013	충남대학교	채소작물의 육종 활성화를 위한 데이터베이스	www.vegetable.or.kr	2017년 기준, 생산 및 수집된 배추와 수박의 육종관련 정보가 반영
2	2016	충남대학교	BrTED(Brassica rapa Transcriptome Expression Database)	brted.cnu.ac.kr	2017년 DATABASE지에 submit 예정

제2절 성과 활용계획

1. 배추 육종 특화 데이터베이스 구축을 위한 육종 기반 정보의 생산 및 수집

채소작물의 육종에 활용할 수 있는 성과로서 5편의 SCI급 논문의 게재와 12건의 특허실적을 달성하였다. 해당 성과들은 과제 수행을 통해 구축한 데이터베이스를 통해 연구자 및 육종가들의 접근이 가능하도록 서비스하여 국내의 채소작물의 분자육종의 활성화를 실현할 수 있도록 할 것이다. 또한 과제 수행 중의 연구에서 개발한 분자마커 및 관련정보들을 분자육종 기술이 부재한 종자회사에게 마커 검정 서비스를 제공할 수 있을 것으로 기대된다.

가. 배추 수집단 표현형 정보 조사 결과

본 실험실에서는 3년에 걸쳐 배추 수집단의 생육조사를 수행하였다. 해당 결과는 23개의 표현형에 대한 정량적 및 정성적 결과를 가지며 정성적 결과의 경우 index에 따른 값을 부여 받아 정량화 되었다. 이 생육조사 결과는 전산화되어 23개 표현형에 대한 유전연구의 부모본 선발과 mapping 집단의 전개 및 GWAS의 표현형 정보로서 활용되어 표현형 관련 candidate gene의 식별에 활용될 수 있을 것으로 기대된다.

나. 배추 유용형질의 분자마커 수집 및 재생산

기보고된 유용 표현형관련 분자마커가 지정하는 유전자의 위치는 현재 공개된 배추의 표준 유전체 정보를 통해 쉽게 접근이 가능하다. 수집한 분자마커의 세부사항을 확인하고, 본 실험실에서 작성한 배추 수집단의 표현형 및 유전적 변이 정보와 수집단의 계통별로 추출한 DNA를 활용하여 기존의 분자마커를 SNP 범용마커로 전환할 수 있다. 개발 및 검증이 종료된 범용마커는 육종공동체에 보급되어 신속한 목표 형질의 선발 및 신품종 육성에 기여할 수 있다.

2. 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스 구축을 위한 생물정보의 생산과 재가공

가. 채소작물 표준 유전체 정보를 활용한 gene annotation

현재까지 알려진 배추 표준 유전자 41,020개와 수박 표준 유전자 23,440개에 대한 gene annotation을 수행하였다. gene annotation에는 KEGG, KOG, GO, PANTHER, uniprot annotation 등 배추와 수박 genome 데이터베이스에서 다루지 않는 추가적인 annotation을 부여하여 사용자가 추가적인 검색 없이 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스를 허브로 삼아 관련 정보를 쉽게 열람할 수 있다. 향후 본 데이터베이스에서 다루는 5대 채소작물의 표준 유전체 정보가 갱신되면 이를 기반으로 재생산된 데이터 또한 분석 을 신규 유전체 정보를 반영하여 재처리하여 갱신된 분석정보를 사용자들에게 제공할

계획이다.

나. 배추 수집단의 re-sequencing을 통한 GWAS 분석 기반의 구축

현재 배추 201계통 및 수박 20계통의 변이 정보가 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스에 반영되어 있다. 이를 통한 변이 정보가 재생산되어 배추와 수박의 수집 집단 내의 변이 정보를 현재 이용가능한 상황이다. 배추의 경우 3년간의 생육조사를 통한 23개의 표현형 정보가 전산화되어 있다. 배추 201 계통의 SNP 정보와 표현형 정보를 GAPIT을 통해 분석하면 표현형별 GWAS가 가능하다. GAPIT의 분석 결과인 맨하탄 플롯과 SNP의 표현형 연관 분석 정보를 통해 표현형에 강하게 연관된 SNP를 보유한 유전자를 식별하고 이를 분자마커로 전환할 수 있다. 이를 통해 GWAS 기반의 형질 관련 SNP 마커의 대량 발굴 체계를 구축할 것이다. 또한 생산된 SNP 마커의 검정 체계를 통해 신규 SNP의 검정력이 평가를 받는 기준을 제시하여 이를 충족하는 분자마커의 특허출원을 실시할 것이다. 특허 출원 및 등록이 완료된 신규 분자마커의 정보는 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스의 분자마커 정보 검색 체계에 편입되어 육종가들에 의해 활용될 수 있도록 할 계획이다.

다. 채소작물의 유용 유전자 식별을 위한 RNA-seq 기반 DEG 연구

채소작물의 조직 및 조건에 따른 전사체 발현양상의 비교분석을 통해 비교 조합에 따른 DEG의 식별 체계가 채소종자의 육종 특화 데이터베이스와 배추 전사체 데이터베이스에 갖추어져 있다. 이를 통해 특정 형질에 관련된 기보고된 연구결과와 데이터베이스에서의 DEG 산출을 통한 candidate gene의 리스트를 쉽게 얻을 수 있다. 추후 DEG를 기반으로 한 네트워크 분석을 도입하여 표면적으로 보이는 표현형에 관련된 candidate gene을 조절하는 패스웨이 상의 upstream에 존재하는 유전자를 규명할 것이다. 그리고 GWAS 구축 과정에서 얻은 표현형별 변이 데이터와 RNA-seq 분석 결과를 통합하여 서로 다른 표현형을 보이는 계통간에서 특정 유전자상의 염기서열이 나타내는 변이가 유전자의 발현 양상을 지배할 수 있다는 것을 규명하고자 한다. 그리고 이러한 유전자상의 변이를 범용 마커로 전환할 수 있는 체계를 구성하여 단순 표현형 연관 마커를 넘어선 유전자 발현 원리에 기반을 둔 근원적인 분자마커 생산을 위한 체계를 구성하고자 한다.

3. 생산 및 수집한 생물정보에 기반한 채소작물의 육종 특화 데이터베이스 구축과 운영

가. 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스

현재, 채소작물의 종자개발을 위한 육종 특화 통합 데이터베이스는 5대 작물에 대한 기본 플랫폼이 완성되었고 1차에서 3차 년도까지 배추에 대한 데이터 구성 및 입력이 완료된 상태이다. 그리고 4차 년도에는 배추의 사례를 기반으로 수박으로부터 표준 유전체의 기본 정보와 annotation, 계통별 변이 정보 및 전사체 발현 정보를 입력한 상태이다. 입력된 정보는 지놈 브라우저를 중심으로 구성한 스키마에 따라 출력되며 BLAST 및 키워드를 통한 검색이 가능하다. 그러나 아직 수박에 대한 데이터 입력의 절대량과 분석 및 검색 기능이 배추의 사례에 비해 미진한 부분이 있다. 따라서 차후의 작업으로 배추와 수박의 데이터베이스 간의 완성도에 대한 차이를 좁혀나갈 것이며 이후 무, 고추, 그리고 파프리카에 대한 정보 생산과 수집을 실시하여 5대 채소 작물에 대한 데이터베이스 구성을 종결하고 향후 새로운 자료가 확보하는 즉시 이를 데이터베이스의 data pool에 반영해나갈 것이다. 이로써 국내의 채소 분자육종 공동체에 각 작물에 대한 육종 관련 정보를 통합적으로 제공하여 신규 마커의 개발과 신품종 육성의 촉진을 기대할 수 있다.

나. 유용 형질의 DEG 기반 전사체 데이터베이스

배추(*Brassica rapa*)를 대상으로 한, 총 10개의 공개된 실험의 92개 전사체 데이터들이 수집되었다. 이를 통해 다양한 조건들에 대한 유전자의 발현량이 산출되었으며 이를 활용한 DEG 분석으로 KEGG와 GO에 대한 enrichment 분석이 데이터베이스내에서 실용화되었다. 이로서, 연구자 및 분자육종가들이 각자가 목적으로 하는 형질 관련 후보 유전자의 선정에 DEG 분석 결과를 사용할 수 있게 되었다. 앞으로도 BrTED는 새롭게 생산된 발현량 데이터를 지속적으로 생산 및 수집하여 더 많은 data pool을 연구 및 육종 공동체에 제공할 수 있도록 할 것이다. 또한 DEG 분석방식을 정밀화하여 더욱 정확한 분석 결과를 제공할 수 있도록 할 것이며 이를 기반으로 한 네트워크 분석을 도입할 예정이다.

제 6 장 연구개발과정에서 수집한 해외과학기술정보

1. 유전연구에 이용되는 polymorphism(SNP, InDel)

Single Nucleotide Polymorphism(SNP)는 DNA 염기서열에서 단일염기의 다형성을 의미하는 용어로서 다른 다형성 유형(SSR, InDel 등)에 비해 유전체에서 가장 많이 존재하는 것으로 평가받는 다형성의 형태이다. 이는 유전체 전반에 걸쳐 분포하며 그 수가 방대하여 특정 유전자에서 allelic variation 구분이 가능하여 각종 진단을 위해 사용되고 있고, 육종에서 형질 연관 마커로서 뿐 아니라 고밀도 유전자지도 작성용으로도 주목 받고 있다. Soybean 유전체 연구 에서는 고밀도 유전자지도 작성 및 Scaffold를 연결하는데 SNP 마커를 이용하였고 *B. rapa* 에서는 등이 EST 염기서열로부터 대량의 SNP를 발굴하여 유전자지도를 작성하였다. 식물유전체에서 SNP의 빈도에 대한 보고는 각기 다양한 데 *B. napus*의 경우 coding region에서 1 SNP/2.1Kb, non-coding region에서 1 SNP/1.2Kb 라고 보고하였고, *B. rapa*에서는 12.6 SNP/Kb 라고 보고하였는데 *B. napus* 와의 이러한 차이는 유전체 전반을 대변하는 데이터가 아니라 *B. rapa*와 *A. thaliana*의 EST 정보를 이용하여 유전자 밀집영역만을 분석하였기 때문으로 해석되었다. Park 등은 8종의 *B. rapa* 를 대상으로 *B. rapa* genome 전반에 걸쳐 분포하는 SNP/InDel 을 분석한 결과 SNP의 빈도는 exon 영역에서 1 SNP/103bp, intron 영역에서는 1 SNP/54bp 였고 InDel 의 빈도는 exon 영역에서는 1 InDel/2.2Kb, intron 영역에서는 1 InDel/135bp 였다고 보고하였다. 한편, Brassica A genome의 pseudochromosome 을 만들기 위해서도 InDel 마커가 이용되었다. Wang 등은 InDel 마커로 작성된 유전자지도와(RCZ16_DH) 동시에 3종의 *B. rapa* 유전자지도(VCS_DH, JWF3P, CK_DH) 정보를 이용하여 scaffolds의 위치를 확인하고 결정하였다. 모두 4종의 유전자지도에서 총 183개의 scaffolds 가 연결되어 전체 252Mb 길이의 물리지도를 보고하였다. 이러한 결과로부터 배추과에서 SNP와 InDel 은 마커로 유용하게 이용될 수 있으며, 특히 coding region에서 고밀도의 SNP/InDel 마커 정보는 *B. rapa* 뿐 아니라 배추속의 다른 근연종에서 농업적으로 중요한 형질들과 관련된 후보유전자의 탐색에도 유용하게 이용될 것으로 기대된다.

2. 현대의 작물 육종 지원 체계

배추에서의 분자마커의 개발 및 육종 및 연구로의 이용은 1980년 후반부터 활성화된 것으로 알려져 있다. 이후, RAPD, RFLP, AFLP, SSR, 그리고 SNP 등 다양한 형태의 분자마커들이 개발되어 배추과 작물의 유전적 다양성 및 육종을 위한 선발(Marker Assisted Selection)에 사용되어 왔다. 국내에서는 현재 포장에서의 교배에 대한 노동력을 절감을 실현할 수 있는 자가불화합성, 웅성

불임성 및 추대, 병저항성 등 다양한 주제에 초점을 맞추어 분자마커의 개발과 육종이 이루어지고 있다.

현제 배추를 비롯한 다양한 경제 작물에서 수행되고 있거나 종료된 표준 유전체 생산 프로젝트는 작물 육종 환경에 일종의 지도로서 활용될 수 있는 표준 유전체 정보를 제공함으로써 분자마커의 제작과정의 효율화에 큰 영향을 미치고 있다. 이를 통해 작물 육종에서 표준 유전체 정보는 농업적 형질과 관련된 변이들 중 핵심 유전자들의 탐색을 용이하게 하고, 나아가 표준 유전체의 존재로부터 파생되는 다양한 분석 기술들이 활용되어 육종 기술 기반의 규모 확장 및 작물 형질 개선의 촉진을 가능하게 한다. 배추의 경우, 표준 유전체 정보와 기보고된 유전자의 기능 관련 정보를 활용하여 농업적 형질에 관련된 핵심 유전자들의 탐색을 돕는 SNP 및 SSR 마커가 개발되어 활용되고 있다.

다수의 cDNA 라이브러리의 병렬적 sequencing을 통한 염기서열에 대한 고속 및 대량 생산 시대를 열은 NGS(Next Generation Sequencing)의 도입과 보편화는 작물 유전체로부터의 자료의 대량 생산을 가능하게 하였다. 작물의 표준 유전체와 re-sequencing 분석를 통해 특정 교배 집단내의 계통들에 대한 변이 정보를 대량으로 손쉽게 얻을 수 있게 되었으며 부모본으로 선정된 계통의 변이 정보를 서로 대조하여 분자마커개발에 활용함으로써 mapping 집단의 세대 진전과 선발에 사용할 수 있는 분자마커를 대량으로 생산할 수 있게 되었다. 이를 통해 특정 형질에 대한 QTL을 식별할 수 있었다. 또한, 근래에 도입된 GWAS(Genome Wide Association Study)는 형질에 연관된 유전적 영역을 in silico 방식으로 쉽게 찾아낼 수 있으며 분석 과정에서 나타난 SNP를 직접적으로 SNP 마커로 전환시킬 수 있기 때문에 향후 분자마커의 대량 생산에서 기본적인 플랫폼으로 사용될 수 있는 체계로 평가받고 있다.

이와 같이 작물의 유전육종을 위한 다양한 체계가 현재 이용 가능한 상황이다. 그러나 다양한 체계들로부터 생산된 육종 관련 정보들이 서로 간의 연관관계를 갖지 못하고 독립적으로 존재하게 된다면, 이는 연구 및 육종 관련 종사자들이 많은 자본과 시간을 투자하여 생산한 정보를 온전히 활용할 수 없게되는 결과를 야기하게 된다. 따라서 생산된 정보간의 관계를 조정하고 이를 일관된 체계상에서 표현하여 이를 필요로 하는 사람들에게 전달할 필요가 있다. 그 결과, 현재 다양한 주제로 작성된 데이터들을 각 데이터의 특성에 맞게 서로 간의 관계망을 구축하고 이를 시각화한 데이터베이스가 각광을 받고 있다.

제 7 장 참고문헌

Barrett, J. C., et al. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2): 263-265.

Barrett, T., et al. (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41(Database issue): D991-995.

Chase, C. D. (2007). Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends Genet 23(2): 81-90.

Cingolani, P., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6(2): 80-92.

Cox, M. P., et al. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11: 485.

Farinho, M., et al. (2004). Mapping of a locus for adult plant resistance to downy mildew in broccoli (*Brassica oleracea* convar. *italica*). Theor Appl Genet 109(7): 1392-1398.

Finn, R. D., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1): D279-285.

Guo, S., et al. (2015). Comparative Transcriptome Analysis of Cultivated and Wild Watermelon during Fruit Development. PLoS One 10(6): e0130267.

Guo, S., et al. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet 45(1): 51-58.

Haas, B. J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8): 1494-1512.

Hatakeyama, K., et al. (2013). Identification and characterization of Crr1a, a gene for resistance to clubroot disease (*Plasmodiophora brassicae* Woronin) in

Brassica rapa L. PLoS One 8(1): e54745.

Huang da, W., et al. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1): 44-57.

Jin, M., et al. (2014). Identification and mapping of a novel dominant resistance gene, TuRB07 to Turnip mosaic virus in *Brassica rapa*. Theor Appl Genet 127(2): 509-519.

Kanehisa, M., et al. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1): D353-D361.

Li, H. and R. Durbin (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754-1760.

Li, R., et al. (2009). SNP detection for massively parallel whole-genome resequencing. Genome Res 19(6): 1124-1132.

Li, X., et al. (2010). Development of a high density integrated reference genetic linkage map for the multinational *Brassica rapa* Genome Sequencing Project. Genome 53(11): 939-947.

Minoche, A. E., et al. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 12(11): R112.

Nayidu, N. K., et al. (2014). *Brassica villosa*, a system for studying non-glandular trichomes and genes in the Brassicas. Plant Mol Biol 85(4-5): 519-539.

Rahman, M., et al. (2007). Development of SRAP, SNP and multiplexed SCAR molecular markers for the major seed coat color gene in *Brassica rapa* L. Theor Appl Genet 115(8): 1101-1107.

Reimand, J., et al. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res 44(W1): W83-89.

Rhee, S. J., et al. (2015). Transcriptome profiling of differentially expressed genes in floral buds and flowers of male sterile and fertile lines in watermelon. BMC Genomics 16: 914.

Schranz, M. E., et al. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci 11(11): 535-542.

Schwelm, A., et al. (2015). The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. Sci Rep 5: 11153.

Trapnell, C., et al. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9): 1105-1111.

Trapnell, C., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3): 562-578.

Wagner, G. P., et al. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131(4): 281-285.

Wang, X., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10): 1035-1039.

Wang, X., et al. (2015). Brassica database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. Database (Oxford) 2015.

Wang, Z., et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1): 57-63.

Zhao, J., et al. (2010). BrFLC2 (FLOWERING LOCUS C) as a candidate gene for a vernalization response QTL in Brassica rapa. J Exp Bot 61(6): 1817-1825.

주 의

1. 이 보고서는 농림축산식품부.해양수산부.농촌진흥청.산림청에서 시행한 Golden Seed 프로젝트의 연구보고서입니다.

2. 이 보고서 내용을 발표할 때에는 반드시 농림축산식품부.해양수산부.농촌진흥청.산림청에서 시행한 Golden Seed 프로젝트의 연구결과임을 밝혀야 합니다.

3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 아니 됩니다.