명지대학교

박사학위 논문

제주도 자생 왕벚나무의 유전체 해독

Draft Genome Sequence of *Prunus* x *nudiflora*, a Natural Hybrid Flowering Cherry

명지대학교 대학원

생명과학정보학과

백 승 훈

지도교수 문 정 환

2019년 8월

제주도 자생 왕벚나무의 유전체 해독

Draft Genome Sequence of *Prunus* x *nudiflora*, a Natural Hybrid Flowering Cherry

이 논문을 박사학위 논문으로 제출함.

2019년 8월

명지대학교 대학원

생명과학정보학과

백 승 훈

제주도 자생 왕벚나무의 유전체 해독

Draft Genome Sequence of *Prunus* x *nudiflora*, a Natural Hybrid Flowering Cherry

명지대학교 대학원
생명과학정보학과
백 승 훈

상기자의 이학박사 학위논문을 인준함.

심사위원장 최 상 봉 (인)
심 사 위 원 권 형 진 (인)
심 사 위 원 문 정 환 (인)
심 사 위 원 김 주 환 (인)
심 사 위 원 이 병 하 (인)

2019년 8월

감사 글

석사 학위를 취득 후 13년 세월이 지난 지금 박사 학위 논문을 작성하였다.
박사 학위 과정 동안에 경험한 지식과 사고의 수준은 석사 시절 보다 많이 달랐다.
연구 주제에 관련된 분석 범위가 넓었으며 결과를 심도 있게 고찰하고 분석 과정에서 발생된 문제점을 해결해 나가기 위해 많은 시간과 노력이 필요했다.
제주도 자생 왕벚나무를 직접 보았을 때 웅장하고 신비롭게 느껴졌다.
자생 왕벚나무 유전체를 연구한 것은 나에게 대단한 행운이다.

소양이 부족한 저에게 열정적으로 지도해 주신 존경하는 문정환 지도교수님께 진심으로 감사를 드립니다. 논문 지도에 참여해 주신 최상봉 교수님, 권형진 교수님, 가천대학교 김주환 교수님, 서강대학교 이병하 교수님께 진심으로 감사를 드립니다.
다양한 조언과 연구에 도움을 주신 김군보 연구교수님, 실험실을 이끌어 가는 주역 조아라, 김문진, 장호열, 손성욱 학우 여러분께 진심으로 감사를 드립니다.

박사 학위를 하겠다고 결정을 했을 때 이해와 힘이 되어준 황혜경 아내에게 고마움과 사랑을 전합니다. 나를 보고 싶어 했고 필요로 할 때 오랫동안 함께 있어 주지 못한 백상윤, 백주엽 아들에게 미안한 마음을 전합니다. 두 아들이 성장하여 나의 박사 학위 논문을 읽을 수 있을 때 내가 무엇을 연구 했는지 이해하고 자랑스러운 아버지로 기억하길 바랍니다. 그리고 백덕현, 손채연 부모님과 백승미 동생에게 고마운 마음을 전합니다.

감사합니다.

백승훈 올림

목 차

그림 목차

표 목차

제주도 자생 왕벚나무의 유전체 해독

백 승 훈

명지대학교 대학원 생명과학정보학과

지도교수 문 정 환

이종간 교배는 식물 다양성을 증가 시키는 중요한 진화적 과정중의 하나이다. 제주도에서 왕벚나무(*Prunus* x *nudiflora*)는 동배수성 잡종으로서 벚나무속(genus *Prunus*)의 유전적 다양성에 기여하며 자생지를 형성하고 있다. 벚나무속은 과수 작물인 체리와 매실을 포함하고 있고, 다수의 종이 아름다운 꽃과 우수한 관상용 특성을 갖고 있기 때문에 전세계적으로 원예 자원으로서 가치가 높다. 제주도 왕벚나무는 자연 발생적으로 형성된 이형접합체 유전체를 갖고 있는 것으로 추정돼 왔다. 따라서 왕벚나무의 유전체 연구는 벚나무속에서 종간 잡종화에 의한 이형접합 유전체의 기능에 대한 형성과 식물유전체학의 진화 생물학적 지식을 제공해 줄 수 있다.

본 논문에서는 장거리 서열에 기반을 둔 유전체 분석 전략을 사용하여 자생 왕벚나무의 높은 이형접합체 유전체를 해독하고 그 정보를 분석하였다. 왕벚나무 유전체는 323.8 Mb 크기의 스케폴드 서열로 조립되었고, 반복서열은 약 47.2%를 포함하고 있다. 유전자 주석 분석 결과 41,294개의 단백질 암호화 유전자를 결정하였다. 이는 유전자 영역의 93% 이상을 포괄한다. 왕벚나무의 모계로 추정되는 올벚나무(*Prunus pendula* f. *ascendens*)와 부계로 추정되는 벚나무(*Prunus jamasakura*)의 유전체 서열을 왕벚나무 유전자 모델과 매핑하여 분석한 결과, 전체 유전자 모델 중 19%는 모계 유래, 21%는 부계 유래, 59%는 모계와 부계의 공통 유전자로 나타났다. 이 결과로 볼 때, 왕벚나무는 모계 올벚나무와 부계 벚나무 사이에서 형성된 1세대 잡종으로 판단된다.

잎, 꽃잎, 암술, 수술, 열매 등 5개 조직에서의 유전자 발현 분석을 실시한 결과 모계 유래 유전자 562개와 부계 유래 유전자 576개가 조직별로 차등 발현하는 것을 확인하였다. 특히 모계에서 유래된 꽃 발달 관련 전사 인자는 영양조직과 생식조직에서 차등 발현 하였으며, 부계에서 유래한 차등 발현 유전자는 꽃가루와 꽃가루관, 세포벽, 2차 대사산물 관련 유전자를 다수 포함하고 있었다. 이는 잡종 강세에 의한 양친 계통

별 특이적 유전자의 발현 양상을 나타낸다.

왕벚나무와 이의 근연종인 올벚나무, 벚나무, 산벚나무(*P. sargentii*), 사옥(*P. jamasakura* var. *quelpaertensis*), 그리고 소메이 요시노(*Prunus* x *yedoensis*) 등 6개 분류군에 속하는 16개 개체의 유전체를 비교 분석한 결과, 다차원척도법의 차원에서 자생 왕벚나무는 모계 올벚나무와 부계 벚나무 양친 그룹의 중간 유전체 특성을 갖는 것을 확인하였다. 또한 제주도 자생 왕벚나무는 소메이 요시노와 유전체 수준에서 뚜렷하게 구분되어 상호 별개의 분류임이 판명되었다.

왕벚나무의 형성에 대한 유전체 기반을 확립하기 위하여 제주도 자연적 생육지 내에서 분포하는 근연종 벚나무 분류군의 자가 불화합성 S-locus 반수체의 구조와 유연관계를 분석하였다. 그 결과 왕벚나무 각 개체는 서로 다른 2개의 S-locus 반수체를 갖고 있으며, 두 개체 중 하나는 벚나무와, 다른 하나는 올벚나무와 S-locus를 공유하고 있었다. 한편 벚나무와 올벚나무간 서로 공유하는 S-locus가 전혀 없었다. 이상의 결과로 볼 때, 벚나무류의 강한 배우자체 자가 불화합성에 의한 동소적 이종간 교배로 인해 자연 잡종화를 통하여 왕벚나무가 생성되었다고 사료된다.

본 연구를 통해 확립한 왕벚나무의 유전체 서열과 유전자 정보는 왕벚나무의 동정과 계통학적 분류, 산업적 활용을 위한 기반으로 활용될 수 있다.

주제어

자생 왕벚나무, 유전체, 이형접합체, 장거리 서열, 어셈블리, 유전체 재분석, 자가 불화합성, 잡종화, 장미과

제 1 장 서 론

제 1 절 연구의 배경 및 필요성

장미과(Rosaceae)는 90개 속에 속하는 약 3,000개 종이 존재하며 많은 종들이 이종간 잡종 또는 속간 잡종이다(Potter et al., 2007). 장미과는 북부 온대 지역에 널리 퍼져 있는 허브, 관목, 수목 등 다양한 식물 종을 포함하고 있다. 일부 종들은 과실과 견과류를 생산하는 식용작물로 경제적, 농업적 중요성 때문에 유전체 해독 연구가 수행되었다. 장미과 중 과수작물의 경우 사과(Velasco et al., 2010), 체리(Shirasawa et al., 2017), 매실(Zhang et al., 2012), 배(Wu et al., 2013; Chagne et al., 2014), 딸기(Shulaev et al., 2011; Hirakawa et al., 2014; Li et al., 2019), 그리고 복숭아(The International Peach Genome Initiative, 2013), 관상용 화훼식물의 경우 장미속 월계화(Raymond et al., 2018), 야생 찔레꽃(Nakamura et al., 2018), 그리고 장미(Hibrand Saint-Oyant et al., 2018) 유전체 연구 결과가 보고되었다. 벚나무속(*Prunus*)은 배나무아과(*Amygdaloideae*)에 속하며 다수의 종이 핵과 과수작물 또는 조경수로 개발되었다. 벚나무속에는 약 250 종이 포함되어 있으며, 대한민국, 일본 그리고 중국 포함하여 동아시아에서 자연적 또는 인공적 교잡으로 개발된 관상용 수목도 다수 있다(Ma et al., 2009; Knight, 1969). 자연 발생적으로 생성된 야생 벚나무와 이종간 교잡에 의해 생성된 벚나무는 오랫동안 동아시아에서 재배되어 왔다. 벚나무는 꽃이 아름다와 전세계적으로 인기 있는 관상용 수목 또는 조경수로 많이 식재되고 있다. 특히, 벚나무 중 소메이 요시노(*P.* x *yedoensis*, Pxy)는 일본내에서 광범위하게 심겨져 있을 뿐만 아니라 미국 워싱턴의 Tidal Basin에도 심겨져 있다(Bailey and Bailey, 1976; Cheng et al., 2000). 소메이 요시노는 1882년~1884년에 Tokyo Ueno 공원에서 뚜렷이 구분되는 3개체가 확인되었고, 그 중 하나가 Somei-yoshino로 보고되었다(Fujino, 1900). 이후 Matsumura에 의해 *Prunus yedoensis* Matsumura라는 학명으로 명명되었다(Matsumura., 1901). 소메이 요시노는 그 기원을 연구하기 위해 형태학적 연구가 실시되었다(Wilson, 1916). 또한 1954년~1963년에는 *P. lannesiana* var. *speciosa* (Oshima-zakura)와 *P. subhirtella* var. *pendula* (Edo-higan)를 교배하여 소메이 요시

노가 부계 오오시마 벚나무(*P. speciosa*)와 모계 올벚나무(*P. pendula* f. *ascendens*) 사이에서 교잡되었고 이즈 반도(Izu peninsula)에서 기원하였다고 추정하였으나 현재까지 자연 생육지는 발견되지 않았다(Takenaka, 1963).

한편, 제주도의 자생 왕벚나무(*P.* x *nudiflora*)는 1908년에 Taquet 신부가 제주도 한라산에서 채집하였다. 1912년에 Koehne은 Taquet 신부의 채집 표본을 바탕으로 *P. yedoensis* Matsumura var. *nudiflora* Koehne 학명을 명명하였다(Koehne, 1912). 1916년에 Nakai는 조선삼림식물편 5집에서 왕벚나무를 *P. yedoensis* Matsumura의 이명으로 처리하였다(Nakai., 1916). 이후 자생 왕벚나무와 일본의 소메이 요시노 사이에서 분류학적 기원과 종의 실체에 대한 논쟁이 제기 되었다. 제주도 자생 왕벚나무와 미국 워싱턴과 일본 도쿄에서 재배되고 있는 소메이 요시노를 분류학적으로 구별하기 위하여 inter-simple sequence repeat (ISSR) 분자마커와 엽록체 DNA의 *rpl*16 분자 마커를 이용한 유전형 분석 연구가 보고되었다(Roh et al., 2007). 또한 자생 왕벚나무의 부계는 벚나무 또는 산벚나무로 추정되며, 모계는 올벚나무로서 왕벚나무와 올벚나무의 엽록체 유체가 동일하다고도 보고되었다(Roh et al., 2007). 한편 20개 Conserved Ortholog Sets (COS) 분자 지표를 이용하여 제주도에서 후보 부모 벚나무종과 함께 왕벚나무 집단의 유전적 구조를 조사한 결과 자생 왕벚나무는 모계 올벚나무(*P. pendula* f. *ascendens*)와 부계 벚나무(*P. jamasakura*)에서 교배되어 유래된 것으로 추정되었다. 특히 자생지 집단에서 야생 왕벚나무의 약 81%는 F1 잡종 가능성이 있고, 나머지 19%는 부계의 유전자형의 추가적인 불균형 유전자 이입에 의한 역교배 잡종으로 확인되었다(Cho et al., 2017). 따라서 제주도 왕벚나무는 자연 생육지에서 생성된 동배수성 종간 잡종으로 사료된다. 자생 왕벚나무와 소메이 요시노의 유전적 특성은 비슷하며, 공통적으로 모계는 올벚나무에서 유래되었으나 자생 왕벚나무의 경우 명확하게 부계는 알져지지 않았다(Cho et al., 2014). 따라서 왕벚나무의 전체 유전체를 해독하여 부계와 모계로 추정되는 종들과 비교하고, 분석하면 왕벚나무의 기원을 추정할 수 있을 것이다.

자가 불화합성(self-incompatibility, SI)은 현화식물이 자가수분을 방지하는 장치로서 교잡을 통해서 번식하게 하는 진화적 시스템이다. 현화식물 중에서 약 60%는 자가불화합성 시스템을 갖고 있는 것으로 추정된다(Hiscock and Kues, 1999). 자가 불화합성에는 배우자체 불화합성(gametophytic self-incompatibility, GSI)와 포자체 불화합성

(sporophytic self-incompatibility, SSI) 시스템으로 나누어진다. GSI와 SSI는 공통적으로 양친에서 유래된 불화합성 특성의 single polymorphic locus (S)에 의해 결정된다. GSI의 경우, 화분과 암술의 불화합성으로 양친으로부터 유래한 반수체(n) S-allele에 의해 조절된다. SSI의 경우, 부모의 꽃밥과 암술의 이배체(2n) S-allele에 의해 조절된다(Hiscock and Tabah, 2003). 장미과(Rosaceae)는 자가 수정을 피하기 위한 양친에게서 유래된 배우자체 자가 불화합성 결정 유전자를 갖고 있다(Fujii et al., 2016). 특히 벚나무는 꽃가루와 암술머리에서 상호작용은 부계 결정요인 반수체 특이적 F-box protein (*SFB*)와 모계 결정요인 S-locus ribonuclease (*S-RNase*) 유전자의 조합에 의해 GSI가 결정된다(Sassa et al., 2010; Vieira et al., 2008). 따라서 자생 왕벚나무의 부계와 모계를 추정하고 소메이 요시노 분류군과의 유연관계를 명확히 구별하기 위해서는 전체 유전체 해독에 기반한 S-locus 구조 분석 연구가 필요하다.

유전체 연구를 위해서는 대상 생물종의 유전체 서열을 해독하고 분석하는 기술이 반드시 필요하다. 지난 2000년 애기장대(Mayer et al., 1999)와 인간(Venter et al., 2001)의 유전체가 해독된 이래 전 세계 연구자들이 다양한 생물종에 대한 유전체 해독 연구를 수행해 왔다. 미국 국립보건원(National Institutes of Health)의 NCBI 데이터베이스에는 유전체 및 유전자 데이터가 등록되어 있고 이로부터 다양한 서열 및 유전자 정보가 활용가능하다. NCBI Genome (www.ncbi.nlm.nih.gov/genome) 데이터베이스에서 2000년부터 2018년까지 각 년도 별로 등록된 유전체의 조립 수준을 컨티그, 스케폴드, 염색체로 구분하여 그 빈도를 비교 분석을 하였다(Figure 1). 총 355 종에 대한 641개 식물 유전체 데이터 중 염색체 수준으로 완성된 유전체는 총 151개(23.6%), 컨티그 수준으로 완성된 유전체는 총 117개(18.3%)이며, 스케폴드 수준으로 완성된 유전체는 총 373개(58.2%)로서 염색체와 컨티그 수준으로 완성된 결과 보다 2배 많았다. 특히 2017년도와 2018년에 스케폴드 조립 서열 등록 빈도가 현저히 높았다. 유전체 조립 결과에서 서열 완성도에 대한 중요한 평가 기준은 연속성 있는 컨티그 서열을 얼마나 길게, 염색체별로 정렬하여 작성 했는가 이다. 스케폴드 수준의 유전체 서열이 다수 등록되는 것은 식물 유전체의 배수성과 반복서열 특성으로 인하여 염색체 수준으로 유전체 조립 완성도를 높이는 것이 어려운 일임을 보여주고 있다.

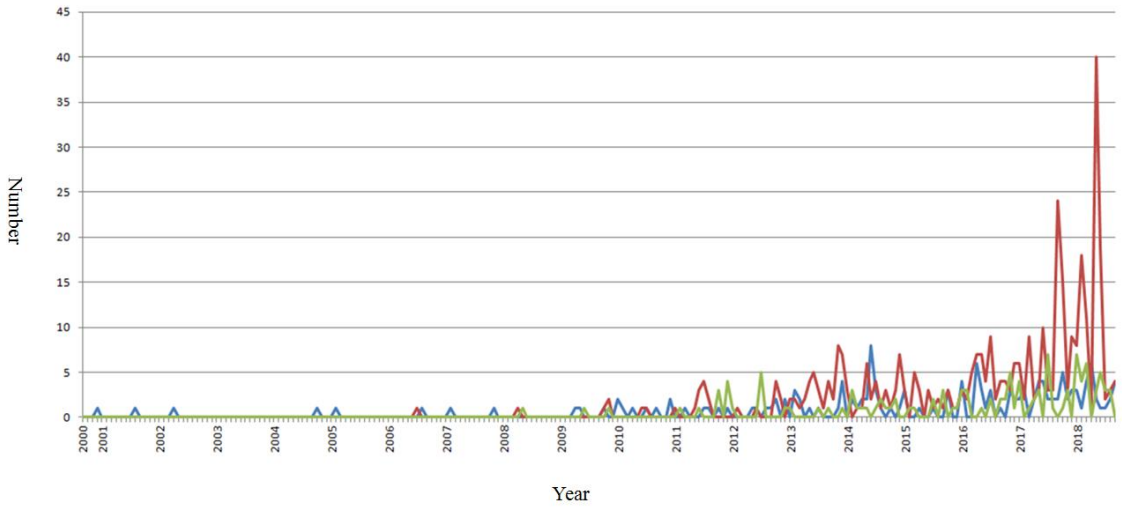


Figure 1. Frequency distribution of assembly levels for the genome sequences reported in NCBI genome database.
Colored lines represent chromosome level (blue), scaffold level (red), and contig level (green), respectively.

현화식물 중 약 80%는 진화 과정 중 배수화 현상을 최소 1회 이상 겪은 것으로 추정된다(Meyers and Levin, 2006). 68개 척추동물과 44개 식물에서 유전체 크기와 반복서열의 비율 상관관계를 분석한 결과 척추동물은 유전체 크기가 증가 할수록 반복 서열 비율은 20% 수준이었으나, 식물 유전체 크기가 증가 할수록 반복 서열 비율도 선형에 가깝게 비례하여 증가하였다(Jiao and Schneeberger, 2017). 식물 유전체의 반복서열은 유전체 해독과 조립을 어렵게 하는 요인으로 지목되어 왔다(Claros et al., 2012). 진핵생물의 유전체 조립에 있어 차세대 유전체 분석 기술(Next Generation Sequencing, NGS)을 이용하여 생산된 대량의 서열 데이터를 가공하고 연결하는 기술적 한계를 극복하기 위해 향상된 유전체 조립 알고리즘이 개발되었다. 대표적인 유전체 조립 알고리즘에는 *de bruijn* graph (DBG) 알고리즘과 overlap layout consensus (OLC) 알고리즘이 있다. DBG 알고리즘과 OLC 알고리즘은 특히 반복서열을 분석하는 방법이 다르다. OLC는 반복서열이 중첩되는 모든 연결 노드(node)를 고려하지만, DBG는 Kmer (substrings of length K) 길이로 연결되는 단일 노드를 분리한다. OLC는 반복적인 서열을 중첩하여 상호 연결되는 모든 노드를 고려하기 때문에 연결 관계에 있는 노드가 많아지고 계산하는 시간이 증가한다. 반면 DBG 경우 단일 노드를 잘라내기 때문에 계산하는 시간은 감소하지만, 연결성이 줄어들어 짧은 길이의 서열을 작성한다. 따라서 유전체 조립 시 시간과 컴퓨팅 자원을 고려할 때 OLC 방법을 이용하는 경우 낮은 커버리지의 long-read 서열을 사용하는 것이 적합하며, DBG 경우 높은 커버리지의 short read 서열을 사용하는 것이 적합하고 알려져 있다(Li et al., 2012).

유전체 조립 알고리즘을 사용하여 NGS 서열로부터 유전체 어셈블리를 제작하는 프로그램으로는 ALLPATH-LG 어셈블러, Celera 어셈블러, FALCON 어셈블러 등이 있다. 각 프로그램의 특징은 다음과 같다. ALLPATH-LG 어셈블러(Gnerre et al., 2011)는 Eulerian *de Bruijn* graph 알고리즘을 기반으로 전체 유전체 조립을 수행 NGS 서열의 오류 수정을 수행한 후 모든 short read 서열에서 Kmer 값에 해당하는 하나의 read 서열이 다른 하나의 read 서열을 커버하는 모든 연결 경로를 찾는다. Kmer 서열이 중첩하는 과정 중 두 개 이상으로 연결될 경우 모든 branch에서 복잡도가 없는 unipath 경로를 선정한다. Unipath 경로에서 서열 매핑률이 높은 것을 추정하여 seed로 선정하며, 10 kb 이하 범위에서 short read 서열과 일관성 있게 연결되는

unipath 경로를 선정한다. 상호 연결되는 unipath 경로에 있는 서열을 이용하여 유전체 조립을 수행하여 컨티그 서열과 스케폴드 서열을 작성한다.

Celera 어셈블러(Miller et al., 2008)는 OLC 알고리즘을 이용한 서열 중첩 과정을 수행하여 서로 다른 NGS 서열을 혼합한 hybrid 유전체 조립 결과를 제공한다. NGS read의 커버리지를 증가시키기 위해 overlap-based trimming 과정을 수행하며, read 중첩 과정에서 정확하게 일치하는 seed를 사용하는 anchors and overlaps 과정을 수행한다. 서열이 서로 중첩되는 노드와 read의 양쪽 끝이 일치하는 서열과 중첩 데이터인 multigraph 결과에서 best overlap graph 과정과 중첩과정에서 서열들을 하나로 연결시키는 graph를 선택하는 unitig construction 과정을 수행한다. 또한 유전체의 반복서열처럼 복잡하게 교차되는 경로를 피하기 위한 heuristics 알고리즘을 이용하여 교차점을 분리하는 unitig splitting 과정을 수행한다. 컨티그 조립, 스케폴드 조립, 조립 서열의 consensus 서열 작성은 hybrid 조립 과정에서 특별한 수정 없이 Celera 파이프라인을 통해 진행 시킬 수 있다.

FALCON 어셈블러는 PacBio의 long read 서열을 조립하는 프로그램이다. Daligner 프로그램을 이용하여 PacBio 서열에 존재하는 오류 서열을 교정하고 정렬된 서열에서 이형접합체 SNP 정보를 포함한 consensus 서열을 생성한다. 어셈블러는 조립 과정에서 이배체 유전체의 haplotype phasing 정보를 직접 통합하여 사용한다. 조립 과정에서 상동성 서열 간에서 높은 이형성에 의한 주요 구조적인 변이 영역이 포함된 haplotype fused 컨티그를 Myers' 방법으로 분리하고 선형적으로 연결되는 string graph 정보를 이용하여 유전체를 조립한다(Chin et al., 2016). 이형접합도가 높은 유전체의 short read 서열을 사용하여 전체 유전체를 조립한 경우 서열 연속성이 있는 컨티그 서열 N50 길이는 50 kb를 넘지 못한다(Kajitani et al., 2014). 반면, long read 서열을 사용하는 FALCON 어셈블러는 애기장대 F1 hybrid (Col-0 x CVI-0) 유전체 조립 결과에서 N50 길이 7.9 Mb 수준의 조립서열을 생성했다(Chin et al., 2016). 또한 이형접합도가 높은 카베르네 소비뇽 포도(*Vitis vinifera* cv, Cabernet Sauvignon) 유전체 조립 결과에서도 FALCON 어셈블러는 N50 길이 2.3 Mb 수준의 완성도 높은 haplotigs 서열을 조립하였다(Chin et al., 2016). 따라서 FALCON 알고리즘을 이용할 경우 왕벚나무 같이 이종간 교배에 의해 형성된 이형접합도가 높은 유전체를 조립하여 완성도가 높은 전체 유전체를 획득할 수 있을 것으로 사료된다.

제 2 절 연구의 목표

자연 발생적 잡종화에 의해 생성된 것으로 추정되는 제주도 자생 왕벚나무는 소메이 요시노와 분류군의 기원, 종의 경계, 근연관계 등에 대한 논쟁이 있어 왔다. 두 분류군간의 유연관계에 대한 유전학적 연구로서 모계 유래된 엽록체 서열을 비교 분석하는 방법이 적용될 수 있으나 부계 기원에 대한 분석에는 한계가 있다. 따라서 왕벚나무의 유전적 실체와 진화적 기원을 규명하기 위하여 유전체 해독을 통한 전체 유전체 분석과 모계와 부계기원 haplotype 유전체의 구분 및 특성 분석이 필요하다.

본 연구에서는 제주도 자생 왕벚나무 기념목 천연기념물 159호 2번 개체의 전체 유전체를 해독하고 조립 완성도가 높은 유전체 초안을 완성하고자 하였다. 자생 왕벚나무 유전체를 기반으로 근연종인 모계 추정 올벚나무, 부계 추정 벚나무, 그리고 소메이 요시노의 전체 유전체를 추가 해독하여 벚나무속 분류군간의 유전자형 차이, 자가 불화합성 유전 좌위의 구조 분석을 실시하여 왕벚나무의 양친 및 소메이 요시노의 유연관계를 확립하고자 하였다. 또한, 유전체 정보를 활용하여 자생 왕벚나무 종판별에 필요한 분자 마커 후보를 선발하고, 장미과 비교 유전체 분석을 통해 벚나무속의 진화 특성 등 밝혀 제주도 자생 왕벚나무의 유전체 특성과 기원을 구명하는 것을 목표로 하였다.

제 2 장 연구 재료 및 방법

제 1 절 연구 재료

1. 식물 재료

가. 제주도 자생 왕벚나무 기념목과 근연종 왕벚나무 및 벚나무류

본 연구에서 전체 유전체 조립에 사용한 식물재료는 제주도 자생지 내에서 국가가 지정한 기념목 천연기념물 제 159호 2번 자생 왕벚나무(*P.* x *nudiflora*, Pxn) Pxn-Jeju2를 사용하였다(Figure 2; Table 1). 자생 왕벚나무 Pxn-Jeju2와 자생지에서 근연종 왕벚나무 유전체 비교를 위해 제 159호 1번(Pxn-Jeju1), 3번(Pxn-Jeju3)과 제주 시도 기념물 제51호 2번(Pxn-Jeju4), 제주시 향토유산 기념물 제3호(Pxn-Jeju5)를 식물 재료로 사용하였다.

왕벚나무 분류군 비교를 위해 미국 워싱턴에 있는 소메이 요시노 Pxy-US1, Pxy-US2와 일본 도쿄에 있는 소메이 요시노 Pxy-JP1, Pxy-JP2를 식물 재료로 사용하였다.

부계와 모계에서 유연관계를 비교하기 위해 모계로 추정되는 올벚나무(*P. pendula* f. *ascendens*, Ppa) Ppa-1, Ppa-2, Ppa-3 그리고 부계로 추정되는 벚나무(*P. jamasakura* var. *jamasakura*, Pjj) Pjj-1, Pjj-2, 사옥(*P. jamasakura* var. *quelpaertensis*, Pjq), 산벚나무(*P. sargentii*, Psa)를 식물재료로 사용하였다.

나. 왕벚나무 전사체 분석을 위한 조직

자생 왕벚나무 표준 유전체에서 유전자 모델 확립, 각 조직 간에서 유전자 차등 발현 분석, 부계와 모계에서 유래된 유전자의 발현 특성을 연구하기 위해 잎눈(foliarbud), 꽃눈(floral bud), 수술(stamen), 암술(pistil), 꽃잎(petal), 열매(berry), 잎(leaf) 등 총 7개 조직을 채집하였다(Figure 3).

Figure 2. Photographs of a Pxn-Jeju2 tree, the reference accession of wild *P.* x *nudiflora* used in this study.

Table 1. Summary information of accessions in closely related *Prunus* taxa

Taxon	Name	Locality	Voucher
P. x *nudiflora*	Pxn-Jeju1	Korean National Monument 159, No. 1, Bongae-dong, Jeju, Korea	159-1
(wild *P.* x *nudiflora*)	Pxn-Jeju2	Korean National Monument 159, No. 2, Bongae-dong, Jeju, Korea	159-2
	Pxn-Jeju3	Korean National Monument 159, No. 3, Bongae-dong, Jeju, Korea	159-3
	Pxn-Jeju4	Jeju Province Monument No. 51, Gwaneumsa Temple, Jeju, Korea	51-2
	Pxn-Jeju5	Jeju Province Local Tangible Heritage No. 3, Odeung-dong, Jeju, Korea	128
P. x *yedoensis*	Pxy-US1	Tidal Basin at the National Mall, Washington D.C., USA	NA69513
(Yoshino cherry)	Pxy-US2	Tidal Basin at the National Mall, Washington D.C., USA	NA69515
	Pxy-JP1	Koishikawa Botanical garden, Tokyo, Japan	JKS2206
	Pxy-JP2	Ueno park, Tokyo, Japan	JKS2215
P. pendula f. *ascendens*	Ppa-1	Harye-ri, Seogwipo-si, Jeju, Korea	60571
	Ppa-2	Gwaneumsa Temple, Jeju, Korea	Gachon-P2
	Ppa-3	Bongae-dong, Jeju, Korea	Gachon-P6
P. jamasakura var. *jamasakura*	Pjj-1	Odeung-dong, Jeju, Korea	63375
	Pjj-2	Odeung-dong, Jeju, Korea	Gachon-P3
P. jamasakura var. *quelpaertensis*	Pjq	Ara 1-dong, Jeju, Korea	63437
P. sargentii	Psa	Odeung-dong, Jeju, Korea	63385

A

B

C

Figure 3. Photographs of flowers (A and B), berries and leaves (C) of Pxn-Jeju2.

제 2 절 연구 방법

1. Genomic DNA와 전사체 RNA의 추출 및 시퀀싱

가. Short read 서열 생산을 위한 DNA 추출 및 시퀀싱

Short read 서열 생산을 위해 자생 왕벚나무 기념목 제159호 2번에서 어린 잎을 채취하였다(Figure 3C). 유전체 DNA (gDNA)는 액체 질소로 곱게 간 잎 조직을 CTAB buffer로 추출한 후 Isopropanol를 첨가하여 -20℃에서 8시간 동안 침전시켜 gDNA를 회수하였다(Khanuja et al. 1999; Murray and Thompson 1980). 수확한 DNA는 Illumina TruSeq DNA Sample Prep Kits를 사용하여 유전체 DNA 삽입체 길이 250 bp, 500 bp의 라이브러리로 제작하고, Illumina의 NextSeq, MiSeq 그리고 HiSeq X 시퀀싱 플랫폼을 사용하여 시퀀싱을 했다. 삽입체의 5' 과 3' 영역에서 NextSeq과 HiSeq X는 150 bp, MiSeq은 300 bp 길이로 paired end (PE) 서열을 생산했다(Bentley et al., 2008).

조립 서열의 커버리지를 높이기 위해 Illumina Nextear Mate Pair Library Prepare Kit를 사용하여 삽입체 길이 3 kb, 5 kb, 10 kb, 15 kb, 20 kb 라이브러리를 제작하였다. 삽입체 DNA 5'과 3' 끝에 biotin을 붙이고 원형의 삽입체를 분절하여 잘라냈다. 작은 fragment DNA 중 biotin이 존재하는 서열을 선별하여 잘려진 양쪽 부위에 어댑터를 부착했다. NextSeq과 HiSeq X 플랫폼을 사용하여 삽입체 양끝을 150 bp 길이로 mate paired end (MP) 시퀀싱을 수행하였다. 또한 삽입체 길이 40 kb Fosmid 라이브러리를 제작하였다. pFosill vector에 ILMN-F와 ILMN-R 프라이머 서열이 포함한 Nb.BbvCI nick 위치에 40 kb 삽입체를 넣어 55,200개 클론을 생산하였다. S1 nuclease 사용하여 nick 위치를 잘라내고, HiSeq X 플랫폼을 사용하여 삽입체 양 끝을 150 bp 길이로 시퀀싱하였다.

나. Long read 서열 생산을 위한 유전체 DNA 추출 및 시퀀싱

Long read 서열 생산을 위해 자생 왕벚나무 기념목 제159호 2번에서 어린 잎을 채취하였다(Figure 3C). gDNA 추출 방법은 고분자량(high molecular weight) 추출법을 적용하였다(Zhang et al., 2012). Polysaccharide 등 long read 서열 해독에 저해 물질을 완전히 제거하기 위해 잎에서 원형질체를 유도한 후 세포핵을 분리하고 정제된 고분자량

gDNA를 추출하였다. Long read 서열 생산은 PacBio RSⅡ 플랫폼을 사용하였다. 삽입체 길이는 20 kb 크기로 라이브러리를 제작하였다. 삽입체 양 끝에 hairpin 어댑터를 붙이고 single-molecule real-time (SMRT) cell 안에 DNA polymerase를 이용하여 삽입체를 복제했다. 복제 과정에서 zero-mode waveguide (ZMW)를 이용하여 DNA polymerase가 합성에 사용하는 dNTP 형광 파동을 구별하는 방법으로 PacBio subread 시퀀싱을 수행했다.

다. 전사체 RNA 추출 및 시퀀싱

전사체 시료는 자생 왕벚나무 기념목 Pxn-Jeju2에서 잎눈(foliar bud), 꽃눈(floral bud), 수술(stamen), 암술(pistil), 꽃잎(petal), 열매(berry), 잎(leaf) 7개 조직을 채취하였다(Figure 3). 잎눈과 꽃눈 조직 샘플을 제외한 각 잎, 열매, 수술, 암술, 꽃잎 조직 샘플은 3회 반복으로 준비하였다. RNA 추출 방법은 각 조직 샘플에서 CTAB/High salt 침전 방법(Tan and Yiap, 2009)으로 total RNA를 추출한 후 Illumina TruSeq Stranded mRNA Sample Preparation Kit를 사용하여 mRNA를 분리하였다. cDNA를 합성하여 삽입체 길이 250 bp와 500 bp로 인덱스 어댑터를 연결하여 라이브러리를 제작했다. Illumina NextSeq과 MiSeq 플랫폼을 사용하여 각각 150 bp, 300 bp 길이로 시퀀싱을 수행하였다.

2. 전사체 서열 데이터의 품질 관리

각 잎눈(foliarbud), 꽃눈(floral bud), 수술(stamen), 암술(pistil), 꽃잎(petal), 열매(berry), 잎(leaf) 7개 조직별로 mRNA-Seq 데이터에서 시퀀싱 어댑터와 낮은 신뢰도를 갖는 서열을 잘라내고, 비핵 유전체 및 rRNA 등의 오염원 서열을 필터링 하기 위해 Trimmomatic v0.32 (Bolger et al., 2014) 프로그램을 사용하여 전사체 데이터 품질 관리를 수행하였다. 입력된 데이터는 각 조직별 서열 데이터를 사용하였다. 신뢰도가 낮은 서열은 4 bp 단위로 읽어 구간별 Phred Score 값이 평균 30 이하인 경우 제거했고, 품질 관리 과정에서 서열 길이가 36 bp 이하인 short read 서열 또한 제거하였다. 그리고 오염원 서열을 제거하기 위해 bowtie2 v2.2.3 (Langmead and Salzberg, 2012) 프로그램을 이용하여 비핵유전체(GenBank accession: NC_026980)와 rRNA 서열에 매핑 된 short read 서열을 제거하였다.

3. 유전체 서열 데이터의 품질 관리

가. 유전체 short read 데이터 품질관리

잎 조직에서 250 bp와 500 bp 라이브러리 시퀀싱 데이터에서 시퀀싱 어댑터와 낮은 신뢰도를 갖는 서열 제거 및 비핵 유전체 서열을 필터링 하기 위해 Trimmomatic v0.32 프로그램을 사용하여 유전체 데이터 품질관리를 수행하였다. 신뢰도가 낮은 서열은 4 bp 단위로 읽어 구간별 Phred Score 값이 평균 20 이하인 경우 제거했고, 품질 관리 과정에서 서열 길이가 36 bp 이하인 short read 서열 또한 제거하였다. 시퀀싱 과정에서 중합효소연쇄반응(polymerase chain reaction) 의해 생성 될 수 있는 중복서열(duplicate read)를 제거하기 위해 FastUniq v1.1 (Xu et al., 2012) 프로그램을 사용하였다. 그리고 비핵 유전체 서열을 필터링하기 위해 bowtie2 프로그램을 이용하여 왕벚나무 엽록체 서열(GenBank accession: NC_026980)에 매핑 된 short read 서열을 제거하였다.

나. 유전체 Mate Paired End 데이터의 품질관리

잎 조직에서 시퀀싱 된 3 kb, 5 kb, 10 kb, 15 kb, 20 kb MP 라이브러리와 40 kb Fosmid end 라이브러리 시퀀싱 데이터에서 시퀀싱 Nextera 어댑터와 중복 서열을 제거하기 위해 NextClip v1.3 (Leggett et al., 2014) 프로그램을 사용하여 품질관리를 수행하였다. 파라미터는 중복서열 제거를 위한 remove duplicates 옵션을 적용하였다. 40 kb Fosmid end 데이터는 cloning vector 어댑터를 제거하고, BWA v0.7.12 (Li and Durbin, 2009) 프로그램을 이용하여 조립 유전체 서열에 매핑되는 PE 서열을 선별하였다.

다. Long read PacBio RsⅡ 데이터의 품질관리

단일분자 실시간 시퀀싱 기술을 이용한 PacBio RsⅡ 시퀀싱 데이터는 SMRT Portal 분석 프로그램을 통해 중합효소 서열(polymerase read)에서 어댑터를 제거하고, 획득한 서브리드(subread) 서열 중 길이가 500 bp 이하인 서열은 제거하였다. 그리고 PacBio subread 서열을 정렬하는 MinHash Alignment Process (MHAP) 알고리즘을 탑재한 PBcR (Berlin et al., 2015) 프로그램을 사용하여 시퀀싱 오류 서열을 교정하였다.

4. 자생 왕벚나무 유전체 크기 추정

유전체 크기 추정은 품질관리를 수행한 PE short read 서열을 사용하였다. 크기 추정에 필요한 K-mer 빈도 값을 계산하기 위해 Jellyfish v2.1.3 (Marcais and Kingsford, 2011) 프로그램을 사용하였다. 입력 파라미터 값은 K 값 17을 적용하였다. K-mer 분포에서 이형접합(heterozygous)과 동형접합(homozygous) 중에서 동형접합에 해당하는 배수 값을 선택하여 유전체 크기를 추정하였다. 유전체 길이 계산은 BGI에서 제시한 방법에 따라 K-mer 빈도의 배수 값 c_{k-mer}, short read 서열의 평균 길이 L, K-mer 값 K 를 사용하여 유전체 크기 대비 커버리지 값 c_{base}를 계산하였고, 유전체 크기 G 를 계산하기 위해 단거리 서열 길이의 총 합 n_{base} 값을 계산하였다(Li et al., 2010).

$$c_{base} = \frac{c_{k-mer} \times L}{L-K+1}$$

$$G = \frac{n_{base}}{c_{base}}$$

유전체 크기를 추정하는 수식

5. 자생 왕벚나무 유전체 분석 전략

자생 왕벚나무 유전체 서열을 해독하여 유전체 조립, 유전자 모델 예측, 근연종 벚나무류 집단에서 유전체 변이 분석을 위해 Figure 4와 같은 분석 전략을 수립하고 연구를 수행하였다.

이형접합성이 높은 유전체의 경우 short read 서열을 사용하여 유전체를 조립할 경우 조립 서열 개수가 많고 서열 길이가 짧은 서열로 조립되는 경우가 있다. 이형접합성(heterozygosity)이 높은 자생 왕벚나무 유전체를 조립하기 위해 PacBio subread 서열만을 사용하여 FALCON 어셈블러를 통해 *de novo* 어셈블리를 수행하였다. 조립 서열의 커버리지를 높이기 위해 MP 서열과 Fosmid end 서열을 사용하여 SOAPdenovo2 v2.04 (Luo et al., 2012) 및 OPERA-LG v2.0.6 (Gao et al., 2016) 프로그램을 통해 스케폴드 조립을 수행하였다. 스케폴드 조립 서열에는 공백 영역이 존재하며, PBJelly v15.2.20 (English et al., 2012) 프로그램을 사용하여 서열 오류가 교정된 PacBio subread 서열로 공백영역을 교정하였다. PacBio subread 서열을 사용한 조립 서열에 존재하는 조립 서열 오류를 교정하기 위해 short read 서열을 매핑 후 Pilon v1.16 (Walker et al., 2014) 프로그램을 사용하여 조립 서열을 교정하였다.

자생 왕벚나무의 유전자 모델 예측 방법은 *ab-initio* 유전자 예측, 전사체 서열을 유전체에 매핑, 그리고 단백질 서열을 유전체에 매핑하여 유전자 예측 방법을 적용하였다. 예측된 유전자 모델은 EVidenceModeler (EVM) (Haas et al., 2008) 프로그램을 통해 일치성 있는 유전자 모델로 통합하여 확립하였다.

소메이 요시노, 올벚나무, 산벚나무, 사옥, 벚나무의 유전체 서열은 상기한 Illumina PE 서열로 생산한 후 Pxn-Jeju2 참조 유전체에 매핑하여 근연종 벚나무 간에서 유전체 구조적 차이와 변이를 분석하였다.

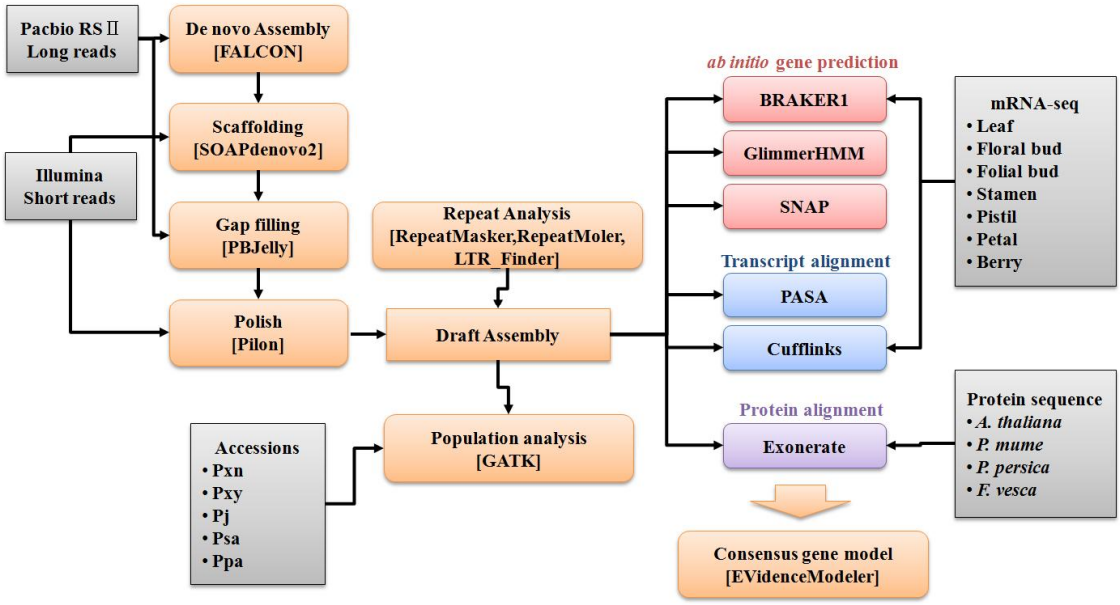


Figure 4. The integrated analysis pipeline for the Pxn genome and the variome of *Prunus* species.

The analysis workflow was shown by arrows. The related software, data and analysis results in each step were indicated in boxes.

6. 자생 왕벚나무 Pxn-Jeju2 전체 유전체 조립

가. PacBio RsⅡ 데이터를 이용한 유전체 조립

유전체 조립은 PacBio RsⅡ subread 서열을 사용하여 FALCON 어셈블러로 조립하였다. FALCON 어셈블러에 사용한 입력 파라미터는 서열 길이가 12 kb 이상인 서열만을 사용하는 length_cutoff = 12000, 서열 중첩 분석 시 seed-read 서열 길이 length_cutoff_pr = 12000, 사전 조립 과정 중 낮은 커버리지 영역에서 seed-read를 분리 또는 잘라내기 위한 설정 falcon_sense_option= --output_multi --min_idt 0.70 --min_cov 4 --local_match_count_threshold 2 --max_n_read 200 --n_core 6, 중첩 분석 시 반복 서열을 고려한 필터링 설정 overlap_filtering_setting = --max_diff 100 --max_cov 150 --min_cov 2 --n_core 24 을 적용하였다. 비핵유전체는 엽록체 서열(GenBank accession: NC_026980)과 가천대학교 식물분류학 연구실에서 제공 받은 미토콘드리아 서열에 NUCmer (Kurtz et al., 2004) 프로그램을 이용하여 조립 서열을 정렬하여 조립 서열이 비핵 유전체 서열에 매핑 커버리지가 50% 이상인 경우 제거했다.

나. Mate Paired End 데이터를 이용한 조립 서열의 스케폴드 제작

FALCON 어셈블러로 조립되지 않은 유전체 서열을 재조립하기 위해 250 bp와 500 bp 라이브러리 PE short read와 3 kb, 5 kb, 10 kb, 15 kb 라이브러리 MP short read를 사용하였다. 스케폴드 유전체 조립은 SOAPdenovo2 프로그램에서 finalFusion, map, scaff 분석 모듈을 단계별로 사용하였고, 파라미터는 K=23 값을 적용하였다. 또한 20 kb MP, 40 kb Fosmid end 데이터는 OPERA-LG 프로그램을 이용하여 전체 유전체 재조립에 사용하였다.

다. PacBio subread 데이터를 이용한 재조립 서열의 공백 영역을 교정

스케폴드 조립 서열에서 공백 서열은 MP short read 서열이 매핑 된 라이브러리 길이가 반영되어 공백 영역을 생성한다. 공백 서열을 제거하기 위해 PBJelly 프로그램을 사용하였다. 서열 오류가 교정된 PacBio subread 서열을 스케폴드 유전체 서열에 매핑하기 위해 BLASR v1.3.1.142244 (Chaisson and Tesler, 2012) 프로그램을 사용하였다. PBJelly 파라미터는 -minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20 -maxScore -500 -nproc 30 -noSplitSubreads을 적용하였다. 분석 과정에서 공백

영역을 지정하는 support 파라미터 값 --capturedOnly --minMapq=250을 적용하여 스케폴드 조립 서열 내에 존재하는 공백 영역 서열을 교정하였다.

라. 전체 유전체 조립 서열 교정

공백 교정이 완료 된 전체 유전체 조립 서열에는 조립 과정에서 PacBio subread의 낮은 커버리지 영역, 서열 중첩 과정에서 발생하는 조립 오류 서열이 존재한다. 이러한 오류와 서열 변이를 교정하기 위해 시퀀싱 오류가 낮고 시퀀싱 커버리지가 높은 short read PE 데이터와 서열 오류를 교정한 PacBio subread 서열을 매핑하여 서열 중첩을 통해 조립 오류 서열을 교정하였다. Short read 서열 매핑은 bowtie2 프로그램을 이용하였고, 파라미터와 옵션은 기본값을 적용하였다. Long read 서열은 BLASR 프로그램을 이용하여 매핑하였고, 파라미터 값은 minPctSimilarity 90 값을 적용하였다. Short read 와 long read 매핑 데이터인 bam 파일을 Pilon 프로그램을 이용하여 전체 유전체 조립 서열에 매핑 된 영역에서 공백과 조립 오류 서열을 교정하였다. 옵션은 diploid, fix=bases,gaps을 적용하였다.

7. 반복서열 및 non coding RNA 예측

가. RepeatMasker와 RepeatModeler를 이용한 반복서열 예측

자생 왕벚나무 유전체에 존재하는 반복 서열을 예측하고 유전자 예측의 정확도를 높이기 위해 RepeatMasker v4.0.5 (Smit and Green, 1996)와 RepeatModeler v1.0.8 (Smit, 2008) 프로그램을 이용하여 반복서열을 분석하였다. RepeatMasker 분석 방법은 RepBase (Bao et al., 2015) 반복서열 라이브러리 데이터를 적용하여 RMBlast 프로그램을 통해 상동성 높은 반복서열을 예측하였고, 파라미터는 -species viridiplantae를 적용하였다. RepeatModeler 분석 방법은 RECON v1.08 (Bao and Eddy, 2002)과 RepeatScout v1.0.5 (Price et al., 2005) 프로그램과 연동하여 반복서열을 *de novo* 예측하고, 기존에 알려진 반복서열과 서열 일치도를 분석하여 새롭게 예측된 반복서열을 판별하였다. 분석 과정은 왕벚나무 유전체 서열을 RepeatModeler 내 BuildDatabase 프로그램을 사용, -name Prunus -engine ncbi 파라미터 값을 적용하여 RMBlast 서열 데이터베이스로 생성하였다. 서열 데이터베이스를 입력 데이터로 사용하고 RepeatModeler 분석을 통해 반복 서열로 판별된 반복서열을 생성하였다. 반복서열로

판별된 서열 중에서 Unknown 으로 판별된 서열은 필터링 하였다. 생성된 서열 라이브러리를 -lib 옵션을 지정하여 RepeatMasker 분석을 통해 유전체에 존재하는 *de novo* 반복 서열을 예측하였다.

나. LTR 반복서열 및 ncRNA 예측

LTR 예측은 LTR_FINDER v1.05 (Xu and Wang, 2007) 프로그램을 사용하였다. 파라미터는 기본값과 테이블 형식으로 결과를 생성하는 -w 2 옵션을 적용하였다. 분석 결과에서 Primer Binding Site (PBS), Polypurine Tract (PPT)를 포함한 5′-LTR 에서 3′-LTR 까지 영역에서 LTR 위치 정보를 추출하였다.

Non coding RNA (ncRNA)는 Infernal v1.1.2 (Nawrocki and Eddy, 2013) 프로그램과 Rfam (Kalvari et al., 2018) 데이터베이스를 연동하여 cmscan 프로그램을 통해 예측하였다. 옵션은 -rfam과 -nohmmonly을 적용하였다. 분석 결과에서 E-value 값 0.01 이하의 ncRNA를 선별하였다.

다. MicroRNA 예측

MicroRNA 영역을 예측하기 위해 miRBase (Kozomara and Griffiths-Jones, 2014) 데이터베이스와 MapMi v1.5.9 (Guerra-Assuncao and Enright, 2010) 프로그램을 사용하였다. MicroRNA 서열은 miRBase 데이터베이스에서 Viridiplantae에 해당하는 73개 종에서 8,496개 서열을 다운로드 하였다. MapMi 프로그램을 이용하여 miRNA 서열을 유전체 서열에 매핑하여 miRNA의 precursor 영역을 포함한 miRNA를 예측하였다.

8. 자생 왕벚나무 전체 유전체에서 유전자 모델 예측

유전자 예측 방법은 유전자 모델을 이용한 *ab-initio* 분석 방법, 근연종 단백질 서열을 유전체에 매핑하여 유전자에서 엑손 영역을 예측하는 분석 방법, 그리고 mRNA-Seq short read 서열 매핑 및 전사체 조립 서열을 자생 왕벚나무에 매핑하여 엑손 영역을 예측하는 분석 방법을 사용하였다. 각 분석 방법을 통해 예측된 유전자 모델 결과를 종합하여 유전자 엑손 영역에 일치성 있는 유전자 모델을 최종 예측하였다. 유전자 모델 예측 과정은 다음과 같다.

먼저, *ab-initio* 유전자 예측은 SNAP (Korf, 2004), GlimmerHMM v3.0.2 (Majoros

et al., 2004), BRAKER1 v1.8 (Hoff et al., 2016) 프로그램을 사용하였다. SNAP을 이용한 분석은 A.thaliana.hmm 애기장대 HMM 파라미터를 적용하여 유전자 ORF가 존재하는 유전자만을 선별하였다. GlimmerHMM 분석은 애기장대 모델 파라미터를 적용하고 -f 옵션을 이용하여 유전자 ORF가 존재하는 유전자만을 선별하였다. BRAKER1 분석에는 mRNA-Seq 데이터를 TopHat2 v2.1.0 (Kim et al., 2013) 프로그램을 이용하여 전체 유전체 조립 서열에 매핑하였고 파라미터는 -microexon-search을 적용하였다. BRAKER1 분석 입력 데이터는 TopHat2 분석 결과 BAM 파일과 반복서열을 소문자로 치환한 전체 유전체 조립 서열과 –softmasking 옵션을 적용하여 AUGUSTUS v3.2.1 (Stanke et al., 2004) 기반 분석 파이프라인에 따라 유전자 모델을 예측하였다.

전사체 데이터를 이용한 유전자 모델 예측 분석 방법에는 Cufflinks v2.2.1 (Trapnell et al., 2010), Trinity v2.4.0 (Grabherr et al., 2011), PASA (Haas et al., 2003) 프로그램을 이용하였다. Cufflinks 분석은 BRAKER1 분석에서 TopHat2의 매핑 결과인 accepted_hits.bam 파일을 이용하여 기본 파라미터 값을 적용하였다. Trinity 분석은 모든 조직의 mRNA-Seq 데이터를 사용하였고 -max_memory 100G, -jaccard_clip 파라미터 값을 적용하였다. PASA 분석에 사용된 입력 데이터는 Trinity를 이용하여 조립한 조립 전사체 서열과, Cufflinks 분석에서 매핑 정보로 분석된 transcripts.gtf 데이터를 사용하였다. 그리고 전체 유전체 서열에 조립 전사체 서열을 매핑하기 위해 GMAP (Wu and Watanabe, 2005), BLAT (Kent, 2002) 프로그램을 사용하여 PASA 분석 파이프라인에 따라 유전자 모델을 예측하였다.

근연종 단백질 서열을 전체 유전체에 매핑하여 유전자를 예측 방법에는 Exonerate v2.2.0 (Slater and Birney, 2005) 프로그램을 사용하였다. 애기장대(*Arabidopsis thaliana*; Lamesch et al., 2012) 35,386개, 복숭아(*Prunus persica*; The International Peach Genome Initiative, 2013) 28,702개, 매실(*Prunus mume*; Zhang et al., 2012) 31,390개, 딸기(*Fragaria vesca*; Shulaev et al., 2011) 34,809개 단백질 서열을 사용하여 전체 유전체 서열에 매핑 후 단백질 서열 일치도 70% 이상 조건에서 유전자 모델을 예측하였다.

각 유전자 예측 분석법을 통해 분석된 유전자의 위치 정보를 가공한 후 EVidenceModeler (EVM) 프로그램을 활용하여 통합하고 일치하는 영역의 유전자 모

델을 선별하였다. EVM 분석에 있어 각 유전자 매핑 분석 결과에 대한 가중치는 *ab-initio* 1, 전사체 매핑 10, 단백질 매핑 5의 파라미터 값을 적용하였다.

EVM 분석을 통해 예측된 유전자 모델에서 단백질 서열을 사용하여 주석 정보를 검색하였다. 주석 정보 검색 방법은 BLASTP (Altschul et al., 1990)프로그램을 사용하였고 E-value $1E^{-10}$ 기준으로 검색하였다. 검색 서열 데이터베이스는 NCBI에서 제공하는 NR와 RefSeq plant, InterPro, UniProt, 복숭아, 매실, 딸기, 사과(*M.* x *domestica*; Velasco et al., 2010) 그리고 애기장대에서 제공하는 단백질 서열을 이용하여 상동성 높은 유전자 주석 정보를 선별하였다. 통합 유전자 모델 결과에서 완전한 ORF가 존재하는 유전자만을 선별하였고, 70% 커버리지 이상 반복서열을 코딩하는 유전자와 단백질 서열 길이가 100 bp 이하인 유전자 모델은 제외하였다.

9. 장미과 비교 유전체 분석

가. 전체 유전자 모델에서 오솔로그 유전자 분석

장미과 유전체에서 공통 또는 유일한 오솔로그 유전자 그룹을 분석하기 위해 OrthoMCL v2.0 (Fischer et al., 2011) 프로그램을 사용하였다. 입력 데이터는 Pxn-Jeju2 유전자 41,294개와 복숭아 28,702개, 매실 31,390개, 체리 43,673개, 딸기 34,809개, 사과 63,540개 유전자의 단백질 서열을 사용하였다. 분석에 사용된 파라미터는 percentMatchCutoff=50, evalueExponentCutoff=E^{-5} 를 적용하였다. OrthoMCL 결과에서 오솔로그 그룹명과 그룹별 유전자 개수를 가공하여 종간에서 공통 또는 유일한 유전자를 선별하였다.

나. 전체 유전자 모델에서 Gene Family 분석

장미과에서 동일한 유전자 기능을 갖는 Gene Family 그룹을 분석하기 위해 PLAZA (Proost et al., 2015) 데이터베이스 Dicot 3.0 분석 결과에서 131,126개 Gene Family 그룹 데이터를 사용하였다. 각 종별로 Gene Family 그룹에 속해 있는 유전자 개수 및 Gene Family 그룹별 enrichment 단백질 도매인 분석 결과 등을 사용하였다. BLASTP 프로그램을 통해 31종 991,325개 단백질 서열에 대한 검색 서열 데이터베이스를 생성 후 E-value $1E^{-10}$ 파라미터에서 자생 왕벚나무, 매실, 체리에 대해 상동성 분석을 수행하였다. BLASTP 분석 결과에서 상동성 높은 31종 유전자와 해당 유전자

의 Gene Family 그룹 정보를 가공하고, 자생 왕벚나무 유전자에 상동성 있는 유전자의 Gene Family 그룹 개수를 계산하여 빈도가 높은 Gene Family 그룹을 선별하였다. 매실과 체리에 대해 동일한 Gene Family 분석을 수행하였다. 동일 Gene Family 그룹 내에 자생 왕벚나무, 복숭아, 매실, 체리에서 유전자 개수가 유의미하게 차이가 있는지 Z-test 통계 분석을 실시하여 P-value 값 0.0001 기준에서 Gene Family를 선별하였다.

다. 왕벚나무 전체 유전체에서 synteny 분석

자생 왕벚나무 스케폴드 유전체 서열을 8개 염색체 형태로 재배열하기 위해 복숭아, 매실, 체리별로 유전체 서열 일치성이 있고 상동성 있는 유전자에서 synteny 영역을 비교 분석 하였다. 자생 왕벚나무 유전체 서열 재배열은 벚나무속 중 유전체 완성도가 가장 높은 복숭아 유전체(The International Peach Genome Initiative, 2013)를 기반으로 분석을 수행하였다. 유전자 상에서 synteny 영역을 분석하기 위해 자생 왕벚나무 유전자의 단백질 서열과 복숭아 유전자의 단백질 서열을 BLASTP 프로그램을 이용하여 E-value $1E^{-10}$ 기준에서 상동성 분석을 하였다. 복숭아 유전체의 유전자 좌표 정보 GFF 파일, 자생 왕벚나무 스케폴드 상에서 유전자 좌표 정보 GFF 파일, BLASTP 결과를 함께 MCScanX toolkit (Wang et al., 2012) 프로그램을 이용하여 복숭아 염색체의 유전자 synteny 영역에서 해당하는 자생 왕벚나무 유전자를 선별하였다. MCScanX 분석에는 match score=50, match size=5, gap penalty=-1 파라미터를 적용하였다. 또한 자생 왕벚나무 스케폴드 서열을 NUCmer 프로그램을 이용하여 복숭아 유전체에 매핑하여 일치하는 영역을 분석하였다. 복숭아 유전체에 유전자 synteny 영역과 스케폴드 서열이 일치하는 영역을 비교 분석하여 8개 염색체에 속하는 pseudomolecule 형태로 배열된 유전체 서열과 유전자 좌표를 작성하였다. 2D Dot-plot 분석을 통해 8개 염색체상에서 배열된 자생 왕벚나무와 복숭아 synteny 영역이 선형을 나타내는지 분석하였다. 또한 체리와 매실 단백질 서열을 BLASTP 프로그램을 이용하여 상호 상동성 분석을 수행한 결과와 염색체상에 배열된 자생 왕벚나무 유전체 좌표 정보를 함께 MCScanX 프로그램 및 2D Dot-plot 분석하여 synteny 영역을 결정하였다.

라. 전체 유전자에서 Ks 분석

장미과 종 유전체의 진화적 분화 수준을 결정하기 위해 유전자 암호화 서열에서 synonymous substitution rate (Ks) 분석을 수행하였다. 유전체 synteny 영역에서 상동성 높은 유전자를 선별하기 위해 복숭아 28,702개, 매실 31,390개, 체리 43,673개, 딸기 34,809개, 사과 63,540개 단백질 서열을 사용하였다. BLASTP 프로그램을 사용하여 E-value $1E^{-10}$ 기준에서 자생 왕벚나무와 자생 왕벚나무, 자생 왕벚나무와 복숭아, 자생 왕벚나무와 매실, 자생 왕벚나무와 체리, 자생 왕벚나무와 딸기, 자생 왕벚나무와 사과의 상호간 상동성 분석을 수행하였다. MCScanX 분석 결과에서 공선적으로 존재하는 상동유전자 단백질 서열을 ClustalW v2.0 (Larkin et al., 2007) 프로그램을 이용하여 다중정렬 하였다. 또한 동일한 유전자 쌍의 CDS 서열을 가공하고 PAL2NAL v14 (Suyama et al., 2006) 프로그램을 사용하여 단백질 서열과 CDS 서열에서 codon 분석을 수행하였다. Ks 값을 계산하기 위해 codon 결과를 사용하여 PAML v4.8 (Yang, 2007) 프로그램에서 codeml 을 통해 계산하였다.

마. Diversification time 분석

벚나무속 분기 시간을 추정하기 위해 자생 왕벚나무, 복숭아, 매실, 체리, 장미과 사과와 딸기 그리고 외집단 콩과 *Medicago truncatula* (Young et al., 2011)에서 보존된 유전자를 사용하였다. OrthoMCL 분석 결과에서 벚나무속 single copy gene 그룹 1,608개를 선별하였다. BLASTP 프로그램을 이용하여 서열 커버리지 70% 이상, E-value $1E^{-10}$ 기준으로 *M. truncatula* 유전자와 상호간 상동성 분석을 수행하여 276개 single copy gene 그룹을 선별하였다. Bayesian evolutionary 분석을 위해 BEAST v1.7 (Drummond and Rambaut, 2007) 프로그램을 사용하였다. 입력 데이터는 GTR + I + G 모델로 BEAUti 인터페이스를 사용하여 가공하였고, 분석 데이터는 Yule speciation tree와 uncorrelated lognormal molecular clock model을 사용하여 BEAST 분석에 적용하였다. *Pentapetalae* 분기 추정에 사용한 방법(Moore et al., 2010)으로 100백만 년 전에서 107백만 년 전까지 균일 분포 장미과와 콩과의 공동 조상의 가계도 그룹을 포함하는 crown age를 한정하였다. 또한, 84.2~92.8백만 년 전에서 장미과의 분화시기(crown age), 51.6~65.2백만 년 전에서 벚나무속의 분화시기에 따라 각 균일 분포를 적용하였다 (Chin et al., 2014). 사후 확률 파라미터는 초기값에 10% 정도

를 보정한 20,000,000 번을 각각 MCMC 분석에 적용하였다. Tracer v1.5 (Rambaut et al., 2018)를 이용하여 두 분석 결과를 통합하고 TreeAnnotator v1.5.4 (Drummond and Rambaut, 2007)를 이용하여 maximum clade credibility tree를 작성하였으며, 계통수는 FigTree v1.4.2 (Rambaut) 프로그램을 이용하여 작성하였다.

10. 근연종 벚나무 속에서 변이 분석

자생 왕벚나무 전체 유전체에서 근연종 벚나무와의 SNP와 Insertion/Deletion (InDel) 변이를 탐색하기 위해 변이 분석 파이프라인을 구축하여 분석을 수행하였다. BWA index 프로그램 사용하여 자생 왕벚나무 전체 유전체 서열을 index 데이터베이스 서열로 생성하였다. BWA MEM 프로그램을 사용하여 -M 옵션을 적용, index 데이터베이스 서열에 각 근연종 벚나무별로 short read 서열을 매핑하고 SAM 파일로 저장하였다. Picard SortSam 프로그램을 이용하여 SORT_ORDER=coordinate 옵션을 적용해서 SAM 파일에 저장되어 있는 short read 매핑 위치를 정렬하였다. Picard MarkDuplicates 프로그램을 이용하여 REMOVE_DUPLICATES=true 옵션을 적용해서 PCR duplicate 오류 서열을 제거하였다. GATK v3.7 (McKenna et al., 2010) RealignerTargetCreator 프로그램을 사용하여 매핑 영역에서 서열을 재정렬 할 구간을 저장하였다. GATK IndelRealigner 프로그램을 사용하여 각 개체별로 저장한 재정렬 구간을 입력하고 consensusDeterminationModel=USE_READS 옵션을 적용하여 매핑 서열을 추가 정렬하였다. 재정렬된 BAM 파일을 모두 입력 데이터로 사용하여 GATK HaplotypeCaller 프로그램을 통해 SNP와 InDel 정보를 추출하였다. 전체 유전체에서 추출된 SNP의 haplotype phase를 결정하기 위해 BEAGLE v4.0 (Browning and Browning, 2007) 프로그램을 이용하여 오류 유전자형을 수정하였다.

모계와 부계에 유래된 유전자를 결정하기 위해 유전자 영역 내에서 부계 벚나무 또는 모계 올벚나무 short read 서열의 매핑 커버리지를 BEDTools v2.25 (Quinlan and Hall, 2010) 프로그램을 이용하여 계산하였고, phased SNP 유전자형이 부계 또는 모계 상호 대비하여 2배 이상 비율을 갖는 유전자를 선별하였다. 그리고 SNP 주석 정보를 분석하기 위해 SnpEff v4.3p (Cingolani et al., 2012) 프로그램을 이용하여 유전자 구조에 영향을 줄 수 있는 SNP와 InDel를 분석하였다.

근연종 벚나무 속에서 유연관계를 확인하기 위해 SNP 유전자형을 이용하여 주성

분 분석을 수행하였다. 주성분 분석에 사용되는 MDS 값은 VCFtools v4.2 (Danecek et al., 2011) 프로그램을 이용하여 PLINK 형식 데이터로 변환한 후 PLINK (Purcell et al., 2007) 프로그램에서 –genome, –cluster, –mds-plot 2 파라미터 값을 적용하여 계산하였다. 한편 근연종 벚나무속 SNP 변이 서열을 MAFFT (Nakamura et al., 2018) 프로그램을 이용하여 다중 정렬을 수행하였고, MEGA7 (Kumar et al., 2016) 프로그램에서 Maximum Likelihood (ML) 알고리즘을 이용한 계통수 분석을 수행하여 근연종 벚나무에서 유연관계를 분석하였다.

11. 자생 왕벚나무 전사체 분석

왕벚나무 Pxn-Jeju2의 잎, 꽃잎, 암술, 수술, 열매에서 부계 벚나무 또는 모계 올벚나무 유래 유전자의 조직 특이적 차등 발현 양상을 분석하였다. 각 5개 조직별로 3회 반복하여 생산한 mRNA-Seq 데이터를 STAR v2.5.2b (Dobin et al., 2013) 프로그램을 이용하여 전체 유전체 서열에 매핑하였다. 유전자 모델의 위치 정보 GFF 데이터와 매핑 결과 BAM 데이터를 함께 사용하여 HTSeq v0.6.0 (Anders et al., 2015) 프로그램을 통해 유전자 CDS 영역에서 발현 값을 추출하였다. DESeq2 (Love et al., 2014) 프로그램을 이용하여 발현 값에 대한 표준화를 수행하였고 상호 조직별로 비교하여 P-value 0.05 이하 기준으로 차등 발현 유전자를 선별하였다.

차등 발현 유전자에서 각각 부계와 모계 유전자를 mclust v5.4 (Scrucca et al., 2016)를 이용하여 적절한 K 값을 예측하여 K-mean 클러스터 분석을 수행하였다. 또한 K 개수로 나누어진 각 클러스터 내 차등 발현 유전자에 대해 DAVID (Huang da et al., 2009) 웹 프로그램을 이용하여 GO enrichment 분석을 수행하였다.

각 5개 조직에서 차등 발현하는 alternative splicing (AS) 전사체를 분석하기 위해 TopHat2 프로그램을 이용하여 유전자 모델의 위치 정보 GFF을 함께 사용하여 전체 유전체 서열에 mRNA-Seq 데이터를 매핑하였다. Cuffllinks 프로그램을 이용하여 유전자 모델 영역에 매핑된 전사체를 재조립하였다. Cuffdiff v2.2.1 (Trapnell et al., 2013) 프로그램을 이용하여 상호 조직별 차등 발현하는 전사체를 분석하였다. Cuffdiff2 결과를 spliceR (Vitting-Seerup et al., 2014) 프로그램을 사용하여 Exon skipping/inclusion (ESI), Mutually exclusive exon (MEE), Mutliple exon skipping/inclusion (MESI), Intron skipping/retention (ISI), Alternative 5‘ splice site

(A5), Alternative 3’ splice site (A3), Alternative transcription start site (ATSS), Alternative transcription terminating site (ATTS)로 구분되는 AS 모델을 판별하고, 각 조직에서 차등하게 발현하는 isoform 전사체를 선별하였다.

제 3 장 연구 결과 및 고찰

제 1 절 자생 왕벚나무 및 근연종 벚나무류 유전체 시퀀싱

1. 자생 왕벚나무 유전체 해독 서열

자생 왕벚나무 Pxn-Jeju2의 전체 유전체를 해독하기 위해 9개 라이브러리에 대해 NGS 시퀀싱을 수행하여 서열 데이터를 생산하고 품질관리를 수행하였다(Table 2).

20 kb 라이브러리는 PacBio RSⅡ 플랫폼을 사용하여 24 Cell에서 18.7 Gb 서열을 생산하였다. PacBio 서열에는 약 11%~15% 정도 시퀀싱 오류 서열이 존재한다(Korlach). 따라서 시퀀싱 어댑터를 제거한 후 PBcR 프로그램을 이용하여 오류 서열을 교정하여 11.5 Gb 서열을 가공하였다. 품질 관리를 수행한 PacBio subread 서열은 예측한 자생 왕벚나무 257 Mb 길이에 대비 44.9 X 커버리지에 해당한다.

500 bp 라이브러리는 300 bp PE 서열 길이로 Illumina MiSeq 플랫폼을 사용하여 총 33.6 Gb 서열을 생산하였다. Phred Score Q20 값을 기준으로 품질관리를 수행하여 총 24.5 Gb 서열을 가공하였다. 250 bp 라이브러리는 150 bp PE 서열 길이로 NextSeq 플랫폼을 사용하여 총 38.1 Gb 서열을 생산하였다. Phred Score Q20 값을 기준으로 품질관리를 수행하여 총 28.4 Gb 서열을 가공하였다.

3 kb, 5 kb, 10 kb, 15 kb 라이브러리는 150 bp MP 서열 길이로 NextSeq 플랫폼을 사용하여 각 21.5 Gb, 23.6 Gb, 21.2 Gb, 23.5 Gb로 총 90 Gb 서열을 생산하였다. PCR duplicate 제거 및 Phred Score Q20 기준으로 품질관리를 수행하여 79.4 X 커버리지에 해당하는 총 20.4 Gb 서열을 가공하였다.

20 kb 라이브러리는 150 bp MP 서열로 HiSeq 플랫폼을 사용하여 총 43.7 Gb 서열을 생산하였다. PCR duplicate 제거 및 Phred Score Q20 값을 기준으로 품질관리를 수행하여 9.3X 커버리지에 해당하는 총 2.4 Gb 서열을 가공하였다.

유전체 조립 커버리지를 높이기 위해 40 kb Fosmid 라이브러리는 55,200 클론에서 총 101.6 Gb의 PE 서열을 생산하였다. Fosmid 어댑터를 제거하고 유전체에 매핑 되지 않은 서열을 필터링 하여 전체 유전체의 55.8 X에 해당하는 총 14.3 Gb 서열을 가공하였다.

이상의 결과, 다양한 라이브러리와 NGS 시퀀싱 플랫폼을 사용하여 예측된 유전체 길이 대비 350.7 X에 해당하는 총 101.6 Gb의 고품질 서열을 왕벚나무 유전체 조립에 사용하였다.

Table 2. Statistics of genome sequence data of wild *P.* x *nudiflora* (Pxn-Jeju2) used in genome assembly

		Raw data			Filtered data		
Platform	Library	Read number	Total bases (bp)	Coverage(X)[a]	Read number	Total bases (bp)	Coverage(X)[a]
PacBio RSII	20 kb SMRTbell	1,960,335	18,769,924,156	73.0	1,739,058	11,549,004,047	44.9
Illumina MiSeq	500 bp PE	112,113,886	33,682,954,486	131.1	100,751,904	24,578,916,946	95.6
Illumina NextSeq	250 bp PE	252,294,608	38,096,485,808	148.2	202,181,044	28,444,220,254	110.7
	3 kb MP	144,800,142	21,550,093,231	83.9	57,592,272	5,466,809,896	21.3
	5 kb MP	159,465,808	23,660,690,862	92.1	57,727,120	5,467,614,507	21.3
	10 kb MP	143,411,270	21,292,872,596	82.9	49,169,962	4,717,174,465	18.4
	15 kb MP	156,596,064	23,541,425,847	91.6	47,214,352	4,745,256,985	18.5
Illumina HiSeq	20 kb MP	289,433,366	43,704,438,266	170.0	23,730,240	2,396,234,365	9.3
	40 kb Fosmid end (55,200 clones)	669,182,682	101,046,584,982	393.1	94,923,990	14,333,522,490	55.8
Total		1,929,258,161	325,345,470,234	1,265.6	635,029,942	101,698,753,955	350.7

PE, paired end; MP, mate paired end.

[a]Genome coverage was calculated with the haploid genome size of wild *P.* x *nudiflora* as 257 Mb.

2. 근연종 벚나무류 유전체 해독 서열

근연종 벚나무류 간에서 유전체 변이 차이와 특성을 분석하고자 제주도 자생지에 있는 왕벚나무와 근연종 벚나무류, 일본에서 채집한 소메이 요시노, 미국에서 채집한 소메이 요시노의 전체 유전체를 시퀀싱하였다(Table 3).

근연종 벚나무류는 500 bp 라이브러리를 제작하여 Illumina PE 시퀀싱을 수행하였고, PCR duplicate 제거 및 Phred Score Q20 값을 기준으로 동일한 품질관리를 수행하였다. Pxn-Jeju1, Pxn-Jeju3, Pxn-Jeju4, Pxn-Jeju5의 왕벚나무 4개체는 Pxn-Jeju2 왕벚나무의 유전체 예측 크기 대비 29.7X～39X 커버리지에 해당하는 8.4 Gb～10.8 Gb 서열을 가공하였다. 미국 워싱턴의 Pxy-US1과 Pxy-US2, 일본 도쿄의 Pxy-JP1과 Pxy-JP2 소메이 요시노는 Pxn-Jeju2 자생 왕벚나무 유전체의 예측 크기 대비 33.6X～121.9X 커버리지에 해당하는 9.5 Gb～30.2 Gb 서열을 가공하였다. 올벚나무, 벚나무, 사옥, 산벚나무는 각각 Pxn-Jeju2 자생 왕벚나무 유전체의 7.5X～30.8X 커버리지에 해당하는 2.3 Gb～8.5 Gb 서열을 가공하였다. 이상의 결과를 종합하면 자생 왕벚나무 Pxn-Jeju2를 제외한 전체 15개 벚나무류에서 품질관리를 수행하여 총 158.7 Gb 서열 데이터를 가공하였다.

Table 3. Summary of accessions and Illumina short-read data used in whole-genome resequencing analysis

Taxon	Name	Locality	Voucher	Sequencing	Bases (Gb)	Coverage (X)
P. x *nudiflora* (wild *P.* x *nudiflora*)	Pxn-Jeju1	Korean National Monument 159, No. 1, Bongae-dong, Jeju, Korea	159-1	HiSeq 4000	8.4	29.7
	Pxn-Jeju2	Korean National Monument 159, No. 2, Bongae-dong, Jeju, Korea	159-2	NextSeq, MiSeq	53.0	206.3
	Pxn-Jeju3	Korean National Monument 159, No. 3, Bongae-dong, Jeju, Korea	159-3	NextSeq	10.8	39.0
	Pxn-Jeju4	Jeju Province Monument No. 51, Gwaneumsa Temple, Jeju, Korea	51-2	HiSeq 4000	8.6	30.8
	Pxn-Jeju5	Jeju Province Local Tangible Heritage No. 3, Odeung-dong, Jeju, Korea	128	NextSeq	10.1	36.7
P. x *yedoensis* (Yoshino cherry)	Pxy-US1	Tidal Basin at the National Mall, Washington D.C., USA	NA69513	HiSeq 2500	9.9	37.2
	Pxy-US2	Tidal Basin at the National Mall, Washington D.C., USA	NA69515	HiSeq 2500	9.5	33.6
	Pxy-JP1	Koishikawa Botanical garden, Tokyo, Japan	JKS2206	HiSeq X Ten	29.2	109.6
	Pxy-JP2	Ueno park, Tokyo, Japan	JKS2215	HiSeq X Ten	30.2	121.9
P. pendula f. *ascendens*	Ppa-1	Harye-ri, Seogwipo-si, Jeju, Korea	60571	HiSeq 4000	8.5	30.8
	Ppa-2	Gwaneumsa Temple, Jeju, Korea	Gachon-P2	MiSeq	2.3	7.5
	Ppa-3	Bongae-dong, Jeju, Korea	Gachon-P6	MiSeq	3.9	15.1
P. jamasakura var. *jamasakura*	Pjj-1	Odeung-dong, Jeju, Korea	63375	HiSeq 4000	8.4	28.6
	Pjj-2	Odeung-dong, Jeju, Korea	Gachon-P3	MiSeq	2.9	7.5
P. jamasakura var. *quelpaertensis*	Pjq	Ara 1-dong, Jeju, Korea	63437	HiSeq 4000	7.7	24.0
P. sargentii	Psa	Odeung-dong, Jeju, Korea	63385	HiSeq 4000	8.3	28.5

3. 자생 왕벚나무 유전체 크기 추정 및 이형접합성 분석

벚나무속 왕벚나무, 벚나무, 올벚나무, 사옥, 산벚나무는 총 8개 염색체(2n = 2x = 16)를 갖고 있으며 전체 유전체의 크기와 염색체 구조가 매우 유사하다(Kim et al., 2012). 자생 왕벚나무 유전체 크기를 추정하기 위해 서열 품질관리 후 커버리지가 높은 short read 250 bp 라이브러리 28.4 Gb PE 데이터를 사용하였다.

자생 왕벚나무 유전체 크기 추정은 Jellyfish v2.1.3 (Marcais and Kingsford, 2011) 프로그램을 이용하여 short read 서열을 17 bp 길이로 나누어(K=17mer), K-mer 서열이 서로 구별되는 K-mer 서열 개수(X축)와 배수(Y축) 값을 계산한 히스토그램을 작성하였다(Figure 5). 그 결과 K-mer 분포는 이형접합성을 나타내는 두 개 피크로 나누어졌다. 이형접합(heterozygous) 서열 분포는 48배수로 K-mer 서열 4,073,522개에서 첫 번째 정점을 나타냈다. 동형접합(homozygous) 서열 분포는 98배수로 K-mer 서열 1,938,118개에서 두 번째 정점을 나타냈다. 동형접합 서열 대비 이형접합 서열 분포는 2배로 자생 왕벚나무 유전체는 잡종(hybrid)으로 확인되었다. 동형접합 정점에서의 분포 값을 사용하여 반수체(haploid) 유전체 크기는 257 Mb로 추정되었으며, 이형접합 정점에서의 분포 값을 사용하여 이배체(diploid) 유전체 크기는 525 Mb로 추정되었다. K-mer 분포 분석으로 추정된 유전체 크기는 flow cytometry assay 실험을 통해 보고된 결과와 유사했다. Flow cytometry 분석에 따르면 하나의 nucleotide 쌍의 평균 무게는 1.023×10^{-9} 이며, DNA의 1 pg 은 0.978×10^{9} bp에 해당한다(Dolezel et al., 2003). 왕벚나무의 DNA content (1C)는 0.29 pg 이며 978 Mb 길이를 곱하였고 유전체 크기는 284 Mb 로 추정되었다(Baek et al., 2018).

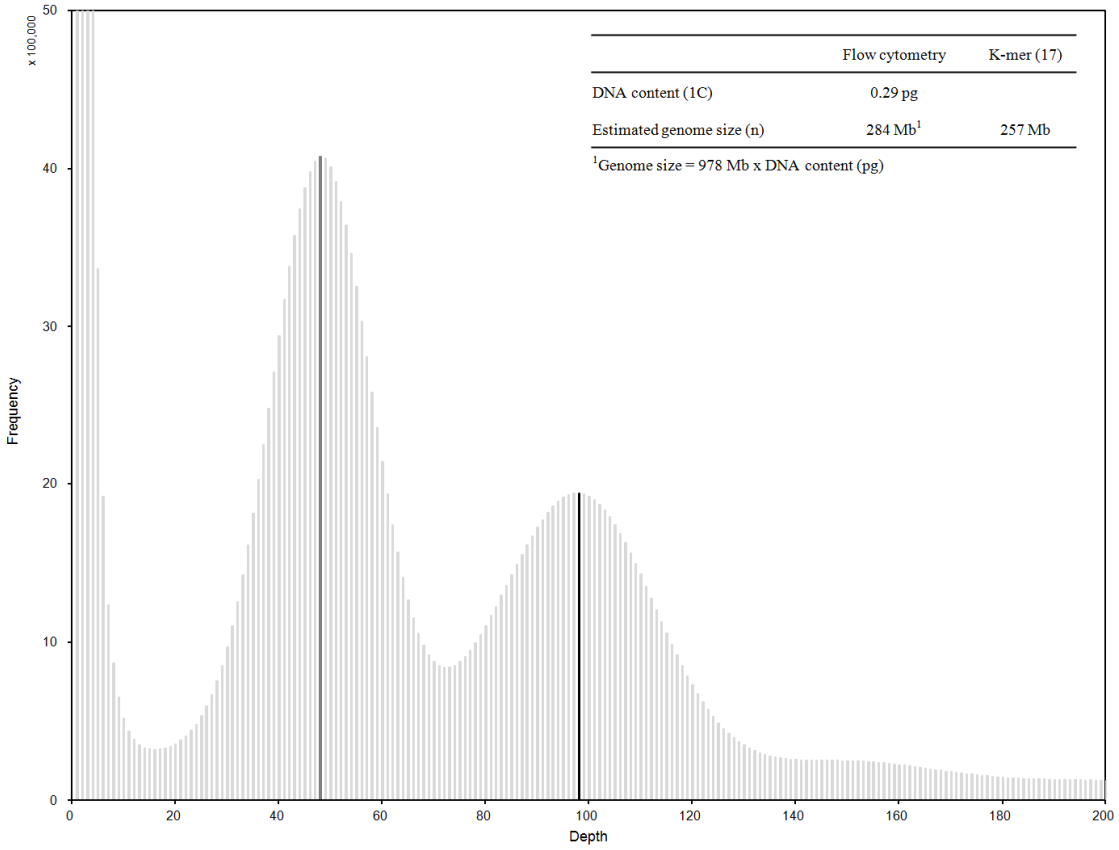
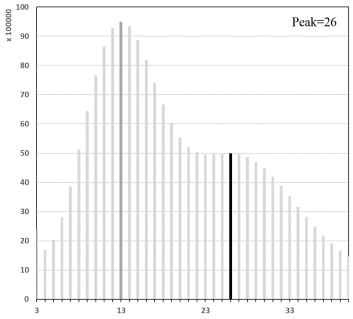


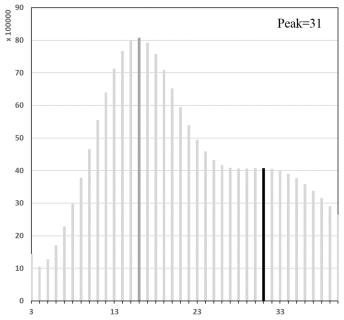
Figure 5. Estimation of the genome size of Pxn-Jeju2 based on K-mer analysis.

The graph represents the volume of K-17mer (Y-axis) plotted against the frequency at which it occurs (X-axis). The gray and black peaks correspond to heterozygous and homozygous reads, respectively. The upper right shows the estimated haploid genome size based on the homozygous K-mer peak as well as flow cytometry analysis.

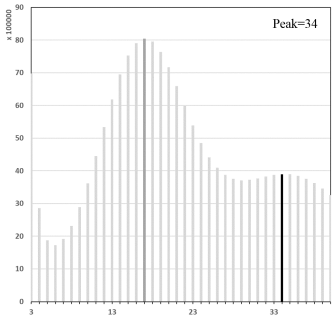
자생 왕벚나무 개체와 소메이 요시노 개체에서도 Pxn-Jeju2 왕벚나무처럼 이형접합성 특성이 나타나는지 K-mer 분포를 분석하였다. Pxn-Jeju1, Pxn-Jeju3, Pxn-Jeju4, Pxn-Jeju5 자생 왕벚나무와 Pxy-US1, Pxy-US2, Pxy-JP1, Pxy-JP2 소메이 요시노의 유전체 크기 추정은 Jellyfish 프로그램을 사용하였다. Illumina short read 서열을 K-mer 17을 적용하여 서로 구별되는 K-mer 서열 개수(X축)와 배수(Y축) 값을 계산한 히스토그램을 작성하였다(Figure 6). 자생 왕벚나무 개체와 소메이 요시노 개체들의 K-mer 분포는 모두 이형접합성을 나타내는 두 개의 피크로 나누어졌다. 예측된 반수체 크기는 248 Mb~282 Mb 범위로 Pxn-Jeju2와 매우 유사한 크기로 계산되었다(Table 4). 동형접합 대비 이형접합 분포는 약 2배로 자생 왕벚나무 개체와 소메이 요시노 개체의 유전체는 모두 잡종(hybrid)으로 확인되었다. 따라서 K-mer 분포 분석을 통해 자생 왕벚나무 및 소메이 요시노는 동배수성(homoploid) 잡종 유전체를 갖고 있는 것을 확인하였다.



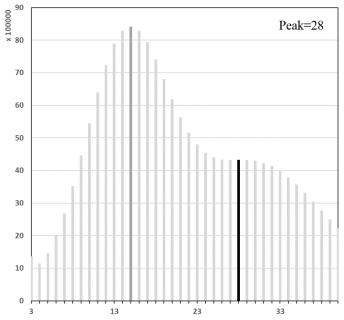
Pxn-Jeju1



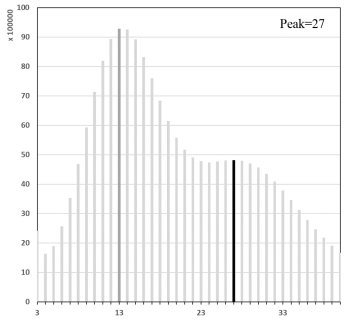
Pxy-US1



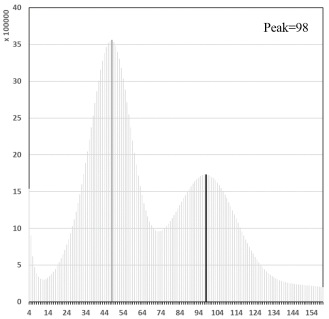
Pxn-Jeju3



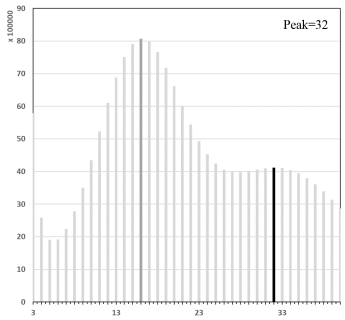
Pxy-US2



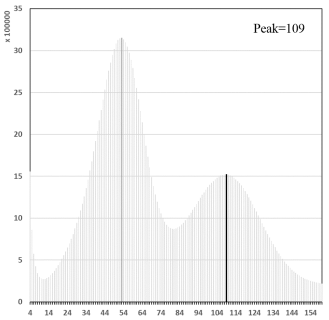
Pxn-Jeju4



Pxy-JP1



Pxn-Jeju5



Pxy-JP2

Figure 6. K-mer plots of wild *P.* x *nudiflora* (Pxn) and 'Yoshino cherry' (Pxy) accessions.

The volumes of Illumina K-mer (K=17 mer, Y axes) are plotted against the frequency where they occur (X axes). Gray line, heterozygous peak; black line, homozygous peak.

Table 4. Estimated genome size of wild *P.* x *nudiflora* (Pxn) and 'Yoshino cherry' (Pxy) accessions

Taxon	Depth		Genome Size	
	Heterozygous	Homozygous	Heterozygous	Homozygous
Pxn-Jeju1	13	26	564,205,380	282,102,690
Pxn-Jeju3	17	34	554,511,832	277,255,916
Pxn-Jeju4	13	27	582,697,410	280,558,012
Pxn-Jeju5	16	32	549,460,939	274,730,469
Pxy-US1	16	31	518,105,164	267,409,117
Pxy-US2	15	28	525,770,960	281,663,014
Pxy-JP1	48	98	544,611,285	266,748,384
Pxy-JP2	53	109	511,102,369	248,517,666

제 2 절 자생 왕벚나무 전사체 시퀀싱

자생 왕벚나무 유전체에서 유전자 모델 예측과 조직간 차등 발현 유전자를 분석하기 위해 전사체 시퀀싱을 수행하였다. 잎, 꽃잎, 암술, 수술, 열매 조직에서 차등 발현 유전자 선별의 재현성을 높이기 위해 동일 조직을 3회 반복하여 전사체를 시퀀싱하고, 꽃눈(floral bud)과 잎눈(foliar bud) 조직을 추가하여 Illumina NextSeq과 MiSeq 플랫폼을 이용하여 총 48.8 Gb 전사체 서열을 해독하였다(Table 5). Illumina flow-cell에서 높은 클러스터 밀도는 GC-rich 서열을 억제한다(Aird et al., 2011). 이때 Phred Score Q30 기준으로 전사체 서열을 품질관리를 하면 GC-bias가 감소되어 전사체 조립 품질을 높일 수 있고 뿐만 아니라 정확도 높은 fold change 값을 추정할 수 있다(Li et al., 2013). 따라서 7개 조직 전사체 데이터를 Phred Score Q30 기준으로 동일한 품질관리를 수행하여 23.7 Gb 서열을 가공하였다.

Table 5. Statistics of transcriptome sequence data of wild *P.* x *nudiflora* (Pxn-Jeju2) used in this study

Tissue	Replicate	Illumina Platform	Raw data		Filtered data	
			Read number	Total Bases (bp)	Read number	Total Bases (bp)
Leaf	1	NextSeq	32,847,784	2,496,431,584	17,921,878	1,293,561,345
	2	NextSeq	34,186,814	2,598,197,864	19,200,738	1,392,808,613
	3	MiSeq	12,401,916	3,246,484,042	7,532,974	1,442,110,068
Petal	1	NextSeq	30,489,072	2,317,169,472	16,543,796	1,196,620,785
	2	NextSeq	27,271,844	2,072,660,144	11,885,078	855,713,305
	3	MiSeq	14,074,138	3,742,118,404	8,709,350	1,683,595,272
Pistil	1	NextSeq	27,938,988	2,123,363,088	15,052,332	1,088,938,243
	2	NextSeq	31,109,160	2,364,296,160	16,676,224	1,202,309,049
	3	MiSeq	13,692,674	3,950,924,684	8,761,358	1,913,564,090
Stamen	1	NextSeq	30,035,180	2,282,673,680	16,623,532	1,201,110,047
	2	NextSeq	30,198,690	2,295,100,440	15,044,868	1,083,491,530
	3	MiSeq	15,056,954	3,913,135,104	9,696,594	1,872,727,978
Berry	1	NextSeq	30,143,606	2,290,914,056	16,326,446	1,178,548,167
	2	NextSeq	32,621,776	2,479,254,976	18,106,088	1,310,520,863
	3	MiSeq	13,363,880	3,597,437,214	8,559,936	1,713,881,345
Floral bud	1	MiSeq	12,563,728	3,334,081,569	8,122,410	1,594,467,047
Foliar bud	1	MiSeq	14,154,026	3,765,756,124	8,902,212	1,744,893,795
Total			402,150,230	48,869,998,605	223,665,814	23,768,861,542

제 3 절 자생 왕벚나무 유전체 조립

1. 자생 왕벚나무 전체 유전체 *de novo* assembly

자생 왕벚나무 Pxn-Jeju2 유전체는 이형접합도가 높아 long read 서열에 대한 OLC 조립 방법으로 전체 유전체를 조립하였다. 먼저 Pxn-Jeju2 유전체 길이 대비 73X 커버리지에 해당하는 18.7 Gb PacBio subread 서열을 FALCON 어셈블러를 이용하여 컨티그로 조립하였다. FALCON 조립 결과 컨티그 서열 개수 총 4,700개, N50 서열 길이 124.4 kb의 총 315.3 Mb 조립 서열을 획득하였다. 조립 서열 중 서열 길이가 1 kb 미만 서열은 제거하였고, 조립 서열에 존재하는 비핵 유전체 서열을 제거하기 위해 NCBI에 등록한 엽록체 유전체 (Genbank accession NC_026980) 서열을 참조서열로 하여 global alignment 분석을 통해 엽록체 서열에 50% 이상 일치성을 갖는 총 182개 컨티그 서열을 제거하였다. 이후 조립 커버리지를 높이기 위해 3 kb, 5 kb, 10 kb, 15 kb, 20 kb 그리고 40 kb Fosmid End 서열 144.5X 커리버지에 해당하는 총 37.1 Gb MP 데이터를 사용하여 스케폴드 서열을 작성하였다. 조립 서열에 존재하는 공백을 교정하기 위해 교정된 44.9X 커버리지에 해당하는 11.5 Gb PacBio 서열을 사용하였고, 조립 오류 서열을 교정하기 위해 Illumina short read PE 서열과 교정된 PacBio long read 서열을 사용하였다. 조립 결과 스케폴드 서열 개수 총 3,185개, N50 길이 198.9 kb로 구성된 총 323.7 Mb의 조립 서열을 작성하였다(Table 6).

최종 조립 결과는 최초의 컨티그 조립에 비해 1,515개 서열이 추가 조립되었고, N50 길이는 74.5 kb 증가하였다. 전체 조립 서열 길이는 추정한 반수체 유전체 크기 257 Mb 대비 120.0%를 차지하며, short read 서열 매핑을 통해 추출된 SNP는 총 2.5 Mb로서 추정한 반수체 길이 대비 1.1%를 포함하는 이형접합 유전체 서열을 포함하고 있다.

전사체 조립으로 선별된 총 84,378개 unigene 서열을 유전체 조립 서열에 매핑하여 유전체 서열과 전사체 서열이 일치하는 유전자 영역을 분석하였다. 전체 unigene 중에서 93.2%에 해당하는 총 78,675개 unigene 서열이 전체 유전체 조립 서열에 매핑되어 그 결과 전체 유전체 조립 서열이 유전자 영역의 93.2% 이상을 포함하고 있는 것으로 나타났다(Table 7).

Table 6. Summary statistics of the draft genome assembly of wild *P.* x *nudiflora*

	Contig		Scaffold	
	Length (bp)	Number	Length (bp)	Number
N90	38,284	2,435	54,586	1,700
N80	59,290	1,770	88,524	1,239
N70	81,939	1,312	124,582	934
N60	106,886	973	158,837	702
N50	132,585	706	198,954	519
Longest	773,088		960,226	
Overall (>1 kb)	318,739,121	4,292	323,781,369	3,185

Table 7. Evaluation of gene space coverage of the wild *P.* x *nudiflora* genome using transcriptome unigenes

Unigene length cutoff	Total number	Total match number	Matched (%)	≥50%(50%～100%) of a sequence covered by a scaffold		≥90%(90%～100%) of a sequence covered by a scaffold	
				Number	Percent (%)	Num6ber	Percent (%)
All	84,378	78,675	93.2	76,356	90.5	71,571	84.8
Length > 300 bp	74,463	69,507	93.3	67,370	90.5	63,142	84.8
Length > 500 bp	57,815	54,434	94.2	52,637	91.0	49,156	85.0
Length > 1,000 bp	35,552	33,782	95.0	32,474	91.3	30,268	85.1

2. 조립 알고리즘별 유전체 조립 서열 비교 분석

이형접합성이 높은 자생 왕벚나무 전체 유전체 조립을 위해 사용된 FALCON 어셈블러 이외 OLC 알고리즘을 적용한 Celera 어셈블러, DBG 알고리즘을 적용한 ALLPATH-LG 어셈블러를 사용하여 전체 유전체 조립 결과에서 완성도를 비교 분석하였다(Table 8). ALLPATH-LG 어셈블러를 이용한 전체 유전체 조립에는 Illumina short read 서열을 사용하였다. Celera 어셈블러를 이용한 hybrid 방식의 전체 유전체 조립에는 Illumina short read 서열과 시퀀싱 오류 서열이 교정된 PacBio subread 서열을 사용하였다. 입력 데이터 PacBio subread 교정 서열은 Celera 어셈블러에서 권장하는 전체 교정 서열 중 25X 해당하는 총 개수 4,594개, 총 길이 125 Mb, N50 길이 26,685 bp로 가장 긴 서열을 사용하였다. ALLPATH-LG 어셈블러를 사용한 전체 유전체 조립 결과에서 컨티그 총 개수는 17,279개, 컨티그 총 길이는 410,681,986 bp, N50 서열 길이는 50,284 bp였다. Celera 어셈블러를 사용한 전체 유전체 결과에서 컨티그 총 개수는 29,033개, 컨티그 총 길이는 199,357,221 bp, N50 서열 길이는 9,064 bp였다. 조립 서열 커버리지를 비교한 결과 추정된 자생 왕벚나무 반수체 257 Mb 크기 대비 ALLPATH-LG 컨티그 서열은 1.6배이며, FALCON 컨티그 서열 보다 약 91 Mb 더 많은 서열이 조립되었다. Celera 컨티그 서열의 커버리지는 추정된 자생 왕벚나무 반수체 257 Mb 크기 대비 0.8배 이며, FALCON 컨티그 서열보다 약 119 Mb 적게 서열이 조립되었다. 컨티그 N50 길이를 비교한 결과 FALCON 컨티그 N50 길이는 132,585 bp 로 가장 길게 조립 되었고 ALLPATH-LG와 Celera에서 N50 길이는 100 kb를 넘지 못하였고, N50~N90에서 FALCON 평균 길이 대비 ALLPATH-LG는 2.7배, Celera 는 13.2배 짧은 길이를 보였다.

컨티그 서열 일치성을 비교 분석하기 위해 NUCmer 프그램을 사용하여 FALCON 컨티그 조립 서열을 참조 서열로 하여 ALLPATH-LG와 Celera 컨티그 서열을 상호 비교하였다(Table 9). FALCON 컨티그 서열에 일치하는 ALLPATH-LG 컨티그 개수는 전체의 98.6%에 해당하는 17,037개, 일치하지 않는 컨티그 개수는 242개였으며, Celera 컨티그 개수는 전체의 98.3%에 해당하는 28,592개, 일치하지 않는 컨티그 개수는 504개였다. ALLPATH-LG 컨티그 서열은 FALCON 컨티그 조립 서열 길이 318,739,121 bp의 96.7%에 해당하는 308,168,923 bp 컨티그 서열이 정렬되었고, 약 10 Mb 서열이 일치하지 않았다. Celera 컨티그 서열은 FALCON 컨티그 서열의 78.3%

에 해당하는 249,689,341 bp 컨티그 서열이 일치성 있게 정렬되었고, 약 69 Mb 서열이 일치하지 않았다. 이상의 결과를 종합할 때, 자생 왕벚나무 전체 유전체 조립 과정에서 PacBio subread 서열만을 사용하여 FALCON 어셈블러를 통해 조립한 결과가 조립 서열 개수가 적고, N50 크기가 가장 컸으며, 서열 연속성이 가장 높아 유전체 조립 완성도가 가장 뛰어났다.

Table 8. Comparison of the genome assembly statistics between OLC- and DBG-based genome assemblers

Contig	FALCON		ALLPATH-LG		Celera	
	Length (bp)	Number	Length (bp)	Number	Length (bp)	Number
N90	38,284	2,435	13,276	8,448	3,478	20,542
N80	59,290	1,770	22,461	6,100	4,769	15,679
N70	81,939	1,312	31,379	4,561	6,030	11,964
N60	106,886	973	40,386	3,408	7,439	8,988
N50	132,585	706	50,284	2,493	9,064	6,555
Longest	773,088		356,966		296,475	
Overall	318,739,121	4,292	410,681,986	17,279	199,357,221	29,033

Table 9. Summary statistics of sequence coverage between OLC- and DBG-based genome assembly

	FALCON	ALLPATH-LG	FALCON	Celera
Total Sequences	4,292	17,279	4,292	29,033
Aligned Sequences	4,197	17,037	4,239	28,529
Unaligned Sequences	95	242	53	504
Total Bases	318,739,121	410,681,986	318,739,121	199,357,221
Aligned Bases	308,168,923	363,830,711	249,689,341	181,774,180
Unaligned Bases	10,570,198	46,851,275	69,049,780	17,583,041

제 4 절 자생 왕벚나무 유전체에서 반복서열 및 유전자 예측

1. 자생 왕벚나무 유전체에서 반복서열 분석

자생 왕벚나무 Pxn-Jeju2의 조립 서열에 존재하는 반복서열을 예측하기 위해 다양한 분석법을 적용한 프로그램을 이용하여 반복서열 종류와 빈도를 분석하였다(Table 10). 먼저 이미 알려진 다른 식물 종의 반복서열을 자생 왕벚나무 조립 서열에 상동성 분석으로 예측하는 RepeatMasker 프로그램을 사용하였다. 또한 조립 서열에서 반복서열 모델을 예측하고 알려진 반복서열과 비교 분석을 통해 판별하는 RepeatModeler 프로그램도 사용하였다. RepeatMasker와 RepeatModeler에서 예측한 반복서열 개수는 총 581,236개이며 반복서열 길이는 총 179.6 Mb였다. 자생 왕벚나무 조립 서열에서 가장 높은 비율을 갖는 반복서열 종류는 Long Terminal Repeat (LTR)로서 총 길이는 72.6 Mb에 달해 유전체 길이 대비 22.8%에 해당했다. 전체 반복서열 종류 중에서 가장 높은 비율을 차지하는 것은 LTR의 하나인 Gypsy 패밀리로 총 길이는 41.8 Mb 이며 전체 반복서열의 13.1%에 해당했다. 두 번째는 Copia 패밀리로서 총 길이 26.4 Mb이며 전체 반복서열의 8.3%를 차지했다. 한편 DNA transposon 종류에서 CMC-EnSpm 패밀리는 총 길이 17.0 Mb이며 전체 반복서열의 5.3%로서 세 번째로 높은 비율을 차지했다. LTR_finder 프로그램을 이용하여 LTR 영역을 추가 분석한 결과 34.7 Mb의 반복서열을 예측하였다. miRbase (Kozomara and Griffiths-Jones, 2014)에서 식물에 존재하는 miRNA 서열을 수집한 후 왕벚나무 유전체 조립 서열에 매핑하여 44.7 kb에 해당하는 총 419개 miRNA 영역을 예측하였다. 이상의 분석 방법을 이용하여 반복서열을 예측한 결과에서 반복서열 위치가 중첩되는 영역을 통합하였다. 그 결과 자생 왕벚나무 유전체에 존재하는 반복서열은 총 길이 150.7 Mb로 전체 조립 서열 대비 47.2% 를 차지하는 것으로 나타났다.

Table 10. Summary of repetitive sequences identified in the draft genome of wild *P.* x *nudiflora*

Class	Family	Count	Base (bp)	Percent
DNA		22,801	4,410,071	1.38%
	CMC-Chapaev	56	30,022	0.01%
	CMC-EnSpm	27,007	17,039,933	5.34%
	Dada	482	139,616	0.04%
	Kolobok-Hydra	113	55,814	0.02%
	Maverick	1,505	224,467	0.07%
	MuLE-MuDR	25,340	8,962,262	2.81%
	Novosib	46	3,206	0.00%
	P	2	284	0.00%
	PIF-Harbinger	13,277	4,700,780	1.47%
	PIF-ISL2EU	107	14,127	0.00%
	Sola	400	53,558	0.02%
	TcMar	553	93,661	0.03%
	TcMar-Stowaway	19	888	0.00%
	TcMar-Tigger	74	35,001	0.01%
	hAT	262	93,134	0.03%
	hAT-Ac	11,833	3,699,185	1.16%
	hAT-Charlie	1,112	186,134	0.06%
	hAT-Tag1	17,582	4,075,560	1.28%
	hAT-Tip100	15,894	3,739,543	1.17%
Total of DNA Repeats		138,465	47,557,246	14.90%
LINE		1	70	0.00%
	DRE	41	15,976	0.01%
	Jockey	1,264	171,243	0.05%
	L1	10,718	5,013,156	1.57%
	L1-Tx1	264	76,471	0.02%
	L2	133	24,774	0.01%
	Penelope	78	18,711	0.01%
	R1	391	96,986	0.03%
	R2	3	149	0.00%
	RTE-BovB	2,927	646,513	0.20%
	RTE-X	5	609	0.00%
	Tad1	211	37,392	0.01%
Total of LINE Repeats		16,036	6,102,050	1.91%
LTR		7,233	1,484,242	0.46%
	Caulimovirus	1,746	2,212,239	0.69%
	Copia	34,740	26,429,739	8.28%
	DIRS	1	50	0.00%
	ERV1	1,668	372,996	0.12%
	ERVK	278	171,658	0.05%
	Gypsy	56,260	41,886,435	13.12%
	Pao	174	53,212	0.02%
Total of LTR Repeats		102,100	72,610,571	22.75%
SINE		554	90,328	0.03%
	B2	2,302	170,956	0.05%
	ID	800	51,059	0.02%
	RTE	2	189	0.00%
	U	69	15,456	0.00%
	tRNA	3,879	350,258	0.11%
	tRNA-RTE	586	111,372	0.03%
Total of SINE Repeats		8,192	789,618	0.25%
Other	Composite	2	271	0.00%
RC	Helitron	7,676	3,776,126	1.18%
Retroposon		1	53	0.00%
Low complexity		30,917	1,596,229	0.50%
Satellite		935	207,391	0.06%
Simple repeat		137,962	5,850,463	1.83%
rRNA		788	2,258,476	0.71%
snRNA		164	46,813	0.01%
Unknown		137,998	38,858,329	12.17%
Total of RepeatMasker and RepeatModeler		581,236	179,653,636	56.28%
Total of LTR_Finder		3,990	34,797,344	10.90%
Total of miRBase		419	44,761	0.01%
Total non-redundant repeat		320,765	150,775,941	47.23%

2. 자생 왕벚나무 유전체에서 유전자 모델을 예측

자생 왕벚나무 유전자 모델은 *ab-initio* 유전자 예측, 근연종 단백질 서열의 상동성 비교, 전사체 서열을 매핑하는 방법을 사용하여 예측하였다(Table 11). *ab-initio* 유전자 예측은 GlimmerHMM과 SNAP 프로그램에서 제공하는 애기장대 모델을 선택하여 수행하였다. 또한 신뢰도 높은 유전자 예측을 위해 BRAKER1 프로그램을 사용하여 mRNA-Seq 데이터를 유전체에 매핑 후 유전자 영역을 예측하였다. 전체 *ab-initio* 유전자 예측 결과 GlimmerHMM에서 39,466개, SNAP에서 55,840개 유전자가 예측되었다. 예측된 유전자의 평균 길이는 1,285~1,951 bp이며 유전자 당 엑손 개수는 3.4~3.9개를 포함하고 있다. 전체 근연종 단백질 서열 비교를 통한 유전자 예측 결과에서 매실 40,439개, 복숭아 37,543개, 애기장대 5,852개 그리고 딸기 10,147개 유전자 상동성이 있는 것으로 예측되었다. 예측된 유전자의 평균 길이는 3,113~3,975 bp이며 유전자 당 엑손 개수는 4.3~5.2개를 포함하고 있다. 전체 전사체 서열 매핑을 통한 유전자 예측 결과에서 PASA 119,754개, Cufflinks 462,37개 유전자가 예측되었다. 예측된 유전자의 평균 길이는 3,030~3,138 bp 이며 유전자 당 엑손 개수는 3.9~4.6개를 포함하고 있다. 근연종 단백질 서열을 사용한 Exonerate, 전사체 서열을 사용한 Cuffllinks와 PASA 분석 결과에서 유전자와 인트론의 평균 길이는 *ab-initio* 분석 방법으로 예측한 경우보다 약 2배 수준으로 길게 예측되었다. 각 분석법을 통해 생성된 유전자 위치 정보를 EVidenceModeler (EVM) 프로그램을 사용하여 유전자 영역이 일치하는 총 41,294개 유전자 모델을 확립하였다.

자생 왕벚나무의 전체 유전자 모델에서 단백질 서열을 BLASTP 프로그램을 이용하여 NCBI에서 제공하는 RefSeq Plant, NR, 장미과를 포함한 근연종 단백질 서열, UniProt 서열 데이터베이스를 대상으로 상동성 높은 유전자를 검색하여 주석 정보를 수집하였다. 또한 InterProScan 단백질 도메인 분석을 통해 주석 정보를 수집하였다. 그리고 mRNA-Seq 서열을 조립 유전체 서열에 매핑 후 유전자 영역에서의 발현 값을 추출하여 유전자의 발현 여부를 분석하였다. 근연종 유전자 모델을 대상으로 한 주석 검색 결과에서 왕벚나무 유전자 모델은 매실 유전자 모델과 84.0%의 높은 유전자 일치성을 보였다(Table 12). RefSeq Plant, NR, Uniprot, InterPro 서열 데이터베이스 대상으로 한 주석 검색 결과 왕벚나무 유전자 모델의 84.3%에 대한 주석을 확인하였다. 따라서 유전자 주석 검색 결과를 통합하여 왕벚나무 유전자 모델 41,294개 중 유전자

기능이 알려진 왕벚나무 유전자 모델은 총 37,444개(90.7%)이며, 기능이 알려지지 않은 유전자 모델은 총 1,781개(4.3%), 가상의 유전자 모델은 총 2,069개(5%)로 확인하였다.

Table 11. Statistics of gene models predicted from the draft genome of wild *P.* x *nudiflora*

Tools	Gene Model	Total gene count	Average gene length (bp)	Average exon length (bp)	Average intron length (bp)	Exon per gene count
GlimmerHMM	*Arabidopsis thaliana*	39,466	1,951	225	371	3.9
SNAP	*Arabidopsis thaliana*	55,840	1,285	200	244	3.4
BRAKER1	mRNA-seq	76,157	1,844	277	349	3.5
Exonerate	*Prunus mume*	40,439	3,685	254	777	4.3
	Prunus persica	37,543	3,975	242	717	4.9
	Arabidopsis thaliana	5,852	3,113	201	486	5.2
	Fragaria vesca	10,147	3,866	235	635	5.2
Cufflinks	mRNA-seq	46,237	3,138	339	433	4.6
PASA	mRNA-seq	119,754	3,030	375	536	3.9
EVM	consensus	41,294	2,154	220	362	4.3

Table 12. Annotation statistics of the wild *P.* x *nudiflora* gene set

Database type	Number genes	Percent (%)
A. thaliana gene models	28,020	67.9
P. mume gene models	34,692	84.0
P. persica gene models	32,729	79.3
F. vesca gene models	31,516	76.3
M. x *domestica* gene models	31,572	76.5
RefSeq Plant	34,807	84.3
NCBI NR	34,790	84.2
Uniprot	22,047	53.4
InterPro	30,930	74.9
mRNA-seq reads of Pxn	33,802	81.9
Known	37,444	90.7
Unknown	1,781	4.3
Hypothetical	2,069	5.0
Total	41,294	100

3. 왕벚나무 근연종 유전체에서 반복서열 및 유전자 모델 비교 분석

Prunus 속에 속해 있는 왕벚나무의 근연종 복숭아, 체리, 매실 각 유전체의 특성을 비교하기 위해 유전자 모델과 반복서열을 분석하였다(Table 13). 복숭아와 체리, 매실의 유전자 모델은 각각 The International Peach Genome Initiative (2013), Shirasawa 등 (2017), 그리고 Zhang 등 (2012)의 결과를 사용하였다. 한편, 복숭아, 체리, 매실에서 일관성 있는 반복서열 발굴을 위해 자생 왕벚나무 반복서열 분석에 사용한 분석 방법을 동일하게 적용하여 반복서열을 분석하였다. 각각의 유전체에서 예측된 반복서열 비율은 복숭아 38.7%, 매실 30.0%, 체리 38.1%이며, 자생 왕벚나무의 반복서열 비율 대비 1.2~1.6배로 큰 차이가 없는 것으로 나타났다. DNA TE 클래스의 반복서열 총 길이는 자생 왕벚나무 대배 1.2~1.9배이며, RNA TE 클래스의 반복서열 총 길이는 자생 왕벚나무 대비 1.3~1.6배였다. 자생 왕벚나무 유전자 모델은 총 41,294개로서 매실 유전자 모델 개수 대비 1.3배, 복숭아 유전자 모델 개수 대비 1.5배로 예측된 유전자 개수가 많았으나, 체리 유전자 모델의 0.9배로 다소 적었다. 자생 왕벚나무 평균 유전자 길이는 2,154 bp이며 유전체에 비해 짧은 엑손 길이로 인해 *Prunus*속 근연종 평균 유전자 길이에 비해 140 bp~453 bp 정도 평균 유전자 길이가 짧게 예측되었다. 자생 왕벚나무 평균 유전자 밀도는 유전자 당 7.7 kb이며 매실에서 평균 유전자 밀도 7.6 kb와 매우 유사한 수준이었다. 이상의 결과를 종합할 때, 왕벚나무와 근연종 벚나무류의 반복서열과 유전자 모델을 비교 분석한 결과는 큰 차이가 없었으며 *Prunus* 속의 유전체 조성은 매우 유사한 것으로 판단되었다.

Table 13. Comparison of repetitive sequences and annotated protein-coding genes in the draft assemblies of four *Prunus* genomes

Genome	Characteristics	*P.* x *nudiflora*	*P. avium*	*P. mume*	*P. persica*
Draft sequences	Size (Mb)	318.7	272.4	237.2	227.2
Repetitive sequences	No. RNA genes[a]	2,187	729	1,541	1,243
	DNA TE (Mb)	47.6	26.5	25.5	38.7
	RNA TE (Mb)	79.5	56.4	49.9	62.6
	Simple repeats (Mb)	5.9	5.3	4.2	4.1
	Other repeats (Mb)	5.6	3.9	3.3	2.7
	Total non-redundant bases (Mb)	150.8	103.9	71.1	88
Protein coding genes	Total number	41,294	43,673	31,390	27,864
	Avr. gene size (bp)	2,154	2,294	2,514	2,607
	No. exons per gene	4.3	3.6	4.6	5.1
	Avr. exon size (bp)	220	248	249	243
	Avr. intron size (bp)	362	417	380	317
	Avr. gene density (kb/gene)	7.7	6.2	7.6	8.2

Statistics for *P. avium*, *P. mume*, and *P. persica* are based on Shirasawa et al. (2017), Zhang et al. (2012), and The International Peach Genome Initiative (2013), respectively, and repetitive sequences were recalculated using the same criterion used to annotate the *P.* x *nudiflora* genome.

[a]Sequences encoding rRNA, tRNA, and miRNA were considered.

제 5 절 장미과 비교 유전체 분석

1. 장미과 유전자 모델에서 오솔로그 유전자 분석

장미과에 속해 있는 복숭아, 체리, 매실, 사과, 딸기의 유전자 모델과 자생 왕벚나무 유전자 모델에서 서열 상동성이 높은 오솔로그 유전자 그룹을 비교 분석하였다. 각 유전체에서 예측된 단백질 서열을 BLASTP 프로그램을 사용하여 서열 상동성 분석을 수행하고 OrthoMCL 프로그램을 통해 총 6종간에서 오솔로그 그룹을 분류했다. 그 결과, 자생 왕벚나무와 장미과 유전체에서 공통 오솔로그 그룹은 9,221개, 오솔로그 유전자는 105,579개로 전체 유전체 모델 총 243,409개 대비 43.4%에 해당했다(Figure 7). 장미과 각 유전체별로 자생 왕벚나무와 공통 오솔로그 그룹이 가장 적은 종은 딸기로 오솔로그 그룹 59개, 오솔로그 유전자 307개였으며, 공통 오솔로그 그룹이 가장 많은 종은 체리로 오솔로그 그룹 1,164개, 오솔로그 유전자 3,402개였다. 자생 왕벚나무에서만 존재하는 오솔로그 그룹에서 유전자 주석 정보를 검색하여 유전자 기능이 알려진 오솔로그 그룹을 분석했다. 자생 왕벚나무에서만 존재하는 오솔로그 그룹은 총 678개, 오솔로그 유전자 총 1,723개(0.7%)이며, 유전자 기능이 알려진 오솔로그 그룹은 273개, 유전자 기능이 확립되지 않는 오솔로그 그룹은 97개, 유전자 기능이 확인되지 않은 오솔로그 그룹 308개로 분류 되었다.

Gene Ontology (GO) enrichment 분석을 통해 자생 왕벚나무에서 유일한 오솔로그 유전자의 유전자 기능 범주 특성을 분석하였다(Figure 8). Biological Process (BP)에서 기능 범주 phosphorylation에 관련된 유전자가 15개로 높은 빈도를 보였고, 이중 LRR receptor-like serine/threonine-protein kinase, wall-associated receptor kinase 기능을 갖는 유전자는 각각 6개, 4개였다. 또한 유의미한 기능 범주 defense response에 관련된 유전자는 12개로 대부분 disease resistance, receptor-like serine/threonine-protein kinase 기능을 갖는 유전자였다.

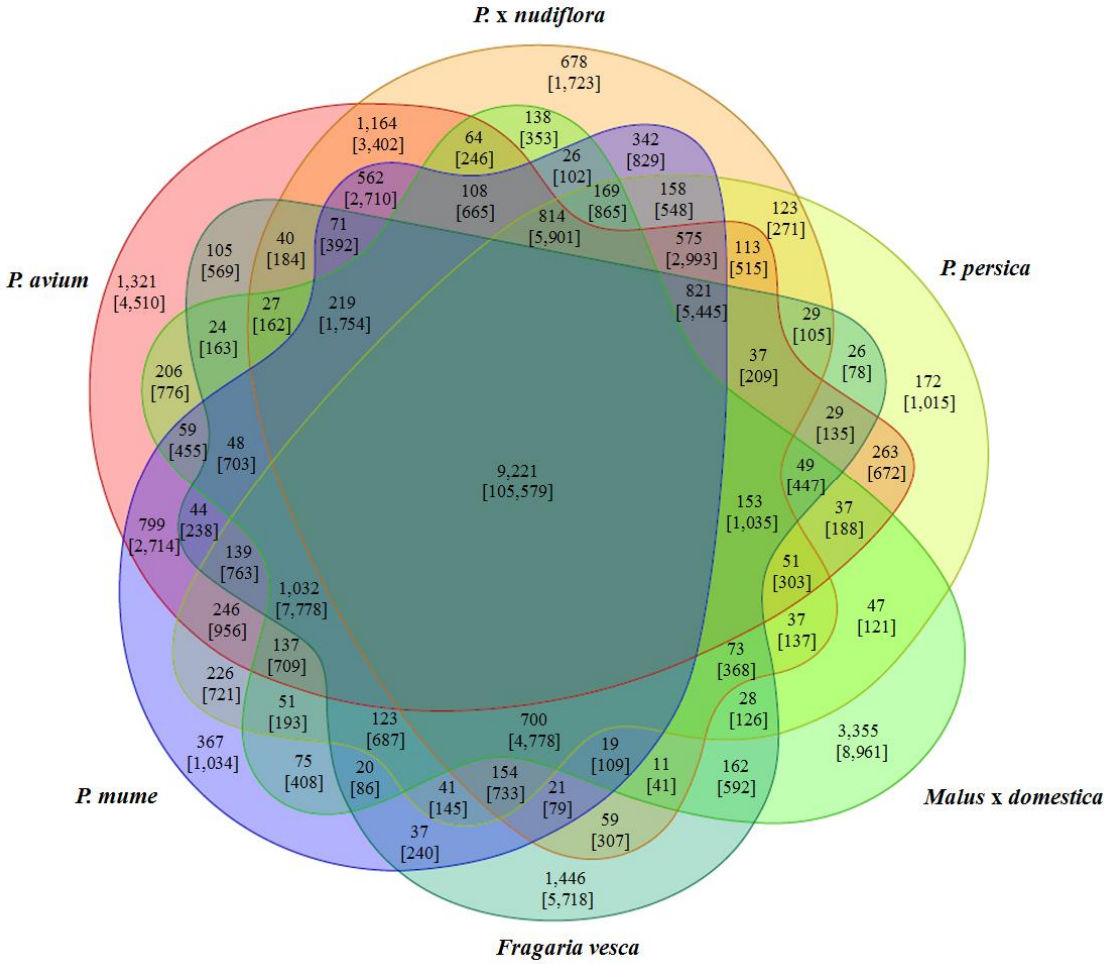


Figure 7. Venn diagram showing the unique and shared gene families between six sequenced genomes of the Rosaceae family.
The number of gene families and genes (in bracket) for each group are shown.

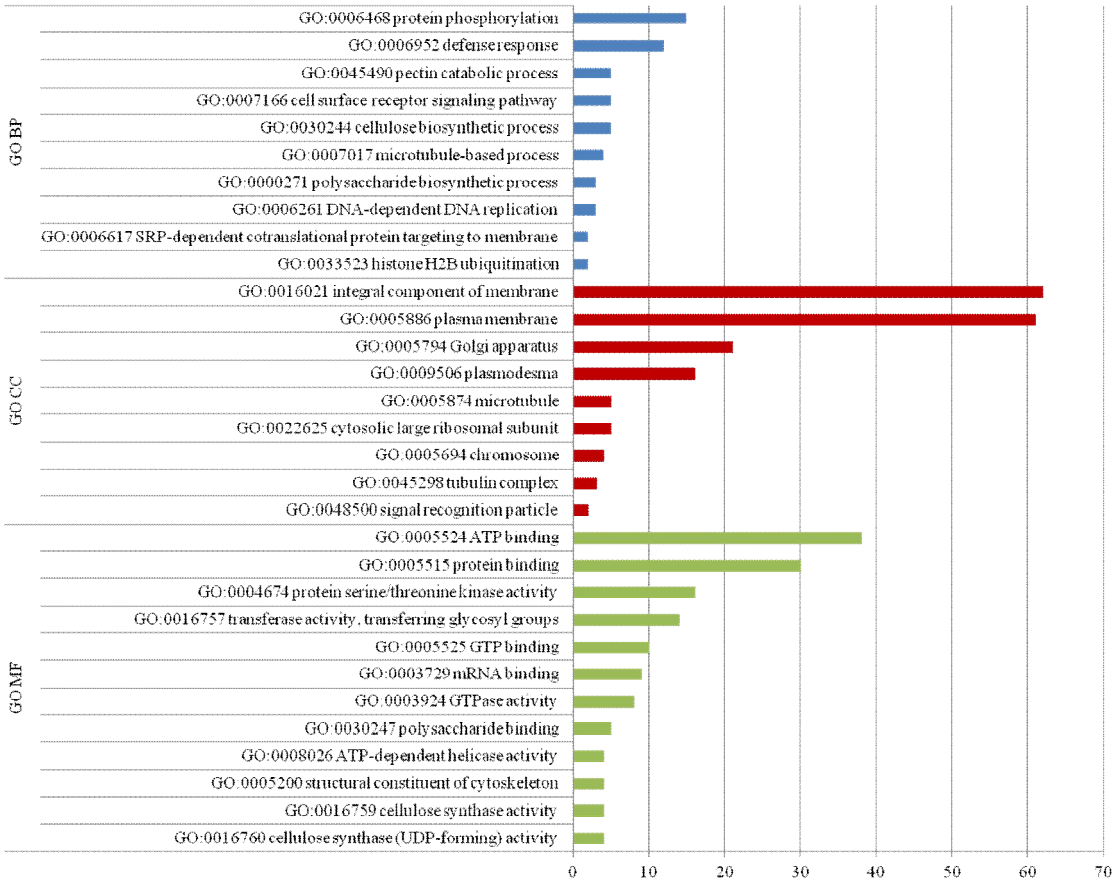


Figure 8. Histogram showing the enriched GO functional category of unique genes in the wild *P.* x *nudiflora* genome.
Colored bars represent functional categories of Biological Process (BP, blue), Cellular Component (CC, red), and Molecular Function (MF, green).

2. 장미과 식물종과 자생 왕벚나무에서 유전자 패밀리 비교 분석

클러스터 방법으로 Tribe-MCL 알고리즘을 적용한 PLAZA 데이터베이스의 특징은 Pfam 데이터베이스와 Prosite 데이터베이스에서 제공하는 단백질 도메인 분석 방법으로 커버되지 않는 유전자 패밀리를 그룹화 할 수 있고 식물 종간에서 계통분석과 진화적 유전자 기능 분석이 가능한 유전자 패밀리 정보를 제공한다(Martinez, 2011). 이에 따라, PLAZA Dicot v3.0 데이터베이스에서 제공하는 복숭아 유전자 패밀리 그룹 정보와 유전자 개수를 수집하였고, PLAZA Dicot v3.0에서 정의한 유전자 패밀리 서열을 대상으로 BLASTP 프로그램을 이용하여 각 유전자 패밀리에 검색 빈도가 높은 유전자를 해당 유전자 패밀리로 정의하였다. Z-test 통계 분석을 통해 자생 왕벚나무 대비 각 복숭아, 매실, 체리에서 유의 수준 0.001 이하의 총 62개 유전자 패밀리를 선별하였다(Table 14). 복숭아, 매실, 체리와 비교하여 자생 왕벚나무에서 유전자 개수가 많은 유전자 패밀리는 총 49개이며 자생 왕벚나무에서 유전자 개수가 적은 유전자 패밀리는 총 13개였다. 자생 왕벚나무에서 유전자를 많이 포함한 유전자 패밀리는 복숭아, 매실, 체리 보다 평균 3.8배 많았으며, 이들 중 특히 꽃의 조기 개화 시간을 조절하는 유전자인 Zinc finger (C2H2 type) transcription factor, Early Flowering 6 유전자 패밀리(Noh et al., 2004)와 꽃 기관 발달과 개화에 관련된 유전자가 포함된 homeobox-leucine zipper family 그룹(Chew et al., 2013)이 여기에 속하였다.

Table 14. Over- or under-represented gene families in the wild *P.* x *nudiflora* genome compared to the *P. avium*, *P. mume*, and *P. persica* genomes

Group ID		Gene family	Number of genes				*p* value		
PLAZA	InterPro		Pxn	Pa	Pm	Pp	Pxn-Pa	Pxn-Pm	Pxn-Pp
HOM03D000001	IPR002885	Pentatricopeptide repeat (PPR) superfamily protein	571	331	313	298	0.000	0.000	0.000
HOM03D000002	IPR000504	NAC transcription factor	438	253	244	211	0.000	0.000	0.000
HOM03D000119	IPR003311	AUXIN RESPONSE FACTOR	58	30	20	18	0.001	0.000	0.000
HOM03D000113	IPR000330	SNF2/Brahma-type chromatin-remodeling protein	53	21	24	23	0.000	0.001	0.001
HOM03D000194	IPR001356	Homeodomain-containing transcription factor, FLOWERING WAGENINGEN	44	22	10	11	0.002	0.000	0.000
HOM03D000428	IPR008928	Plant neutral invertase family protein	42	17	13	16	0.000	0.000	0.000
HOM03D000357	IPR001965	Zinc finger (C2H2 type) transcription factor, Early Flowering 6	31	8	9	8	0.000	0.000	0.000
HOM03D001412	IPR001356	Homeodomain-like transcriptional regulator	27	13	6	3	0.009	0.000	0.000
HOM03D000507	IPR001623	Chaperone DnaJ-domain superfamily protein	21	9	7	7	0.004	0.000	0.001
HOM03D001304	IPR000897	P-loop containing nucleoside triphosphate hydrolases superfamily protein	21	5	4	2	0.000	0.000	0.000
HOM03D003274	IPR024946	DNA GYRASE A	17	5	3	2	0.000	0.000	0.000
HOM03D000610	IPR005690	Outer membrane GTPase protein	16	5	5	4	0.000	0.001	0.000
HOM03D001915	IPR016102	Succinyl-CoA ligase, alpha subunit	16	5	3	5	0.000	0.000	0.001
HOM03D001724	IPR027705	SPFH/Band 7/PHB domain-containing membrane-associated protein family	15	5	5	4	0.001	0.003	0.000
HOM03D000677	IPR014019	Actin-binding FH2 (Formin Homology) protein	13	4	4	4	0.000	0.002	0.003
HOM03D002238	IPR001025	BAH domain	13	2	2	2	0.000	0.000	0.000
HOM03D003041	IPR009067	Histone acetyltransferase of the TAFII250 family	13	2	3	1	0.000	0.000	0.000
HOM03D000820	IPR013083	RING/FYVE/PHD zinc finger superfamily protein	11	4	3	3	0.008	0.001	0.002
HOM03D001331	IPR005516	Remorin family protein	11	4	3	3	0.008	0.001	0.002
HOM03D002176	IPR012935	C3HC zinc finger-like	10	1	2	2	0.000	0.000	0.000
HOM03D001856	IPR007461	RING/FYVE/PHD-type zinc finger family protein	9	3	2	2	0.007	0.000	0.000
HOM03D002014	IPR000225	Armadillo/beta-catenin-like repeat	9	2	2	2	0.000	0.000	0.000
HOM03D003353	IPR019137	Transcription activators	9	3	1	1	0.007	0.000	0.000

HOM03D005496	IPR007185	DNA polymerase epsilon subunit B2	9	1	1	2	0.000	0.000	0.000
HOM03D006297	IPR002058	Nucleotidyltransferase family protein	9	1	1	1	0.000	0.000	0.000
HOM03D000686	IPR004274	C-terminal domain phosphatase-like 3	8	2	2	2	0.001	0.002	0.003
HOM03D001826	IPR003114	Phox-associated domain	8	2	2	2	0.001	0.002	0.003
HOM03D002009	IPR019448	Myosin heavy chain-related protein	8	2	2	2	0.001	0.002	0.003
HOM03D002626	IPR001739	Methyl-CPG-binding domain 8	8	2	2	2	0.001	0.002	0.003
HOM03D002991	IPR015943	BTB/POZ domain with WD40/YVTN repeat-like protein	8	2	2	2	0.001	0.002	0.003
HOM03D003996	IPR012953	Transducin/WD40 repeat-like superfamily protein	8	2	2	2	0.001	0.002	0.003
HOM03D004469	IPR011989	SH3 domain-containing protein	8	1	2	1	0.000	0.002	0.000
HOM03D004516	IPR003594	Zinc finger, C3HC4 type (RING finger) family protein	8	1	1	1	0.000	0.000	0.000
HOM03D004833	IPR012340	Organellar single-stranded DNA binding protein	8	2	1	1	0.001	0.000	0.000
HOM03D004379	IPR013197	RNA polymerase III subunit RPC82 family protein	7	1	1	1	0.000	0.000	0.000
HOM03D005211	IPR003675	Alpha/beta-Hydrolases superfamily protein	7	1	1	1	0.000	0.000	0.000
HOM03D003069	IPR005314	Homolog of separase	6	1	1	1	0.000	0.000	0.000
HOM03D007410	IPR003690	Plastid transcriptionally active 15	6	1	1	1	0.000	0.000	0.000
HOM03D001104	IPR003333	Cyclopropane-fatty-acyl-phospholipid synthase	5	1	1	1	0.001	0.003	0.003
HOM03D002754	IPR021133	CLIP-associated protein	5	1	1	1	0.001	0.003	0.003
HOM03D003144	IPR007452	Embryo defective 2410	5	1	1	1	0.001	0.003	0.003
HOM03D003169	IPR001394	Ubiquitin carboxyl-terminal hydrolase-related protein	5	1	1	1	0.001	0.003	0.003
HOM03D003487	IPR004516	Class II aaRS and biotin synthetases superfamily protein	5	1	1	1	0.001	0.003	0.003
HOM03D003514	IPR000195	Ypt/Rab-GAP domain of gyp1p superfamily protein	5	1	1	1	0.001	0.003	0.003
HOM03D003934	IPR001503	Fucosyltransferase 13	5	1	1	1	0.001	0.003	0.003
HOM03D004324	IPR018027	GLU-ADT subunit B	5	1	1	1	0.001	0.003	0.003
HOM03D004420	IPR013917	flavodoxin family protein / radical SAM domain-containing protein	5	1	1	1	0.001	0.003	0.003
HOM03D005672	IPR005496	Integral membrane TerC family protein	5	1	1	1	0.001	0.003	0.003
HOM03D006176	IPR020046	5′-3′ exonuclease family protein	5	1	1	1	0.001	0.003	0.003
HOM03D000005	IPR000157	WRKY transcription factor	112	310	287	209	0.000	0.000	0.000
HOM03D000010	IPR027417	NB-ARC domain-containing disease resistance protein	81	242	149	123	0.000	0.000	0.000
HOM03D000144	IPR001757	ATPase E1-E2 type family protein	45	61	100	21	0.000	0.000	0.009
HOM03D000112	IPR002110	Ankyrin repeat family protein	37	75	73	75	0.000	0.000	0.000
HOM03D000085	IPR001509	NAD(P)-binding Rossmann-fold superfamily protein	35	45	41	39	0.004	0.002	0.004
HOM03D000103	IPR013126	Heat shock protein 70 family protein	27	53	41	35	0.000	0.000	0.001
HOM03D000180	IPR006580	General transcription factor 2-related zinc finger protein	16	184	23	67	0.000	0.004	0.000

HOM03D000222	IPR000008	C2 calcium/lipid-binding plant phosphoribosyltransferase family protein	9	19	17	16	0.002	0.002	0.003
HOM03D000077	IPR001611	F-box/RNI-like/FBD-like domains-containing protein	8	17	14	18	0.004	0.008	0.000
HOM03D000371	IPR001611	ADR1 family NB-LRR immune receptors	6	19	17	19	0.000	0.000	0.000
HOM03D000278	IPR003480	HXXXD-type acyl-transferase family protein	3	21	16	16	0.000	0.000	0.000
HOM03D000285	IPR025558	Zinc ion binding	2	32	9	23	0.000	0.003	0.000
HOM03D000434	IPR001764	Glycosyl hydrolase family protein	1	10	12	15	0.001	0.000	0.000

3. 장미과에서 자생 왕벚나무의 진화적 분기 시기 추정

자생 왕벚나무의 진화적 분기 시기를 추정하기 위해 양지꽃족(*tribe Potentilleae*) 딸기와 사과나무족(*tribe Maleae*) 사과, 복숭아아속(subgenus *Amygdalus*) 복숭아, 벚나무속(*Prunus*) 매실, 유전자를 이용하여 장미과에서 진화적 유연관계를 분석하였다. 분기 분석을 위해 복숭아, 매실, 체리, 딸기, 사과별로 자생 왕벚나무와 서열 상동성 있는 유전자를 선별하였고 유전자 코딩 영역에서 동의치환을 분석하여 synonymous substitutions per synonymous site (Ks) 값을 결정하고 분기를 추정하였다(Figure 9).

자생 왕벚나무와 딸기에서 Ks 분포에서 정점에 있는 Ks 값은 0.48, 자생 왕벚나무와 사과에서 Ks 분포에서 정점에 있는 Ks 값은 0.35로 양지꽃족과 사과나무족은 백악기(Cretaceous) 효신세(Paleocene) 약 88~61백만 년 전에 분기된 것으로 추정되었다(Chin et al., 2014). 따라서 자생 왕벚나무와 복숭아에서 Ks 분포에서 정점에 있는 Ks 값은 0.05, 자생 왕벚나무와 매실에서 Ks 분포에서 정점에 있는 Ks 값 또한 0.05로 복숭아와 매실의 Ks 값은 동일하였다. 벚나무속 유전체간에서 Ks 분포는 벚나무속 종이 최근에 분기된 것으로 사료된다. 자생 왕벚나무에서 paralogs 유전자들의 Ks 분포에서 정점에 있는 Ks 값은 0.02, 자생 왕벚나무와 체리에서 Ks 분포에서 정점에 있는 Ks 값은 0.01로 이들 분류군이 연속적으로 분기된 것으로 나타났다.

콩과 모델 식물 *Medicago truncatula*를 외집단 분류군으로 포함하여 총 7종의 오솔로그 유전자를 비교 분석하여 276개 single copy gene을 선별하였다. 이 유전자들은 다중 정렬을 통해 376,758 bp 길이로 정렬되었고, 이중 33.7% 해당하는 127,606개가 variable sites이며 7.7%에 해당하는 29,134개가 Parsimony-informative site로 나타났다. BEAST 프로그램을 이용하여 장미과에 대한 분자 연대측정(Molecular dating)을 수행하였다(Figure 10). 장미과내에서 벚나무속 자생 왕벚나무, 복숭아, 매실, 체리의 분기 연대는 약 66.2백만 년 전으로 추정되었으며, 64.2~67.4백만 년 전 구간의 higher posterior densities (HPD)는 95%에 해당했다. 벚나무아속(subgenus *Cerasus*)과 복숭아아속의 분기된 연대는 44.0백만 년 전으로 추정되었으며, 42.2~45.6백만 년 전 구간의 HPD는 95%에 해당했다. 벚나무아속의 자생 왕벚나무와 체리의 연대는 35.9백만 년 전으로 추정되었으며, 34.4~37.3백만 년 전 구간의 HPD는 95%에 해당했다. 이상의 결과는 동아시아에서 벚나무속이 35백만 년 전에 빠르게 분기가 일어난 것으로 보고된 결과와 일치하였다(Chin et al., 2014).

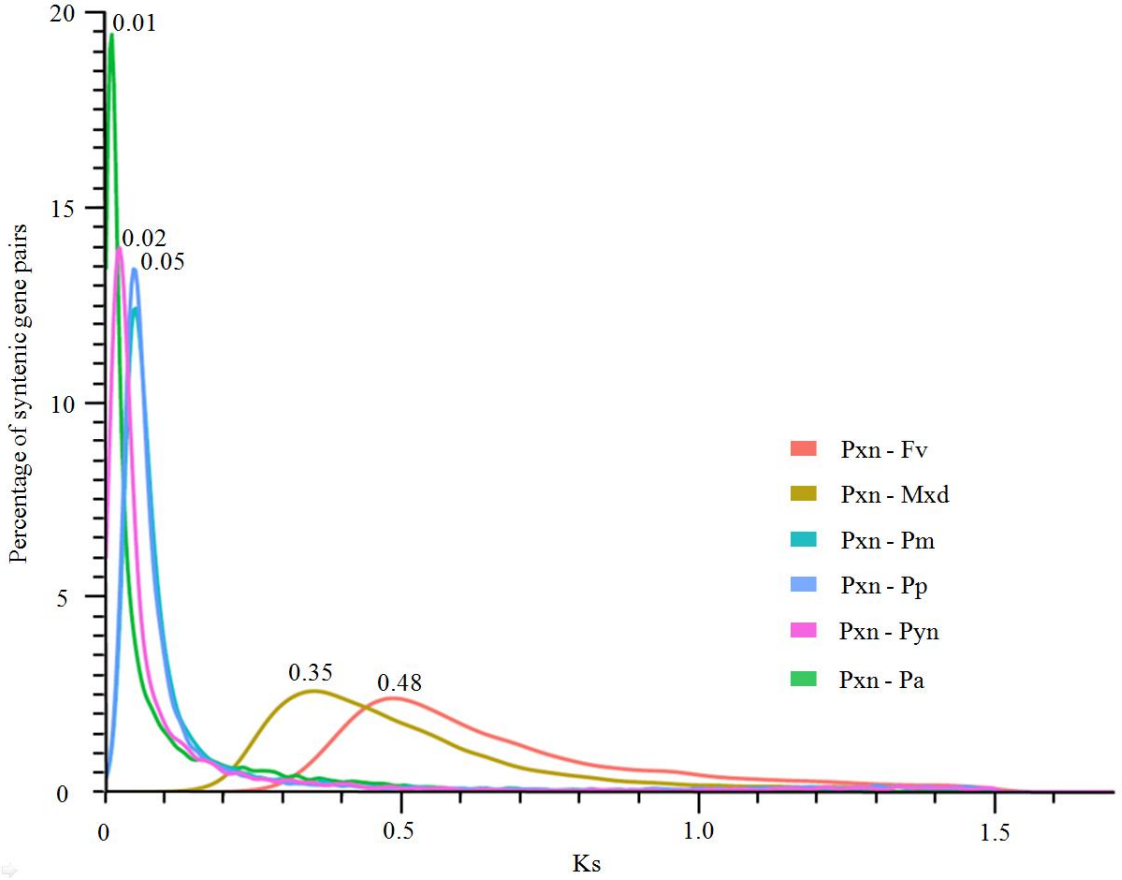


Figure 9. Distribution of Ks in between Rosaceae species.
Distribution of Ks values obtained from comparisons of orthologous gene sets between six genomes of the Rosaceae family and paralogous gene sets in wild *P.* x *nudiflora*. Pxn *P.* x *nudiflora*, Fv *Fragaria vesca*, Mxd *Malus* x *domestica*, Pa *P. avium*, Pm *P. mume*, Pp *P. persica*.

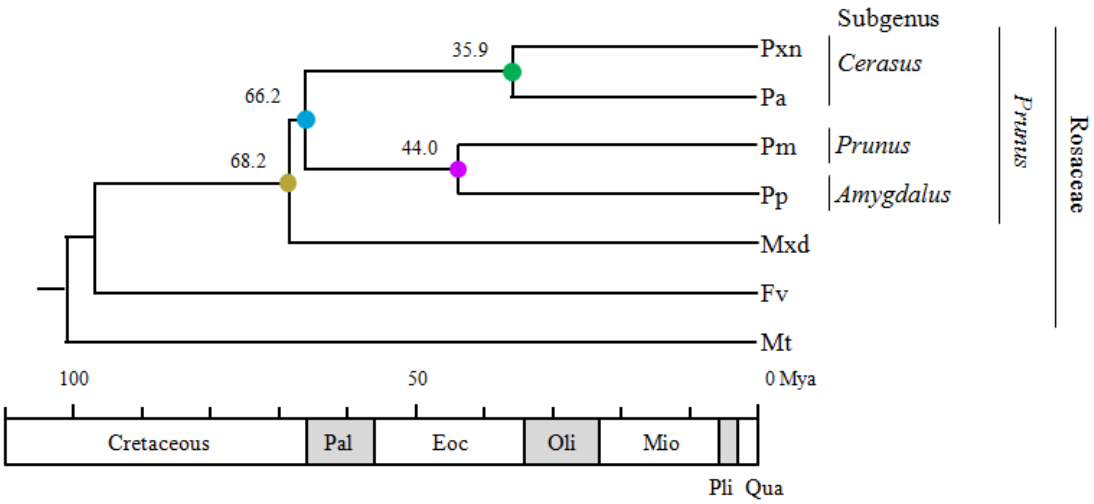


Figure 10. Genome evolution of *Prunus* species.
Estimation of dates for speciation events are given in millions of years based on Bayesian evolutionary analysis of 276 conserved single copy genes. Pxn *P.* x *nudiflora*, Fv *Fragaria vesca*, Mt *Medicago truncatula*, Mxd *Malus* x *domestica*, Pa *P. avium*, Pm *P. mume*, Pp *P. persica*. Pal Paleocene, Eoc Eocene, Oli Oligocene, Mio Miocene, Pli Pliocene, Qua Quaternary.

제 6 절 왕벚나무 유전체의 부계와 모계 기원 분석

1. 왕벚나무 유전체의 부계와 모계 유래 유전자 선별

FALCON 어셈블러는 PacBio subread 서열을 OLC 방법으로 연결하는 과정에서 structure variation (SV)이 발생한 조립 서열을 생성한다. 유전체 조립 서열은 모계 유전형과 부계 유전형이 섞인(haplotype-fused) 특성을 갖는다. 자생 왕벚나무는 올벚나무와 벚나무 사이에서 1세대 잡종으로 보고한 기존 연구결과(Cho et al., 2017)를 참고하여 올벚나무와 벚나무의 short read 서열을 왕벚나무 유전체에 각각 매핑하고 유전자 영역에서 부계 유전자형 빈도가 높은 서열을 포함한 유전자를 부계 유래, 모계 유전자형 빈도가 높은 서열을 포함한 유전자를 모계 유래, 부계와 모계 유전자형이 비슷한 수준으로 존재하는 경우 공통 유전자로 분류하는 서열 phasing을 실시하였다. 또한 short read 서열 매핑 커버리지를 계산하여 올벚나무 서열만 커버하는 유전자를 유일한 모계 유래 유전자로 정의하였고, 벚나무 서열만 커버하는 유전자를 유일한 부계 유래 유전자로 정의하였다(Figure 11). 예를 들어 Pxn_C0035.77 유전자는 서열 매핑 결과, 올벚나무 유전형과 벚나무 유전형을 모두 갖는 short read 서열이 매핑 되었으나 벚나무 서열간에 SNP가 대부분 나타나 유전자 영역의 서열은 올벚나무 유전자형이 반영된 모계 유래 유전자로 판정하였다. Pxn_C114.21 유전자는 서열 매핑 결과에서 자생 왕벚나무와 올벚나무 short read 서열이 매핑 되었고 벚나무 short read 서열은 매핑 되지 않아 올벚나무 서열로만 조립된 모계 유래 유전자로 판정하였다. 한편, Pxn_C1298.72 유전자는 자생 왕벚나무와 벚나무, 올벚나무 short read가 모두 매핑 되었으나 올벚나무 서열 간에 SNP가 대부분 나타나 부계 유래 유전자로 판정하였다. Pxn_C1396.58 유전자는 매핑 결과에서 자생 왕벚나무와 벚나무 short read 서열만 매핑 되어 유일한 부계 유래 유전자로 판정하였다. Pxn_C2023.7 유전자는 올벚나무, 벚나무, 자생 왕벚나무 매핑 결과에서 부계 모계 유래 유전자형이 모두 관찰되고 벚나무와 올벚나무 서열에서 모두 SNP가 나타나 양친에서 함께 유래한 유전자로 판정하였다.

이상과 같이 short read 서열의 매핑 커버리지와 부계와 모계 유전자형을 분석한 결과, 모계 유래 유전자 총 8,030개, 부계 유래 유전자 총 8,809개 유전자를 선별하였다 (Table 15).

따라서 자생 왕벚나무 유전체는 양친 모두에서 유래한 유전자는 59.2%, 모계 유래 유전자 19.4%, 부계 유래 유전자 21.3%로 구성되어 있고, 부계 유전자 개수와 모계 유래 유전자 개수가 유사하여 벚나무와 올벚나무의 1세대 잡종임이 밝혀졌다.

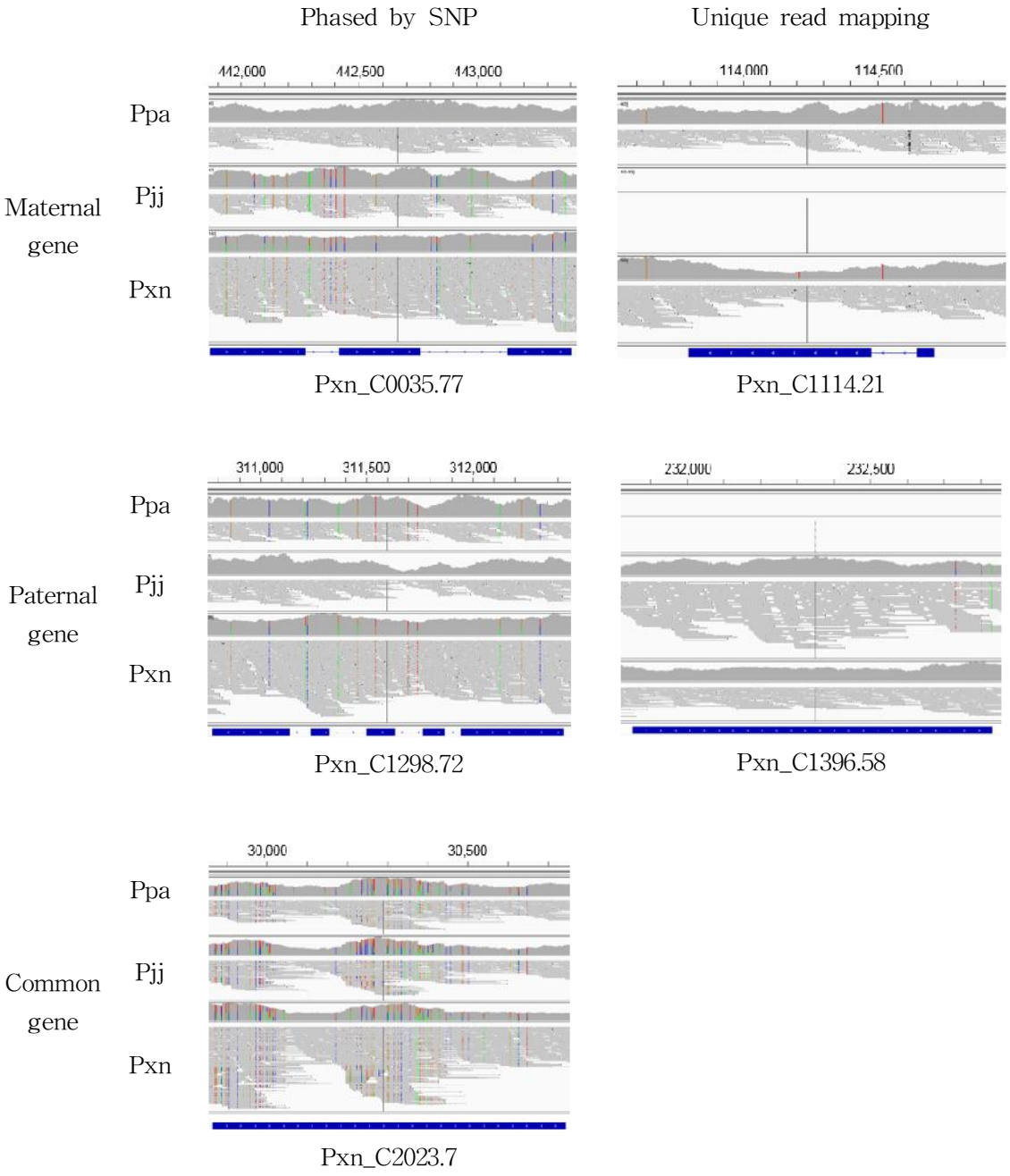


Figure 11. Examples of haplotype-phased gene models.
Gene models predicted from the initial “haplotype-fused” assembly are phased according to read mapping and SNP analysis using the Illumina short-read sequences of putative parental species. Genes were phased into one parental haplotype if a gene was aligned only by reads from one parental species (unique mapping) or had at least two fold as many supports for SNPs by reads of one parental species (phased by SNP). Genes with similar supports of read mapping for both parental species are defined as common type. Colored dots denote SNPs identified in the aligned reads.

Table 15. Classification of wild *P.* x *nudiflora* genes based on sequence phasing of the draft assembly by mapping of Illumina short-read sequences from putative parental species, maternal *P. pendula* f. *ascendens* and paternal *P. jamasakura*

Type	Maternal gene		Paternal gene		Common gene
	Unique	Phased	Unique	Phased	
Number	548	7,482	1,353	7,456	24,455
Ratio (%)	1.3	18.1	3.3	18.1	59.2

2. 장미과 유전체와 자생 왕벚나무 유전체 서열 비교 분석을 통한 어셈블리의 염색체별 정렬

장미과 복숭아, 체리, 매실 유전체는 각각 8개 염색체로 구성되어 있다. 자생 왕벚나무 예측 유전자 모델이 각 장미과 유전체의 상동 염색체상에서 synteny 영역과 서열 일치성을 비교 분석하기 위해 자생 왕벚나무 스케폴드 조립 서열을 각 염색체별 pseudomolecule 형식으로 재배열을 하였다. 8개 염색체별로 스케폴드 조립 서열을 재배열하기 위해 염색체 수준의 조립서열이 보고되어 있는 복숭아 유전체를 기준으로 설정하였다. 자생 왕벚나무 유전자 모델의 단백질 서열을 BLASTP 프로그램을 이용하여 상동성 분석을 수행하고, MCScanX 프로그램을 이용하여 유전자 좌표 정보와 BLASTP 분석 결과에서 synteny 영역을 분석하였다. 또한 NUCmer 프로그램을 이용하여 자생 왕벚나무 스케폴드 조립 서열 중 복숭아 유전체 서열에 일치하는 서열을 선별하였다.

복숭아 유전체에 대한 유전자 synteny 영역과 서열 일치성 비교 분석 결과를 바탕으로 왕벚나무 전체 조립 서열 길이에 대비 87.0%에 해당하는 총 281.6 Mb 서열을 각 염색별로 재배열하였다. 재배열 스케폴드 서열 개수는 총 2,462개이며, 부계 유래 유전자 총 8,148개, 모계 유래 유전자 총 7,549개, 공통 유전자 총 21,106개로 전체 유전자 모델 개수에 89.1%에 해당한다(Table 16). 각 염색체별로 유전자 위치를 부계 유래 유전자는 파란색, 모계 유래 유전자는 적색, 공통 유전자는 회색으로 시각화한 결과 복잡한 모자이크 패턴을 보이며, 부계와 모계로부터 무작위 배열된 잡종 유전체 특성을 보였다(Figure 12).

복숭아, 체리, 매실 각각의 유전체에 자생 왕벚나무 재배열 유전체 서열을 NUCmer 프로그램을 이용하여 서열 일치성과 MCScanX 프로그램을 이용하여 유전자 synteny 영역을 비교 분석하였다. 각 염색체 별로 유전자 synteny 위치를 2D Dot plot 으로 시각화한 결과, 복숭아 유전체에 서열 정렬 54.5%에 해당하는 synteny 유전자는 총 15,077개(Figure 13; Table 17), 체리 유전체에 서열 정렬 66.4%에 해당하는 synteny 유전자는 총 14,031개(Figure 14; Table 17), 매실 유전체에 서열 정렬 56.1%에 해당하는 synteny 유전자는 총 14,061개로 모두 동일 선형 형태를 보였다(Figure 15; Table 17). 매실 유전체의 경우 복숭아와 체리 유전체 보다 선형성이 낮았으며, 특히 Pm2와 Pm4에서 inversion이 관찰되었다.

Table 16. Summary statistics of phased-genes and number of scaffolds in 8 chromosome

Chr	Total bases	Scaffold count	Paternal	Maternal	Common	Total gene
Pxn1	61,276,499	501	1,803	1,659	4,785	8,247
Pxn2	33,419,000	291	944	828	2,406	4,178
Pxn3	29,201,006	246	935	864	2,236	4,035
Pxn4	38,749,813	362	925	862	2,664	4,451
Pxn5	25,844,825	224	778	799	2,115	3,692
Pxn6	36,107,850	311	1,129	1,049	2,734	4,912
Pxn7	29,834,691	262	875	818	2,049	3,742
Pxn8	27,160,671	265	759	670	2,117	3,546
Total	281,594,355	2,462	8,148	7,549	21,106	36,803

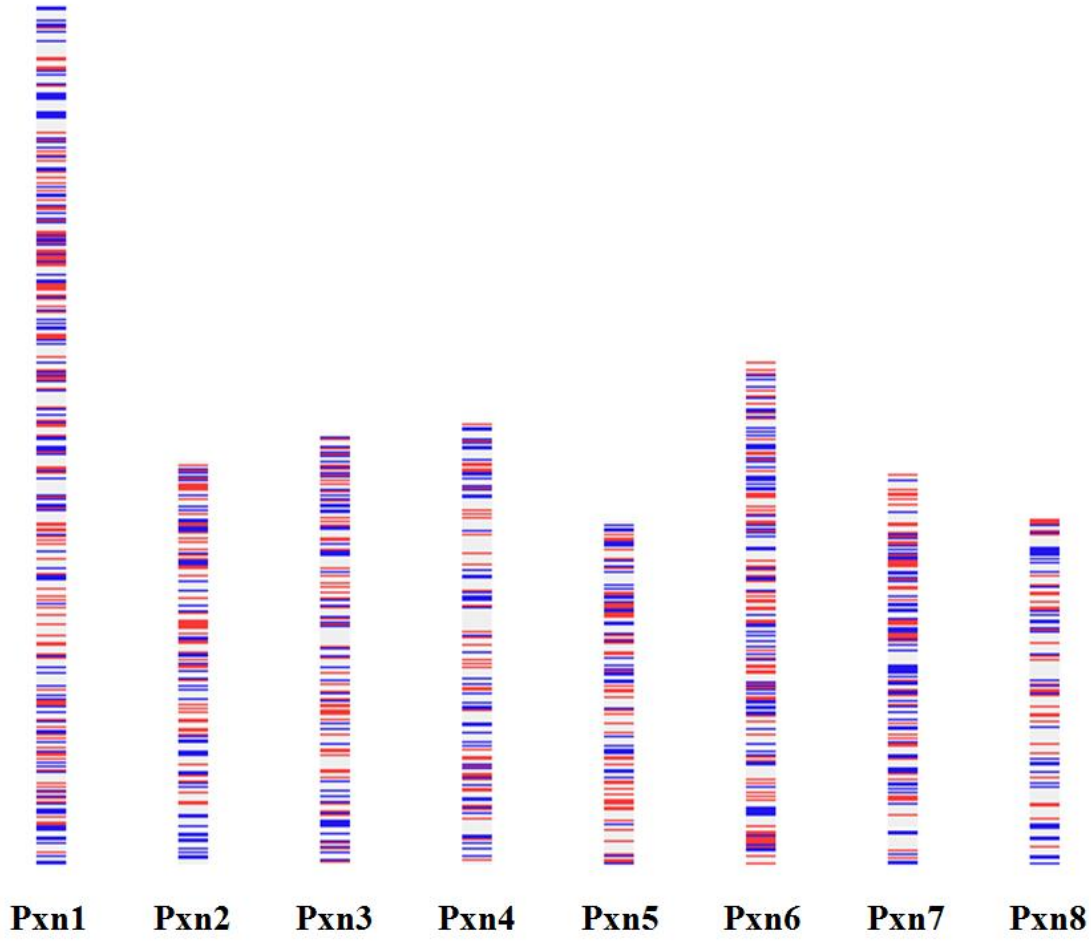


Figure 12. Distribution of haplotype-phased genes in the tentative chromosomes of wild *P.* x *nudiflora*.
Colored lines represent maternal-phased genes (red), paternal-phased genes (blue), or common genes (gray).

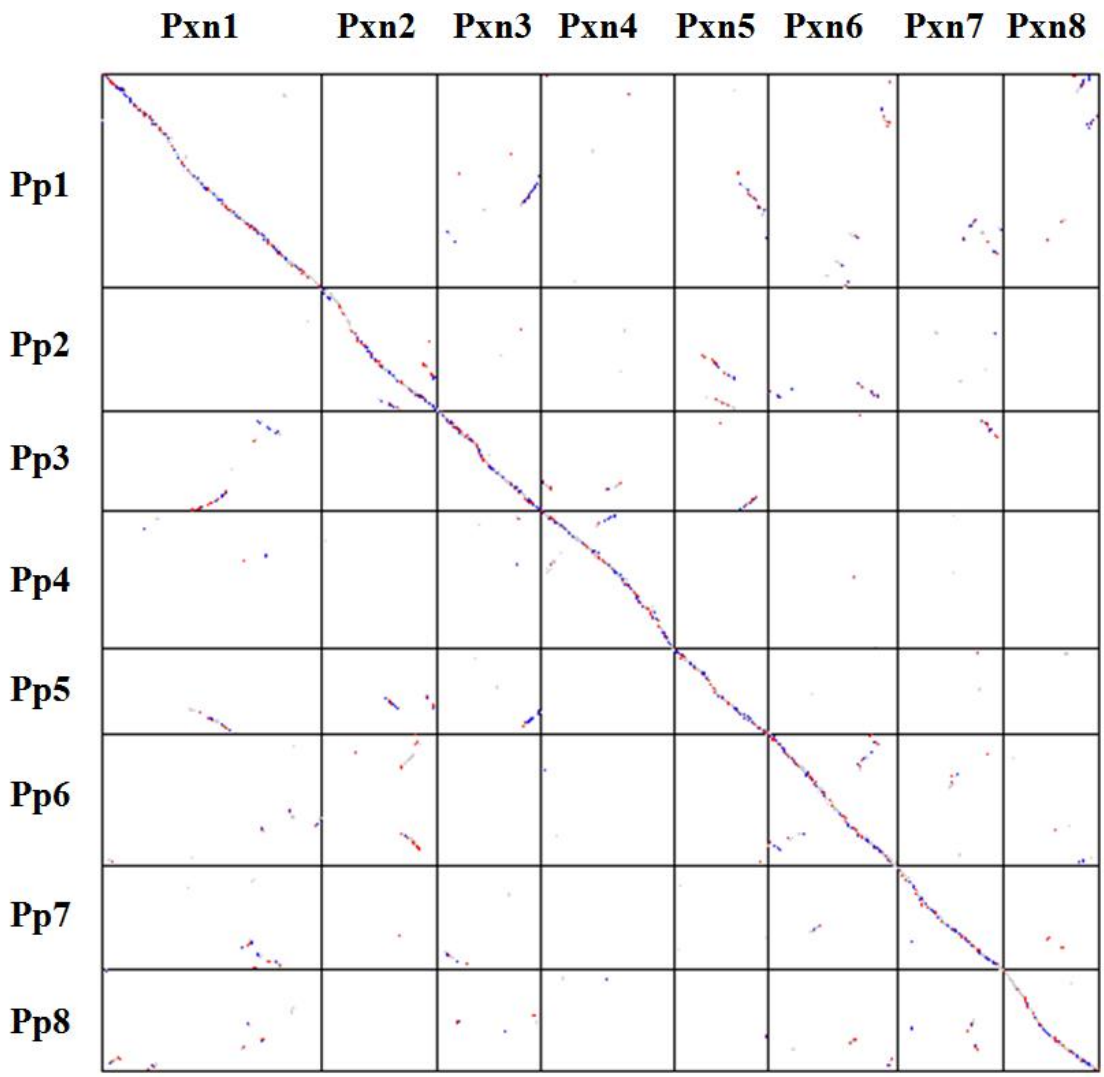


Figure 13. Chromosomal arrangement of the gene-phased genome assembly of wild *P.* x *nudiflora* (Pxn) onto the *P. persica* (Pp) genome.
Colored dots represent maternal-phased genes (red), paternal-phased genes (blue), or common genes (gray).

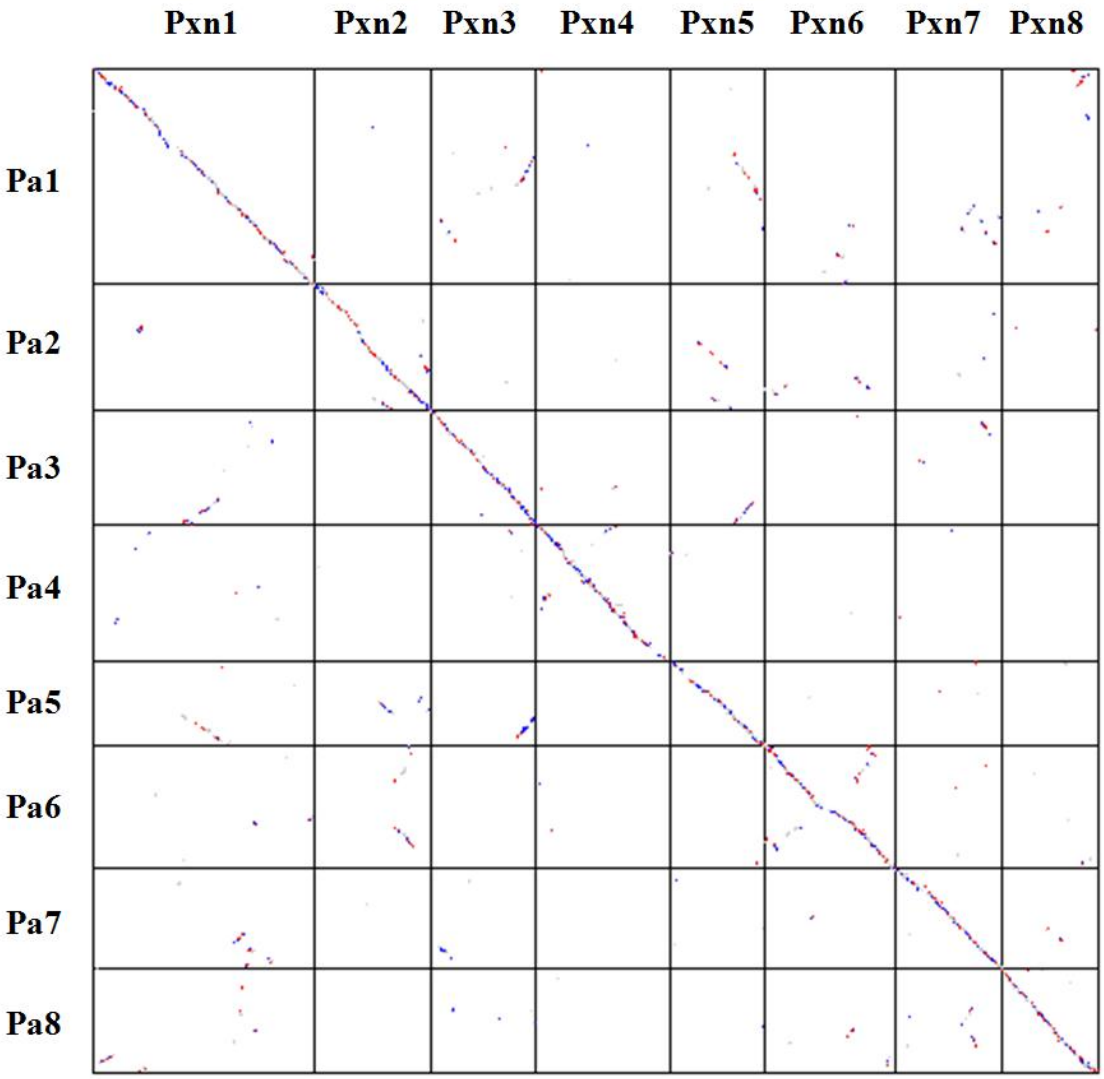


Figure 14. Chromosomal arrangement of the gene-phased genome assembly of wild *P.* x *nudiflora* (Pxn) onto the *P. avium* (Pa) genome.
Colored dots represent maternal-phased genes (red), paternal-phased genes (blue), or common genes (gray).

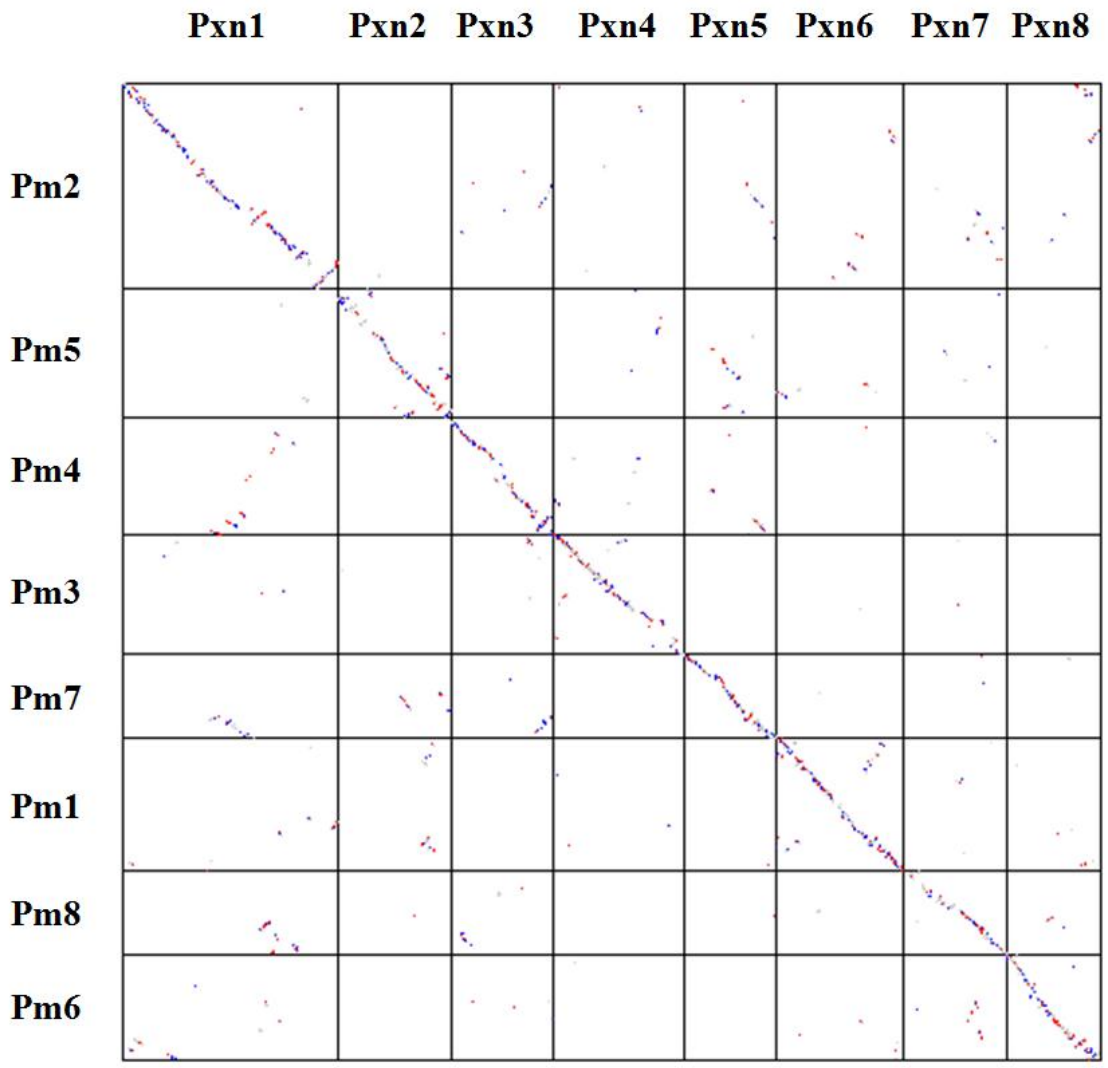


Figure 15. Chromosomal arrangement of the gene-phased genome assembly of wild *P.* x *nudiflora* (Pxn) onto the *P. mume* (Pm) genome.
Colored dots represent maternal-phased genes (red), paternal-phased genes (blue), or common genes (gray).

Table 17. Coverage of individual chromosomes of peach (Pp), sweet cherry (Pa), and Chinese plum (Pm) showing synteny with the counterpart of wild *P.* x *nudiflora* (Pxn) genome

Species	Chr.	Total length (bp)	Sequences syntenic with Pxn counterparts (bp)	Percent coverage	No. syntenic genes
P. persica	Pp1	47,851,208	28,130,621	58.8%	3,427
	Pp2	30,405,870	15,111,843	49.7%	1,682
	Pp3	27,368,013	13,428,324	49.1%	1,636
	Pp4	25,843,236	13,959,725	54.0%	1,648
	Pp5	18,496,696	10,861,775	58.7%	1,496
	Pp6	30,767,194	17,074,354	55.5%	2,114
	Pp7	22,388,614	12,810,257	57.2%	1,634
	Pp8	22,573,980	11,550,115	51.2%	1,440
	Sum	225,694,811	122,927,014	54.5%	15,077
P. avium	Pa1	43,232,855	30,582,266	70.7%	3,160
	Pa2	25,254,475	16,188,987	64.1%	1,624
	Pa3	22,613,589	14,628,953	64.7%	1,571
	Pa4	27,279,932	16,819,579	61.7%	1,547
	Pa5	17,020,956	11,832,132	69.5%	1,392
	Pa6	24,611,171	16,450,029	66.8%	1,885
	Pa7	19,892,082	13,645,244	68.6%	1,486
	Pa8	20,769,356	13,096,358	63.1%	1,366
	Sum	200,674,416	133,243,548	66.4%	14,031
P. mume	Pm1	26,753,124	15,321,425	57.3%	1,941
	Pm2	42,086,871	25,110,563	59.7%	3,106
	Pm3	24,358,621	13,482,748	55.4%	1,538
	Pm4	23,936,025	12,399,586	51.8%	1,603
	Pm5	26,141,170	13,766,270	52.7%	1,670
	Pm6	21,292,300	11,172,574	52.5%	1,442
	Pm7	17,044,613	9,715,088	57.0%	1,355
	Pm8	17,249,491	10,531,026	61.1%	1,406
	Sum	198,862,215	111,499,280	56.1%	14,061

3. 왕벚나무의 부계 및 모계 특이적 유전자 발현 분석

근연종 벚나무류의 short read 서열 매핑을 통해 자생 왕벚나무는 올벚나무와 벚나무에서 교잡된 잡종 식물임을 확인하였으며, 매핑서열의 SNP 분석을 이용하여 부계 유래 유전자와 모계 유래 유전자를 구분하였다. 이를 바탕으로 자생 왕벚나무 유전자 중 각 조직에서 차등 발현하는 유전자를 부계 또는 모계 기원으로 나누어 잡종 강세 또는 잡종 약세 특성을 분석하였다.

잎, 꽃잎, 암술, 수술, 열매 조직에서 3회 반복 생산한 mRNA-seq 서열을 유전자 모델에 매핑 하여 유전자 발현 값에 대한 재현성과 정확도를 높였다. 유전자 발현 값에 대해 각 조직별로 pearson correlation 이용한 상관분석을 수행하였고, 동일 조직 반복샘플 내에서 상관계수는 최소 0.92 이상으로 재현성이 높음을 확인하였다(Figure 16). 5개 조직을 모두 비교한 결과 차등 발현 유전자 총 2,565개를 선별하였다. Short read 서열의 SNP를 이용한 양친계통 판정 결과를 바탕으로 모계 유래 차등 발현 유전자 562개, 부계 차등 발현 유전자 576개 유전자를 선별하였다. 각 부계와 모계 유래 차등발현 유전자를 각각 K-mean 클러스터 분석을 수행하여 조직 특이적인 클러스터 그룹을 분석하였다(Figure 17). K-mean 클러스터는 분석 이전에 K 값을 정의해야 하므로 mclust 프로그램을 이용하여 최적에 K 값을 예측하였고 부계 K값은 7, 모계 K값은 10을 적용하였다. 부계와 모계 클러스터 그룹 중 5개 조직에서 공통 또는 유일한 클러스터 그룹을 비교하였다. 수술 조직에서 모계 클러스터 9번, 10번과 부계 클러스터 5번, 6번 그룹이 공통된 차등 발현 양상을 보였다. 잎 조직에서 모계 클러스터 2번, 5번과 부계 클러스터 7번 그룹이 공통된 차등 발현 패턴을 보였다. 열매 조직에서 모계 클러스터 7번과 부계 클러스터 4번 그룹이 공통된 차등 발현 패턴을 보였다. 특히, 모계 클러스터 3번은 암술 조직에서 조직 특이적인 차등 발현 양상을 보였다.

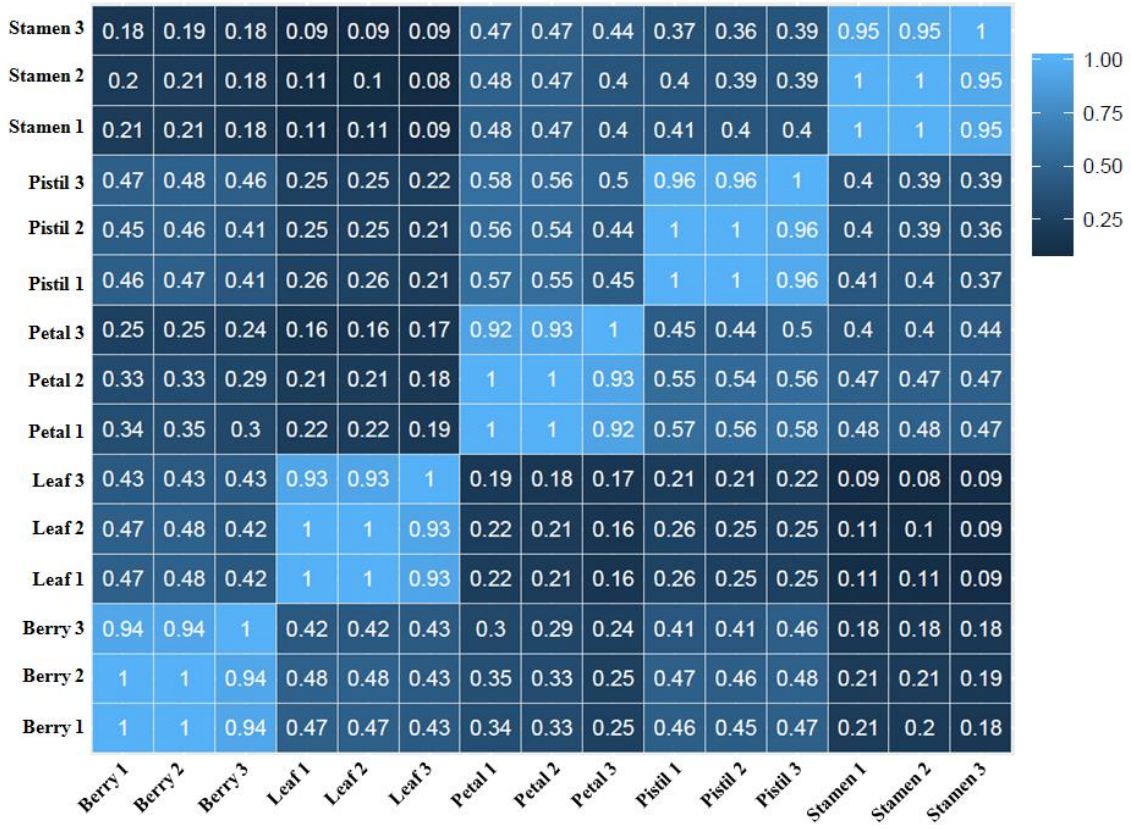


Figure 16. The correlation matrix of reproducibility for expression value by DESeq2 normalization.

The Closer to 1 value is the higher reproducibility level with gradation of blue color in matrix.

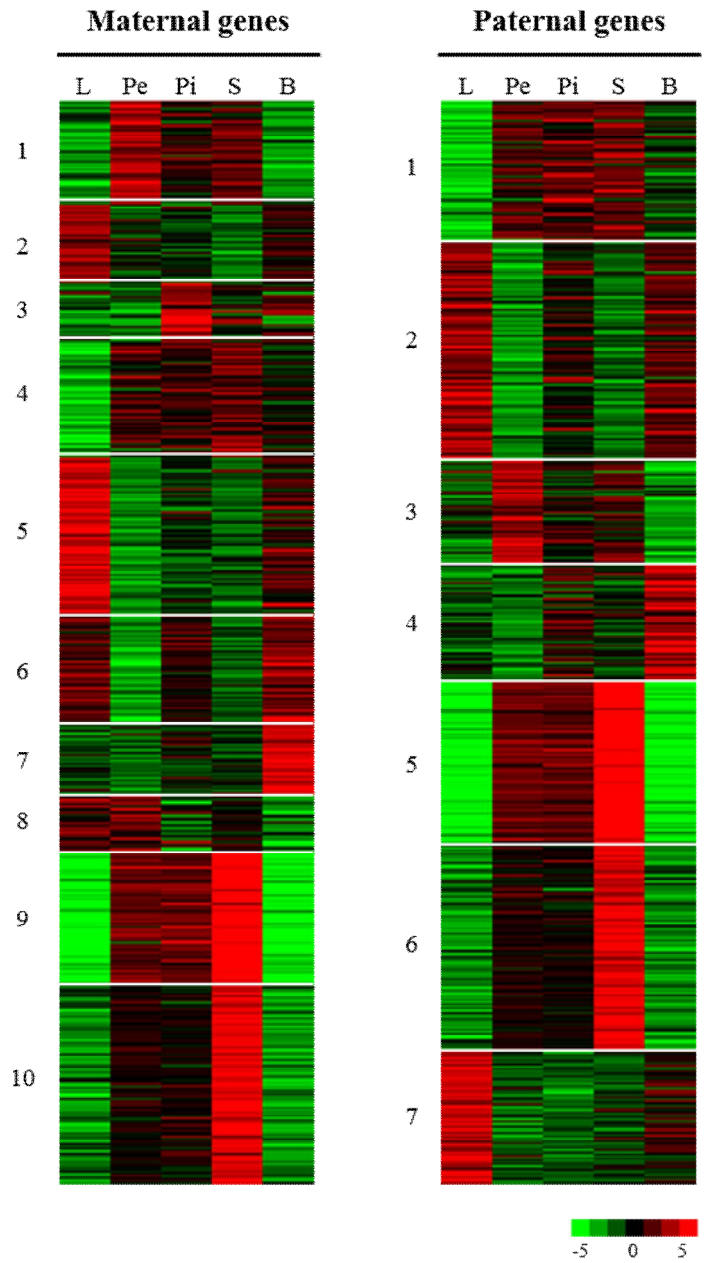


Figure 17. The clustering of differentially expressed genes (DEGs) in each paternal and materal.

Heat maps representing the expression of 562 maternal- and 576 paternal-phased genes, which are identified as differentially expressed genes in the mRNA-seq analysis, in different tissues. The normalized count values of a given gene from three independent biological replicates across all samples were used as a normalization factor. The vertical axes organize genes according to co-expression. The horizontal axes represent five tissues: leaf (L), petal (Pe), pistil (Pi), stamen (S), and berry (B).

모계 유래 차등 발현 유전자 562개와 부계 유래 차등 발현 유전자 576개의 클러스터 분석을 통해 발현 패턴이 구별되는 모계 10개 그룹, 부계 7개 유전자 그룹의 GO와 pathway에 따른 유전자 기능 범주를 분석하였다.

각 클러스터 그룹에서 차등 발현 유전자에 상동성 높은 애기장대 유전자 아이디를 DAVID 프로그램에 통해 유의수준 P-value 값 0.05 이하의 각 기능 범주에 해당하는 GO 아이디 및 기능명, KEGG 아이디 및 pathway 명, 해당 기능 범주에 속해 있는 차등 유전자를 추출하였다. 모계 10개 클러스터에서 GO Biological Process (BP) 총 23개, Cellular Component (CC) 총 20개, Molecular Function (MF) 총 13개, KEGG pathway 총 6개의 유의미한 유전자 기능 범주를 선별하였다(Table 18). 동일한 방법으로 부계 7개 클러스터에서 유의미한 BP 총 30개, CC 총 21개, MF 총 28개, KEGG pathway 총 6개를 선별하였다(Table 19).

5개 조직에서 공통 또는 특이적인 유의미한 유전자 기능 범주에 속한 모계와 부계 유래 유전자 발현 패턴을 비교 분석하였다(Figure 18). 모계 클러스터 3번 그룹의 발현 패턴은 부계 클러스터에는 없는 클러스터 그룹으로 다른 조직 보다 암술 조직에서 1.4~2.2배 조직 특이적 발현 패턴을 보였다. 모계 클러스터 3번 그룹에 차등 발현 유전자는 전사인자 기능을 갖고 있으며 GO enrichment 분석 결과 주로 flower development 기능 범주에 속하였다. Pxn_C3613.4 유전자는 꽃 발달 과정에서 세포 증식과 분화를 조절하는 전사인자이다(Krizek and Eaddy, 2012). Pxn_C1544.45 꽃의 줄기 세포 성장과 분화 시간을 유도하는 Arabidopsis zinc finger repressor KNUCKLES (KNU) 유전자의 활성을 조절한다(Sun et al., 2014). Pxn_C0156.4 유전자는 저온 반응과 꽃의 개화시기를 조절한다(Seo et al., 2009). 따라서 flower development 기능 범주에 속한 유전자들은 꽃의 세포 성장, 분화, 개화시기를 조절할 것으로 추정되었다. GO enrichment 분석 결과에서 pollen tube growth와 pollen development 기능 범주에 관련된 클러스터 그룹은 부계 클러스터 5번, 6번과 모계 클러스터 9번, 10번으로, 다른 조직보다 수술 조직에서 1.7배~5.3배 차등 발현 패턴을 보였다. Cation/H(+) antiporter는 PH 조절과 칼륨 이온 운송에 관여하는 유전자로 모계와 부계에서 공통으로 수술조직에서 차등한 발현 패턴을 보였다. AGL18은 화분 성숙과 현화식물에서 수배우체 발달에 필수적인 유전자로 부계에서만 차등한 발현을 보였다(Verelst et al., 2007). 부계 클러스터 4번과 5번, 모계 클러스터 7번에서 cell wall biogenesis,

organization, modification 기능 범주에 관련된 유전자 중 다른 조직 보다 수술 조직에서 2.2~3.5배 차등 발현하는 부계 유전자 그룹으로, 세포벽 펙틴 내에서 demethylesterification을 통해 세포벽을 변형시키는 pectinesterase 유전자 패밀리를 포함한다. 또한 부계 클러스터 4번과 모계 클러스터 7번 그룹의 유전자는 열매 조직에서 조직 특이적 발현 패턴을 보였고, 대부분 secondary metabolites biosynthesis에 연관되어 있다. 전체적으로 볼 때, 모계 유래 차등 발현 유전자 보다 부계 유래 차등 발현 유전자가 유전자 클러스터 별로 유전자 기능 범주에서 높은 빈도로 나타났다.

Table 18. Summary statistics of GO and Pathway enrichment in the maternal clusters

Cluster	Category	Term	Count	PValue
Cluster 1	GO BP	GO:0055114 oxidation-reduction process	10	0.001
		GO:0010025 wax biosynthetic process	2	0.041
		GO:0006629 lipid metabolic process	3	0.044
	GO CC	GO:0009505 plant-type cell wall	6	0.000
		GO:0048046 apoplast	6	0.002
		GO:0005618 cell wall	6	0.004
		GO:0005886 plasma membrane	14	0.011
	GO MF	GO:0005506 iron ion binding	5	0.004
		GO:0004497 monooxygenase activity	3	0.044
	KEGG Pathway	ath00073 Cutin, suberine and wax biosynthesis	2	0.050
Cluster 2	GO BP	GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	3	0.012
		GO:0010072 primary shoot apical meristem specification	2	0.019
	GO CC	GO:0009941 chloroplast envelope	5	0.004
		GO:0009507 chloroplast	12	0.005
		GO:0009570 chloroplast stroma	4	0.046
	GO MF	GO:0016491 oxidoreductase activity	4	0.016
		GO:0005524 ATP binding	8	0.037
Cluster 3	GO BP	GO:0005987 sucrose catabolic process	2	0.011
		GO:0009908 flower development	3	0.020
	GO MF	GO:0004575 sucrose alpha-glucosidase activity	2	0.015
	KEGG Pathway	ath01110 Biosynthesis of secondary metabolites	5	0.026
Cluster 4	GO BP	GO:0048765 root hair cell differentiation	2	0.034
	GO MF	GO:0015171 amino acid transmembrane transporter activity	3	0.009
	KEGG Pathway	ath00500 Starch and sucrose metabolism	3	0.038
Cluster 5	GO BP	GO:0019253 reductive pentose-phosphate cycle	3	0.002
		GO:0055114 oxidation-reduction process	10	0.016
		GO:0042372 phylloquinone biosynthetic process	2	0.029
		GO:0010205 photoinhibition	2	0.041
	GO CC	GO:0009507 chloroplast	28	0.000
		GO:0031969 chloroplast membrane	7	0.000
		GO:0009570 chloroplast stroma	10	0.000
		GO:0009535 chloroplast thylakoid membrane	7	0.000
		GO:0009941 chloroplast envelope	6	0.008
		GO:0009579 thylakoid	4	0.012
		GO:0016021 integral component of membrane	20	0.033
		GO:0030095 chloroplast photosystem II	2	0.039
	GO MF	GO:0050278 sedoheptulose-bisphosphatase activity	2	0.021
	KEGG Pathway	ath00710 Carbon fixation in photosynthetic organisms	5	0.000
		ath01200 Carbon metabolism	6	0.002
		ath01100 Metabolic pathways	12	0.031
Cluster 6	GO BP	GO:0055114 oxidation-reduction process	9	0.002
		GO:0005975 carbohydrate metabolic process	4	0.034
		GO:0009695 jasmonic acid biosynthetic process	2	0.045
	GO CC	GO:0009507 chloroplast	18	0.000
		GO:0005737 cytoplasm	14	0.035

	KEGG Pathway	ath01110 Biosynthesis of secondary metabolites	9	0.014
Cluster 7	GO BP	GO:0009834 plant-type secondary cell wall biogenesis	5	0.000
		GO:0045492 xylan biosynthetic process	3	0.001
		GO:0080167 response to karrikin	4	0.002
		GO:0010413 glucuronoxylan metabolic process	2	0.004
		GO:0010417 glucuronoxylan biosynthetic process	2	0.022
		GO:0009698 phenylpropanoid metabolic process	2	0.031
		GO:0055114 oxidation-reduction process	7	0.032
		GO:0019853 L-ascorbic acid biosynthetic process	2	0.038
	GO CC	GO:0005576 extracellular region	11	0.004
		GO:0005794 Golgi apparatus	7	0.007
		GO:0000139 Golgi membrane	4	0.014
	GO MF	GO:1990538 xylan O-acetyltransferase activity	2	0.015
Cluster 8	GO CC	GO:0005886 plasma membrane	7	0.077
Cluster 9	GO BP	GO:0009860 pollen tube growth	3	0.022
	GO CC	GO:0016324 apical plasma membrane	2	0.029
		GO:0016021 integral component of membrane	17	0.034
	GO MF	GO:0016301 kinase activity	6	0.038
Cluster 10	GO BP	GO:0007018 microtubule-based movement	4	0.001
		GO:0006812 cation transport	3	0.017
	GO CC	GO:0005871 kinesin complex	4	0.001
		GO:0005886 plasma membrane	21	0.003
		GO:0070382 exocytic vesicle	2	0.006
		GO:0045177 apical part of cell	2	0.014
		GO:0016324 apical plasma membrane	2	0.045
	GO MF	GO:0008017 microtubule binding	5	0.001
		GO:0003777 microtubule motor activity	4	0.001
		GO:0005215 transporter activity	5	0.020
		GO:0005524 ATP binding	14	0.022
		GO:0016887 ATPase activity	4	0.036

Table 19. Summary statistics of GO and Pathway enrichment in the paternal clusters

Paternal	Category	Term	Count	PValue
Cluster 1	GO BP	GO:0009813 flavonoid biosynthetic process	4	0.006
		GO:0052696 flavonoid glucuronidation	3	0.032
		GO:0055114 oxidation-reduction process	8	0.045
	GO CC	GO:0005576 extracellular region	15	0.001
	GO MF	GO:0016758 transferase activity, transferring hexosyl groups	4	0.007
		GO:0004497 monooxygenase activity	4	0.008
		GO:0080044 quercetin 7-O-glucosyltransferase activity	3	0.026
		GO:0080043 quercetin 3-O-glucosyltransferase activity	3	0.027
		GO:0052793 pectin acetylesterase activity	2	0.028
		GO:0016757 transferase activity, transferring glycosyl groups	5	0.030
		GO:0015020 glucuronosyltransferase activity	2	0.035
		GO:0016705 oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3	0.048
		GO:0052689 carboxylic ester hydrolase activity	3	0.049
Cluster 2	GO BP	GO:0010236 plastoquinone biosynthetic process	3	0.000
		GO:0006633 fatty acid biosynthetic process	4	0.017
		GO:0006855 drug transmembrane transport	3	0.036
		GO:0009624 response to nematode	3	0.040
	GO CC	GO:0009507 chloroplast	42	0.000
		GO:0009941 chloroplast envelope	16	0.000
		GO:0009535 chloroplast thylakoid membrane	12	0.000
		GO:0009570 chloroplast stroma	13	0.000
		GO:0031969 chloroplast membrane	6	0.000
		GO:0009534 chloroplast thylakoid	6	0.001
		GO:0009706 chloroplast inner membrane	4	0.002
		GO:0009536 plastid	6	0.004
		GO:0009543 chloroplast thylakoid lumen	3	0.033
		GO:0016021 integral component of membrane	28	0.034
	GO MF	GO:0005247 voltage-gated chloride channel activity	2	0.033
	KEGG Pathway	ath01110 Biosynthesis of secondary metabolites	11	0.020
		ath01100 Metabolic pathways	15	0.031
Cluster 3	GO BP	GO:0055114 oxidation-reduction process	8	0.005
		GO:0009733 response to auxin	4	0.011
		GO:0006970 response to osmotic stress	3	0.017
		GO:0046777 protein autophosphorylation	3	0.020
		GO:0009813 flavonoid biosynthetic process	3	0.023
		GO:0009688 abscisic acid biosynthetic process	2	0.024
	GO MF	GO:0016705 oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	4	0.002
		GO:0005506 iron ion binding	5	0.003
		GO:0004497 monooxygenase activity	4	0.004

		GO:0020037 heme binding	5	0.005
		GO:0019825 oxygen binding	4	0.009
		GO:0045549 9-cis-epoxycarotenoid dioxygenase activity	2	0.013
	KEGG Pathway	ath00941 Flavonoid biosynthesis	3	0.001
		ath01110 Biosynthesis of secondary metabolites	9	0.003
		ath00945 Stilbenoid, diarylheptanoid and gingerol biosynthesis	3	0.011
Cluster 4	GO BP	GO:0009834 plant-type secondary cell wall biogenesis	4	0.000
		GO:0030244 cellulose biosynthetic process	3	0.005
		GO:0055114 oxidation-reduction process	9	0.007
		GO:0052386 cell wall thickening	2	0.009
		GO:0071555 cell wall organization	4	0.030
		GO:0009833 plant-type primary cell wall biogenesis	2	0.042
	GO CC	GO:0000139 Golgi membrane	4	0.030
	GO MF	GO:0016760 cellulose synthase (UDP-forming) activity	3	0.001
		GO:0016759 cellulose synthase activity	3	0.003
		GO:0046872 metal ion binding	10	0.005
		GO:0050664 oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor	2	0.015
		GO:0016757 transferase activity, transferring glycosyl groups	5	0.020
		GO:0004601 peroxidase activity	3	0.025
		GO:0000977 RNA polymerase II regulatory region sequence-specific DNA binding	3	0.029
	KEGG Pathway	ath00904 Diterpenoid biosynthesis	2	0.049
Cluster 5	GO BP	GO:0045490 pectin catabolic process	10	0.000
		GO:0042545 cell wall modification	8	0.000
		GO:0009860 pollen tube growth	4	0.003
		GO:0050829 defense response to Gram-negative bacterium	2	0.024
		GO:0007338 single fertilization	2	0.034
		GO:0080092 regulation of pollen tube growth	2	0.050
	GO CC	GO:0005618 cell wall	11	0.000
		GO:0009505 plant-type cell wall	9	0.000
		GO:0071944 cell periphery	5	0.000
		GO:0005576 extracellular region	16	0.001
		GO:0016021 integral component of membrane	20	0.023
		GO:0016324 apical plasma membrane	2	0.035
		GO:0090404 pollen tube tip	2	0.035
	GO MF	GO:0045330 aspartyl esterase activity	8	0.000
		GO:0030599 pectinesterase activity	8	0.000
		GO:0046910 pectinesterase inhibitor activity	6	0.000
	KEGG Pathway	ath00040 Pentose and glucuronate interconversions	7	0.000
Cluster 6	GO BP	GO:0006813 potassium ion transport	4	0.001
		GO:0042391 regulation of membrane potential	3	0.004
		GO:0006885 regulation of pH	3	0.005
		GO:0009860 pollen tube growth	4	0.006

		GO:0006812 cation transport	3	0.017
		GO:0009651 response to salt stress	6	0.019
		GO:0009555 pollen development	4	0.024
	GO CC	GO:0016021 integral component of membrane	30	0.000
		GO:0090404 pollen tube tip	3	0.001
		GO:0005887 integral component of plasma membrane	6	0.001
		GO:0005794 Golgi apparatus	9	0.027
	GO MF	GO:0030551 cyclic nucleotide binding	3	0.004
		GO:0005451 monovalent cation:proton antiporter activity	3	0.004
		GO:0005249 voltage-gated potassium channel activity	3	0.005
		GO:0016301 kinase activity	9	0.017
		GO:0005524 ATP binding	14	0.034
Cluster 7	GO BP	GO:0006810 transport	5	0.011
		GO:0009658 chloroplast organization	3	0.039
	GO CC	GO:0009534 chloroplast thylakoid	8	0.000
		GO:0009570 chloroplast stroma	9	0.000
		GO:0009507 chloroplast	20	0.001
		GO:0009579 thylakoid	4	0.011
		GO:0010598 NAD(P)H dehydrogenase complex (plastoquinone)	2	0.029
		GO:0009505 plant-type cell wall	4	0.042

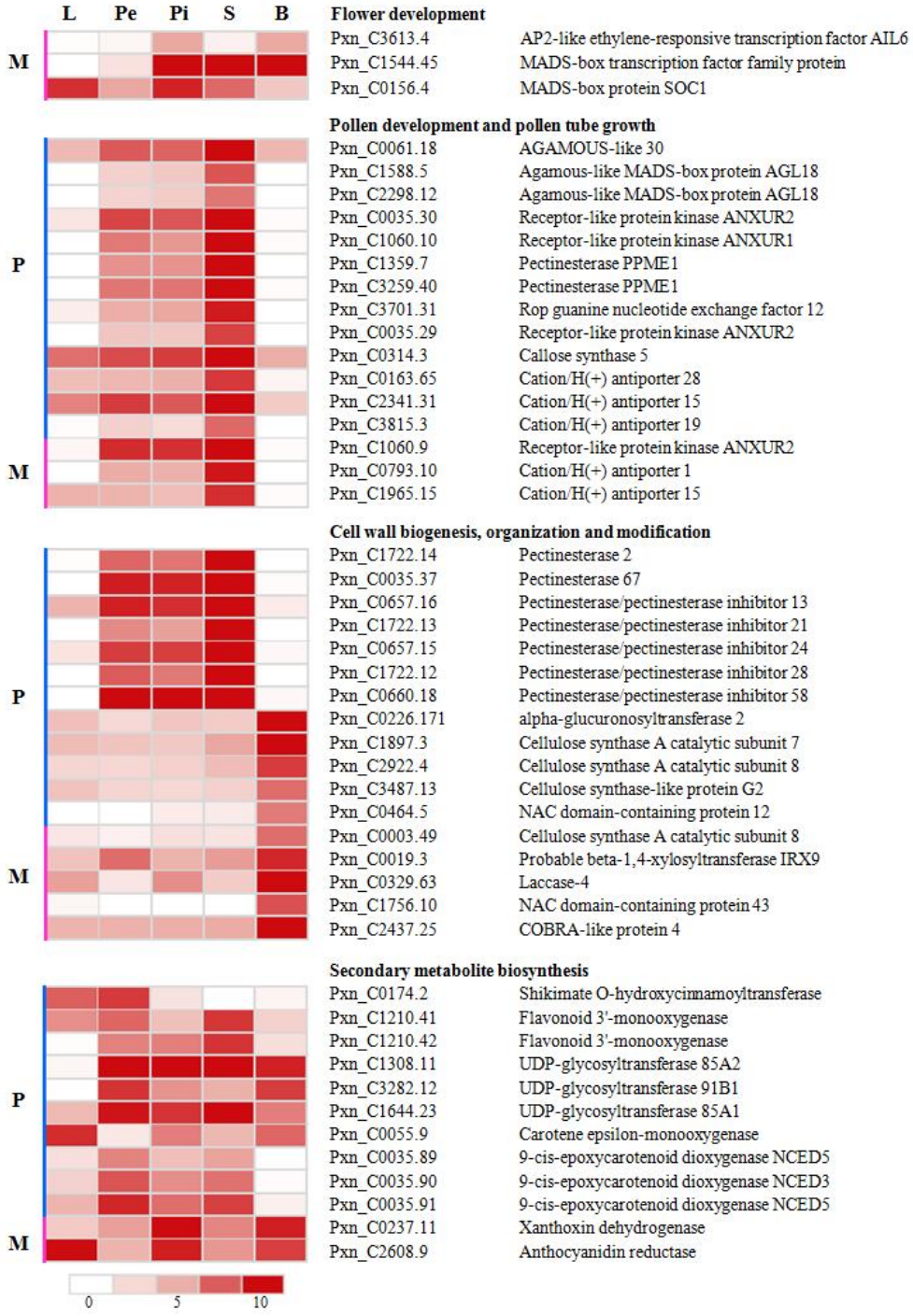


Figure 18. Heat maps showing the differential expression of selected categories of genes related to development and secondary metabolite biosynthesis.
The average normalized count values represent the relative expression across tissues. M maternal-phased genes, P paternal-phased genes.

4. 왕벚나무 유전자의 alternative splicing 분석

자생 왕벚나무 유전자 모델 중 유전자 전사 과정에서 유전자를 구성하는 다수의 엑손이 선택적으로 재조합 되어 동일한 유전자 구조상에서 다양한 전사체를 생성하는지 확인하기 위하여 각각 잎, 꽃잎, 암술, 수술, 열매 조직에서 발생할 수 있는 8 종류의 alternative splicing (AS) 변이체를 분석하였다. AS 분석은 Tophat 프로그램을 사용하여 각 조직별로 mRNA-seq read를 유전자 모델에 매핑하여 수행했다. Cufflinks 프로그램을 사용하여 유전자 모델에 매핑 된 short read의 커버리지와 매핑 영역에서 중첩 영역 고려하여 동일 유전자 영역에서 다양한 isoform 전사체의 위치 정보와 전사체의 FPKM 발현 값을 추출하였다. 차등 발현 isoform 전사체 분석 결과를 SpliceR 프로그램을 이용하여 유전자 영역에서 AS 종류에 해당하는 조직 특이적인 차등 발현 isoform 전사체를 판별하였다.

AS에 해당하는 차등 발현 isoform 전사체는 총 6,090개이며, 차등 발현 isoform 전사체에 해당하는 유전자는 총 4,287개를 선별하였다(Table 20). 5개 조직 중 특히 수술 특이적인 86개 AS isoform이 나타나 수술의 조직 특이성이 가장 높았다. 각 조직 특이적 차등 발현 isoform 전사체에 대해 유전자 기능 범주에 대한 GO enrichment 분석을 수행하였다(Figure 19). 기능 범주에 해당하는 총 18개 전사체 중 12개 isoform 전사체는 alternative transcription start site 방식으로 각각 타 조직 대비 평균 2.9배 차등발현 패턴을 보였다. 특히, 타 조직 대비 수술 조직에서 3.0~4.9배 수준으로 pollen tube growth에 관련된 isoform이 차등발현 하였다.

Table 20. Summary of alternative splicing events identified in protein-coding genes

Alternative splicing events	Gene	Isoform	Tissue-specificity				
			Leaf	Petal	Pistil	Stamen	Berry
Exon skipping/inclusion	1,285	1,505	9	6	9	16	2
Mutually exclusive exon	11	13	0	0	0	0	0
Mutliple exon skipping/inclusion	171	199	2	0	1	1	3
Intron retention	743	818	11	3	1	1	1
Alternative 5′ splice site	879	974	9	5	4	6	3
Alternative 3′ splice site	1,538	1,692	17	2	7	21	6
Alternative transcription start site	2,966	3,416	40	10	15	54	40
Alternative transcription terminating site	2,355	2,417	33	8	13	42	14
Total number of non-redundant type	4,287	6,090	73	15	29	86	27

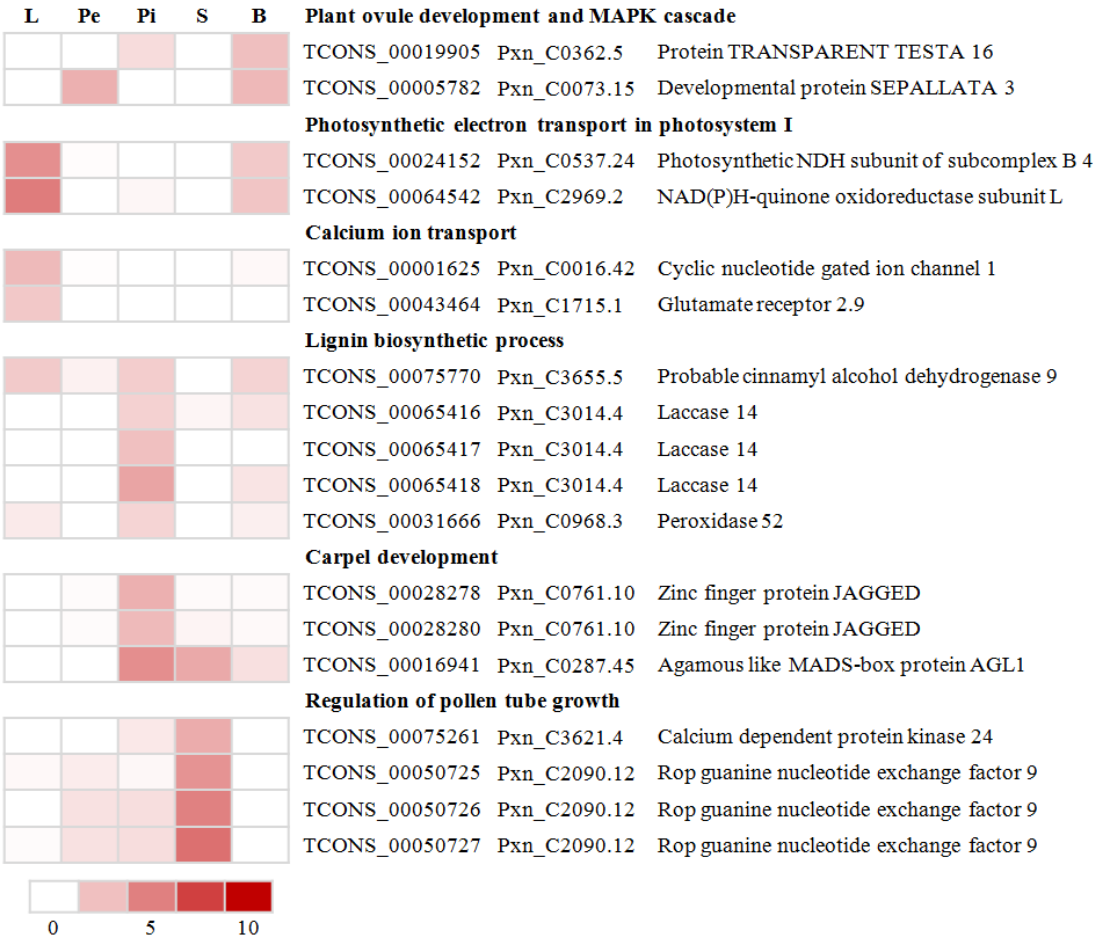


Figure 19. Heat maps showing the differential expression of enriched functional categories of AS isoforms.

The average normalized count values represent the relative expression across tissues.

제 7 절 자생 왕벚나무와 근연종 벚나무류 사이의 유전체 변이 비교 분석

1. 자생 왕벚나무, 소메이 요시노, 근연종 벚나무류에서 변이 분석

제주도 자생 왕벚나무 5개체, 미국 소메이 요시노 2개체, 일본 소메이 요시노 2개체, 올벚나무 3개체, 벚나무 3개체, 사옥 1개체, 산벚나무 1개체 총 16 개체 근연종 벚나무류에서 유전체상의 변이를 분석하였다(Table 21). 먼저 각 개체에서 시퀀싱한 short read 서열을 BWA MEM 프로그램을 이용하여 자생 왕벚나무(Pxn-Jeju2) 전체 유전체에 매핑을 하였고, duplicate read 서열을 제거하여 변이 추출의 오류를 제거하였다. 자생 왕벚나무(Pxn-Jeju2) 유전체 매핑 데이터에서 각 개체별로 shot read 서열의 매핑 커버리지를 분석하였다. 자생 왕벚나무 개체군에서 매핑 커버리즈는 94.5%~98.6%로 높은 매핑 비율을 확인하였다. 소메이 요시노 개체군 매핑 커버리지는 94.9%~95.0%로 자생 왕벚나무 개체보다 다소 낮은 매핑 비율을 보였다. 올벚나무, 벚나무, 사옥, 산벚나무에서 매핑 커버리지는 72.0%~86.9%로 자생 왕벚나무와 소메이 요시노 매핑 커버리지 보다 매핑 비율이 낮았다. 자생 왕벚나무 유전체 조립 서열은 올벚나무 유전형과 벚나무 유전형이 반영된 haplotype-fused 특성을 갖고 있어 벚나무류에서 낮은 매핑 커버리지 보인 것으로 판단된다.

GATK 프로그램 HaplotypeCaller를 이용하여 총 16 개체에서 SNP와 InDel로 구분되는 변이를 추출하였다. 총 변이 개수는 76,427,804개로서 InDels 13,663,164개와 SNP 62,764,640개로 나타났다. Heterozygous SNP 개수는 homozygous SNP 개수 보다 1.7배이며, 전체 변이에서 52%에 해당했다. 소메이 요시노에서 heterozygous SNP 평균 개수는 3,491,676개이며 자생 왕벚나무 heterozygous SNP 평균 개수 2,931,544개 대비 1.2배 높은 비율을 보였다. 특히 Pxn-Jeju5 왕벚나무의 경우 heterozygous SNP와 homozygous SNP, InDels 변이 비율은 소메이 요시노와 매우 유사했다. 왕벚나무의 모계 계통 올벚나무의 평균 변이 개수는 3,284,021개로 부계 계통 벚나무의 평균 변이 개수 3,944,728개 보다 낮았다.

변이 서열상에서 개체 차이를 분석하기 위해 다중 서열 정렬 방법을 이용하였다. 변이 데이터에서 VCFtools 프로그램을 이용하여 동일 위치에 있는 3,936,303 bp SNP 변이 서열을 추출하였다. MAFFT 프로그램을 이용하여 변이 서열을 다중 정렬한 후 MEGA 7 프로그램에서 maximum likelihood (ML) 알고리즘을 사용하여 계통수 분석

을 수행하였다. ML 계통수에서 부계 그룹(벚나무, 사옥, 산벚나무)과 모계 그룹(올벚나무)은 서로 다른 두 그룹으로 나누어졌으며, 자생 왕벚나무 그룹과 소메이 요시노 그룹은 모계 올벚나무 그룹과 부계 벚나무류 중간에 위치하고 있으나 서로 다른 그룹으로 뚜렷하게 나누어졌다(Figure 20). 특히, 자생 왕벚나무 중 Pxn-Jeju5는 SNP와 InDels 비율이 소메이 요시노와 매우 유사 할 뿐만 아니라 계통수 분석 결과에서도 소메이 요시노 그룹에도 속해 있어 자생 왕벚나무가 아니라 소메이 요시노로 판단되었다.

SNP 데이터를 multidimensional scaling (MDS)에 의한 분류군간 차이를 분석하기 위해 주성분 분석(principal component analysis, PCA)을 수행하였다. PLINK 프로그램을 이용하여 16 개체 변이 데이터에서 MDS 값을 계산하였고 R 패키지에서 ggplot 프로그램을 사용하여 PCA plot 상에서 각 분류군을 표시하였다(Figure 21). 그 결과, 자생 왕벚나무 개체는 모계 올벚나무 그룹과 부계 벚나무류 그룹의 중간에 위치하고 있으며, 계통수 분석 결과와 동일한 그룹으로 구분할 수 있었다. 또한 Pxn-Jeju5 개체는 소메이 요시노 그룹에 함께 묶여 있었다.

따라서 16개 개체의 벚나무 분류군에서 유전체 변이 비율 비교, 계통수 분석 및 주성분 분석 결과를 종합적으로 고려할 때, 제주도 자생 왕벚나무는 모계 올벚나무와 부계 벚나무 사이에서 교잡된 잡종 세대이며, 소메이 요시노는 자생 왕벚나무와 뚜렷하게 구분되는 서로 다른 유전형을 가진 계통임이 확인되었다.

Table 21. Summary of SNP and InDel variations in *Prunus* species

Taxon	Genome coverage (%)	Read mapped rate (%)	Variation				SNP in CDS					InDel in CDS		SNP+InDel	
			Hetero SNP[a]	Homo SNP[b]	InDel	Overall	Silence	Nonsense	Missense	Splicing site	Ti /Tv[a]	In-frame	Frame shift	Intron	Intergenic
Pxn-Jeju1	95.4	94.4	2,838,985	1,096,549	895,813	4,831,347	148,300	3,661	182,107	1,158	1.56	4,340	22,281	255,057	4,214,443
Pxn-Jeju2	98.6	95.6	2,596,728	145,287	764,810	3,506,825	99,482	2,560	123,473	851	1.52	2,967	20,894	180,133	3,076,465
Pxn-Jeju3	95.7	93.9	2,946,741	1,262,874	932,447	5,142,062	153,376	3,679	187,285	1,259	1.57	4,421	22,453	269,721	4,499,868
Pxn-Jeju4	95.3	94.5	2,896,646	1,223,511	916,719	5,036,876	152,756	3,652	186,991	1,222	1.57	4,337	22,262	266,970	4,398,686
Pxn-Jeju5	94.5	93.9	3,378,621	1,657,723	1,063,188	6,099,532	176,065	4,532	218,722	1,501	1.58	5,187	23,776	313,087	5,356,662
Pxy-US1	95.0	94.6	3,464,156	1,629,328	1,072,679	6,166,163	175,976	4,596	220,018	1,527	1.58	5,244	23,717	315,788	5,419,297
Pxy-US2	94.9	94.5	3,445,971	1,626,652	1,068,387	6,141,010	175,718	4,600	219,181	1,490	1.58	5,195	23,653	314,928	5,396,245
Pxy-JP1	94.9	91.3	3,527,140	1,722,941	1,109,412	6,359,493	182,173	4,620	225,405	1,538	1.58	5,329	24,706	324,046	5,591,676
Pxy-JP2	94.9	92.6	3,529,437	1,724,190	1,110,969	6,364,596	182,189	4,639	225,537	1,557	1.58	5,343	24,780	324,679	5,595,872
Ppa-1	82.2	93.3	1,197,274	1,428,540	595,858	3,221,672	87,944	2,581	116,320	819	1.6	2,857	14,995	165,140	2,831,016
Ppa-2	72.0	94.0	726,505	1,263,932	482,136	2,472,573	74,707	1,966	94,308	660	1.56	2,216	13,013	136,354	2,149,349
Ppa-3	84.6	96.0	1,883,696	1,522,293	751,829	4,157,818	118,186	3,088	148,056	1,001	1.59	3,340	17,363	218,562	3,648,222
Pjj-1	86.9	93.8	1,975,049	1,823,000	776,618	4,574,667	144,854	3,306	173,239	1,120	1.55	4,011	17,776	238,548	3,991,813
Pjj-2	76.1	94.8	1,139,292	1,592,058	583,439	3,314,789	120,474	2,333	138,885	837	1.52	3,104	14,767	183,366	2,851,023
Pjq	86.4	94.0	1,946,474	1,769,052	763,036	4,478,562	145,116	3,211	172,508	1,119	1.55	3,924	17,563	233,810	3,901,311
Psa	86.6	93.7	1,960,926	1,823,069	775,824	4,559,819	144,689	3,312	173,070	1,131	1.55	3,957	17,591	240,595	3,975,474
Total	-	-	39,453,641	23,310,999	13,663,164	76,427,804	2,282,005	56,336	2,805,105	18,790	-	65,772	321,590	3,980,784	66,897,422

[a]Heterozygous SNP rate, proportion of heterozygous SNPs in a genome.

[b]Homozygous SNP rate, proportion of homozygous SNPs in a genome.

[c]Ti, transition; Tv, transversion.

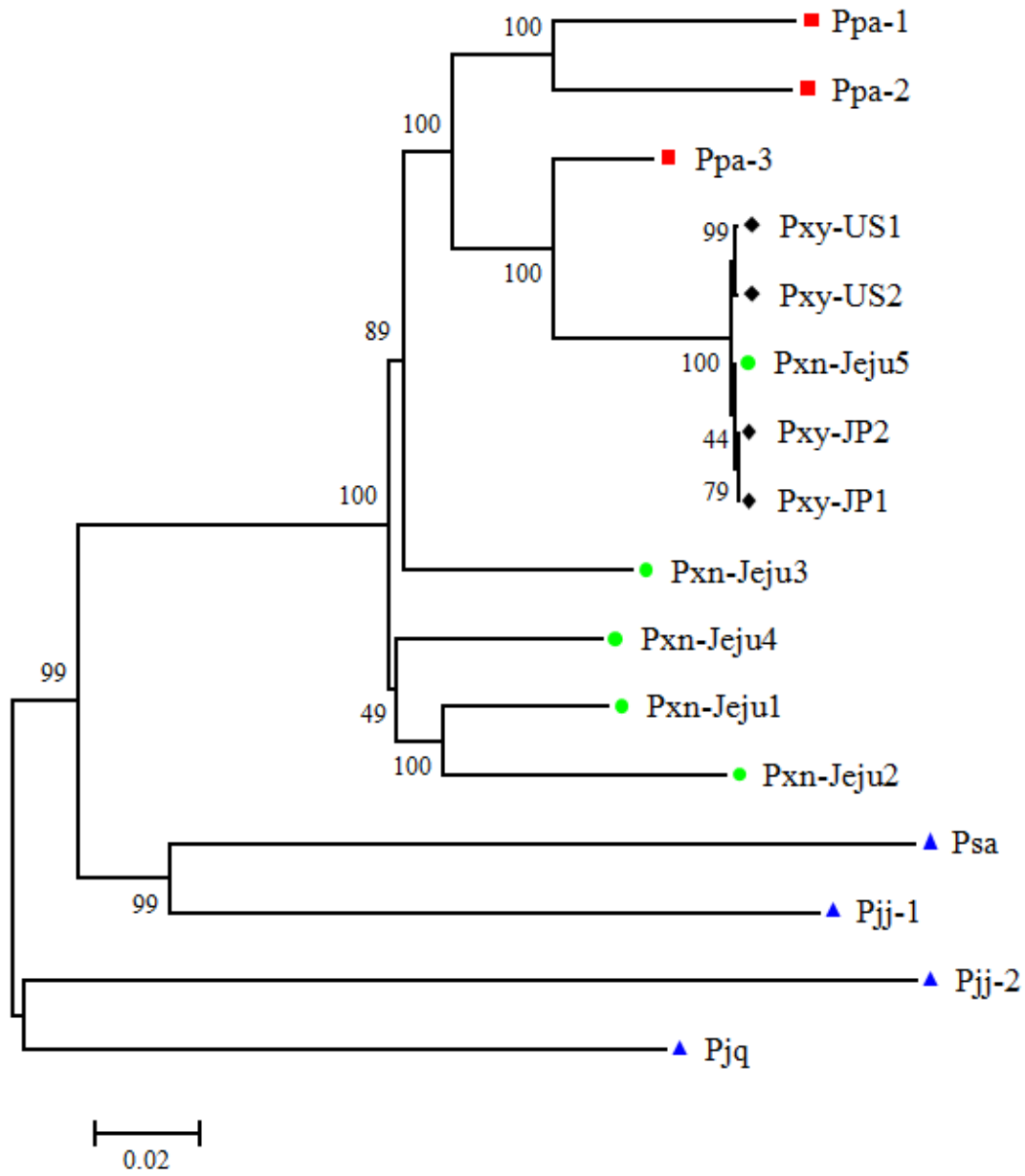


Figure 20. A maximum likelihood tree of *Prunus* accessions based on SNPs/InDels identified by variome analysis.

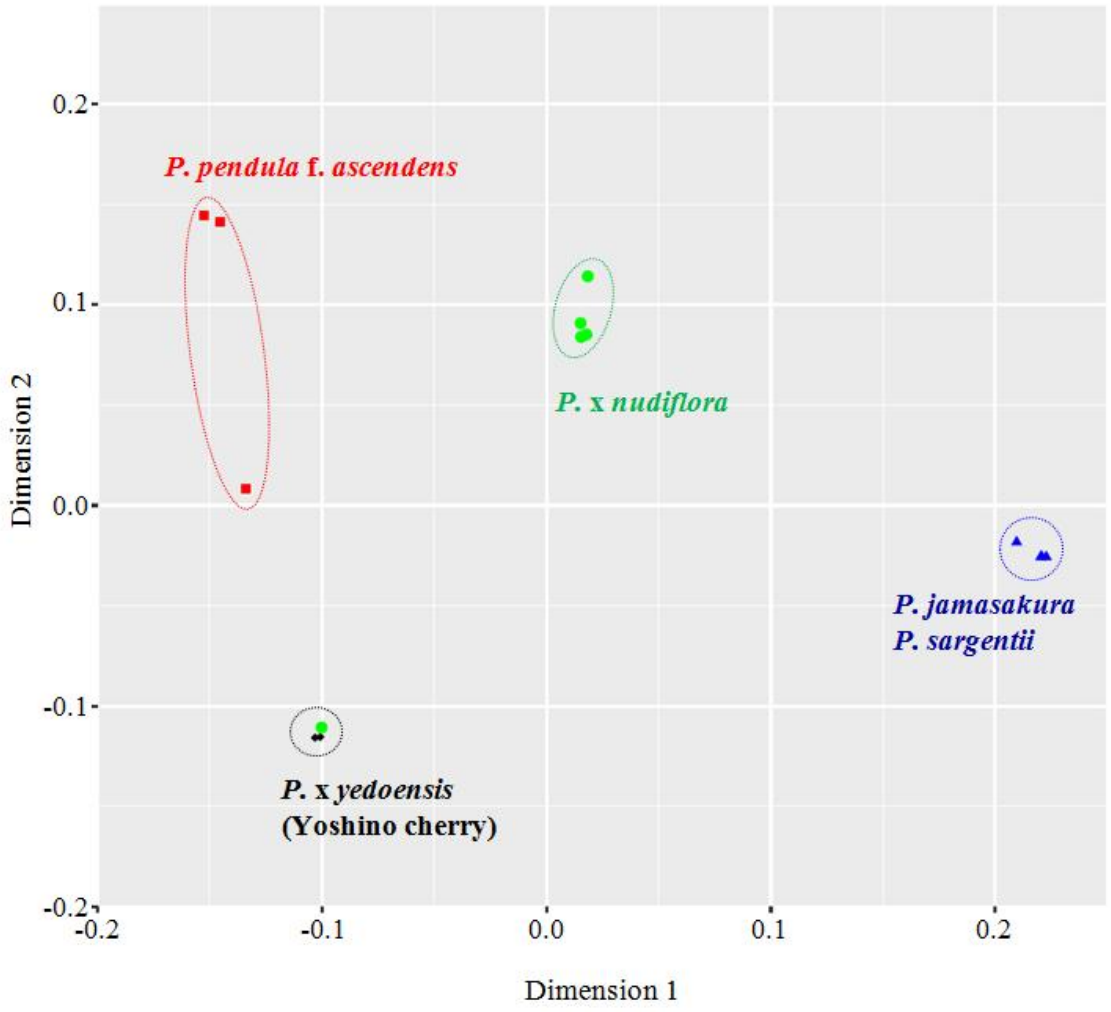


Figure 21. Multidimensional scaling of *Prunus* accessions.
Closely related accessions of Ppa (red square symbol), Pxn (green circle symbol), Pj or Psa (blue triangle symbol), and Pxy (black diamond symbol) are grouped together using dotted circles.

2. 근연종 벚나무에서 자가불화합성 S-locus의 반수체 분석

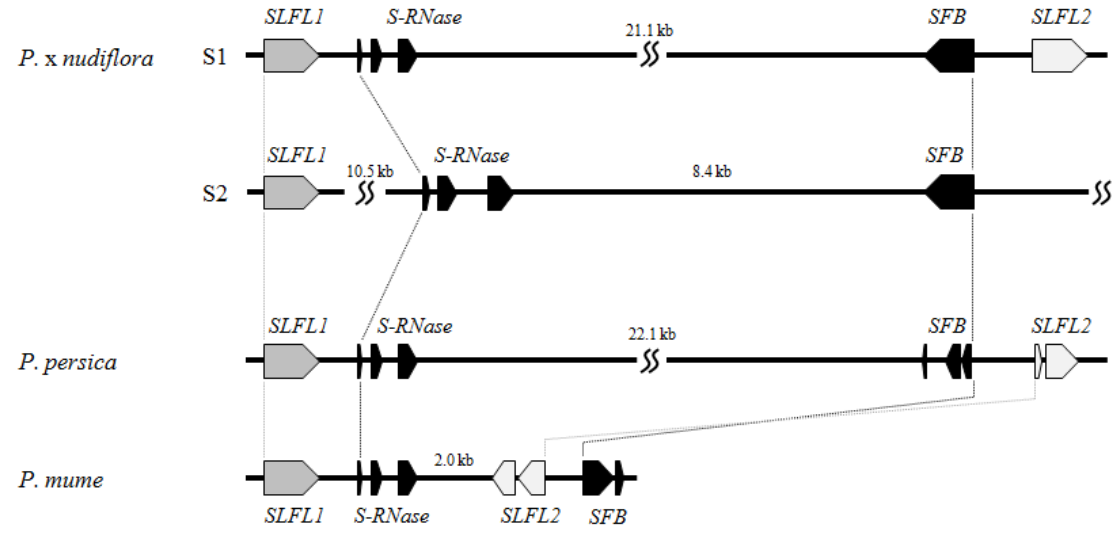
제주도 자생지내 3 km 반경으로 분포하는 부계 벚나무와 모계 올벚나무 그리고 자생 왕벚나무에서 자가불화합성 S-locus 영역의 유전자 구조를 비교 분석하였다(Figure 22).

자생 왕벚나무에서 반수체 S1에서 S-locus 구조는 *S-locus F box-like1* (*SLFL1*), *S-RNase*, *SFB*, *S-locus F box-like2* (*SLFL2*)로 이루어져 있다. 반수체 S2에서 S-locus 구조는 11 kb 길이에 해당하는 *SLFL2* 유전자 영역은 예측되지 않았다. 장미과의 복숭아와 매실에서 S-locus 영역을 탐색하여 비교한 결과, 복숭아 S-locus에서 *SLFL1, S-RNase*, *SFB*, *SLFL2* 유전자 순서는 왕벚나무의 S1과 모두 일치하였으나, 매실의 경우 S-locus 영역에 포함된 유전자 모두 존재하였지만 *SLFL2*과 *SFB*의 순서가 뒤집힌 구조를 보였다(Figure 22A). *S-RNase* 유전자는 암술에서 조직 특이적 발현을 보이며(McClure and Franklin-Tong, 2006), *SFB* 유전자는 꽃가루와 꽃밥에서 조직 특이적 발현을 보이는 것으로 보고되었다(Ushijima et al., 2003). *S-RNase* 유전자 발현은 반수체 S1에서 FPKM 274.9, S2에서 FPKM 568.2로 총 7개 조직 중 암술에서만 조직 특이적 발현을 보였으며, *SFB* 유전자 발현은 반수체 S1에서 FPKM 2.8, S2에서 FPKM 6.1로 수술 조직에서만 조직 특이적 발현을 보여(Figure 22B) 이들이 자가불화합성 S-locus임을 확인하였다. 사과속(*Malus*)와 벚나무속(*Prunus*)에서 *SFB* 유전자 발현은 *S-RNase* 유전자 보다 70배 낮게 발현된다고 보고되었는데(Aguiar et al., 2015), *SFB* 유전자 평균 발현 값이 *S-RNase* 유전자 보다 95배로 낮아 자생 왕벚나무에서도 S-locus 유전자의 발현 양상을 확인하였다.

S-RNase 유전자와 *SFB* 유전자 서열 비교 및 계통수 분석을 통해 왕벚나무와 근연종 벚나무 간에서 15개로 구분되는 반수체 S-locus 연결 관계를 분석하였다(Figure 23). S1 반수체는 Pxn-Jeju1과 Pxn-Jeju2에서 공유하고 있고, S3 반수체는 Pxn-Jeju5와 모계 Ppa-3, S4 반수체는 Pxn-Jeju4와 부계 Pjj-1, S8 반수체는 Pxn-Jeju3, Pxn-Jeju4와 모계 Ppa-3, S10 반수체는 Pxn-Jeju3과 부계 Pjj-1, Pjj-2, S12 반수체 Psa와 Pjq에서 연결 관계를 보였다. 특히 S8 반수체 모계 Ppa-3과 Pxn-Jeju3, Pxn-Jeju4, S10 반수체 부계 Pjj-1, Pjj-2와 Pxn-Jeju3, S4 반수체 부계 Pjj-1은 Pxn-Jeju4에서 상호 연결되어 Pxn-Jeju3, Pxn-Jeju4의 부계는 Pjj-1, Pjj-2, 모계는 Ppa-3의 S-loucs로 확인되었다.

또한 엽록체 전체 서열 비교 분석 결과에서 차이가 나는 서열 개수는 모계 Ppa-3 대비 Pxn-Jeju3에서 14개, Pxn-Jeju4에서 10개로 다른 왕벚나무 보다 2.0배~7.8배 낮게 나타났다(Table 22). 왕벚나무와 근연종 벚나무에서 S-locus 연결 관계 분석을 통해 자생 왕벚나무는 부계 벚나무와 모계 올벚나무로 부터 생성된 이종간 동소적 동배수성 잡종(sympatric homoploid hybrid)으로 사료되었다.

A



B

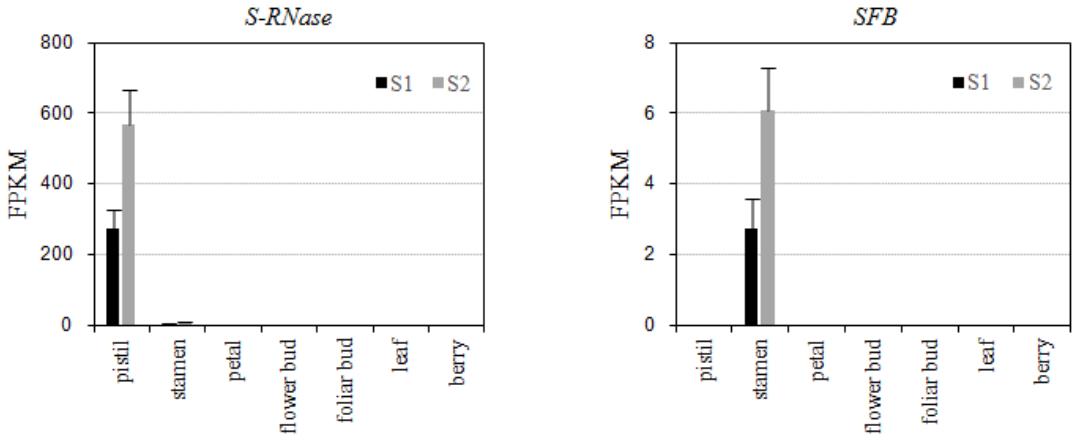


Figure 22. Characterization of S haplotypes in flowering *Prunus* species.

A. Microsynteny at the S-locus regions between wild *P.* x *nudiflora* and fruit crop *Prunus* (*P. persica* and *P. mume*).

There are two S haplotypes in the heterozygous Pxn genome (31 kb of S1 and 35 kb of S2) compared to a single S haplotype in the homozygous Pp and Pm genomes. Syntenic genes are connected with lines.

B. Expression levels of the S-locus genes.

Relative expression levels of the *S-RNase* and *SFB* genes in different tissues are presented by the average fragments per kilobase million (FPKM) value from three independent biological replicates.

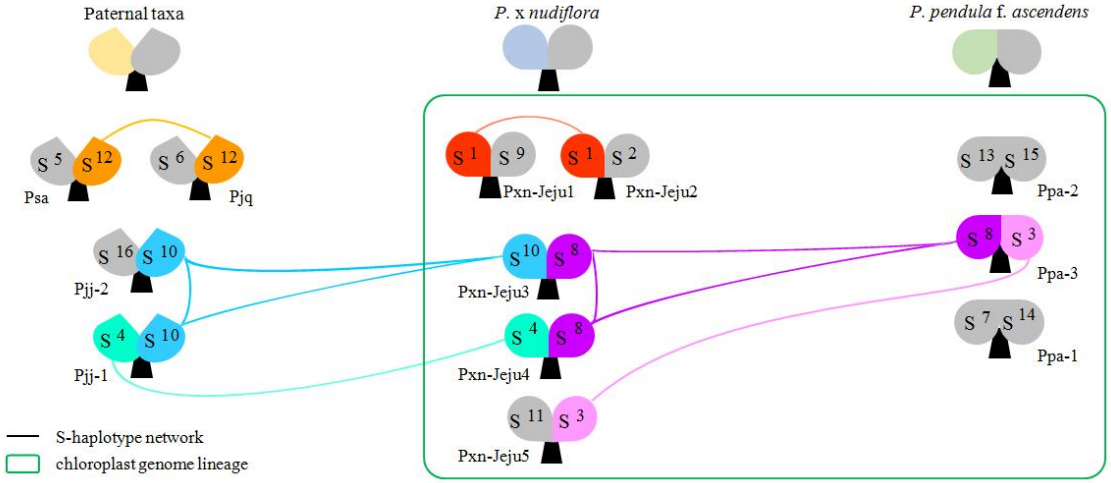


Figure 23. S haplotype network in a natural *Prunus* population.
A total of 15 S haplotypes from 12 accessions, which are distributed sympatrically in a natural habitat on Jeju Island, were identified. Accessions are placed according to their relative geographic location in the natural habitat. Shared S haplotypes between accessions are connected with lines of the same color. Chloroplast genome lineage, showing < 10 nucleotide differences in the proteincoding sequences of the whole chloroplast DNA is also presented in the green box.

Table 22. Comparison of the chloroplast genomes between *Prunus* accessions

Accessions	Ppa-1	Ppa-2	Ppa-3	Pxn-Jeju1	Pxn-Jeju2	Pxn-Jeju3	Pxn-Jeju4	Pxn-Jeju5	Pjj-1	Pjj-2	Pjq	Psa
Ppa-1		3	0	1	0	1	0	3	81	83	97	96
Ppa-2	47		3	4	3	4	3	0	84	86	100	99
Ppa-3	13	56		1	0	1	0	3	81	83	97	96
Pxn-Jeju1	69	34	78		1	2	1	4	82	84	98	97
Pxn-Jeju2	18	63	29	85		1	0	3	81	83	97	96
Pxn-Jeju3	7	48	14	70	23		1	4	82	84	96	97
Pxn-Jeju4	3	46	10	68	19	4		3	81	83	97	96
Pxn-Jeju5	47	0	56	34	63	48	46		84	86	100	99
Pjj-1	631	607	640	632	628	630	630	607		14	24	23
Pjj-2	604	582	613	607	601	603	603	582	214		30	29
Pjq	596	574	605	599	593	597	595	574	164	177		1
Psa	604	582	613	607	601	601	603	582	172	187	12	

Pairwise nucleotide distances between chloroplast DNA sequences of 12 *Prunus* accessions are presented.

X axis : Whole chloroplast DNA, Y axis: 79 protein coding genes.

Upper half, numbers of nucleotide differences in the coding sequences of 79 protein coding genes.

Bottom half, numbers of nucleotide differences in the whole chloroplast DNA sequences.

제 8 절 자생 왕벚나무 판별 분자마커 후보 유전자의 선발

자생 왕벚나무와 소메이 요시노의 종 판별을 위해 Conserved Ortholog Set (COS) 유전자를 기반으로 분자 마커를 선발하였다.

COS 유전자 선별을 위해 자생 왕벚나무와 복숭아, 매실, 체리의 유전자가 암호화하는 단백질 서열을 OrthoMCL 프로그램으로 비교하여 5,751개의 single copy gene COS 유전자를 선별하였다. COS 유전자의 인트론 영역 중에서 자생 왕벚나무와 소메이 요시노 간에 차이가 있는 InDel 서열을 분석하였다. 16개체 벚나무류의 변이 분석 결과에서 InDel 서열은 자생 왕벚나무(Pxn-Jeju1, Pxn-Jeju2, Pxn-Jeju3, Pxn-Jeju4)에서 참조 서열에 동형접합이며 소메이 요시노(Pxy-US1, Pxy-US2, Pxy-JP1, Pxy-JP2) 그리고 Pxn-Jeju5에서 참조 서열에 이형접합을 나타내는 COS 유전자 영역에 포함된 InDel 영역 총 2,769개, COS 유전자 총 1,682개를 선별하였다(Table 23).

자생 왕벚나무와 소메이 요시노에서 차이가 있는 InDel의 길이 분포는 최소 2 bp에서 369 bp 까지 다양하게 나타났으며, 10 bp 단위 서열 길이 분포 중 InDel 길이 2 bp~9 bp에 해당하는 인트론 영역은 전체 대비 70.2%인 1,944개로 가장 높은 분포를 나타냈고, 나머지 10 bp~369 bp에 해당하는 인트론 영역은 COS 유전자 653개에서 825개 존재했다. COS 유전자의 인트론 영역을 대상으로 PCR을 통해 증폭된 서열이 전기 영동을 수행한 젤 상에서 자생 왕벚나무와 소메이 요시노 사이에 크기 차이를 확인 할 수 있는 InDel 서열 길이 10 bp 이상, 인트론 길이 600 bp 이하 기준을 적용하여 총 348개 COS 유전자를 선정하였다. 이렇게 선발한 COS 유전자 중 염색체 별로 Pxn1 80개, Pxn2 46개, Pxn3 33개, Pxn4 38개, Pxn5 36개, Pxn6 39개, Pxn7 39개, Pxn8 28개로 총 339개 COS 유전자의 위치를 결정하였으나, 나머지 9개 유전자는 스케폴드에 위치하여 제외하였다(Figure 24). 8개 염색체상의 위치가 결정된 COS 유전자 339개는 분자 마커 후보로 최종 선정하였다.

Table 23. Summary statistics of InDel length distribution in COS genes between wild *P.* x *nudiflora* and Somei-yoshino

Cutoff InDel length (bp)	Total Intron	Total gene	Intron length ≦ 600 bp		Intron length ≦ 1 kb	
			Intron count	Gene count	Intron count	Gene count
2~9	1,944	1,335	1,289	958	1,673	1,173
10~19	407	361	221	204	306	277
20~29	146	142	73	72	100	97
30~39	102	99	48	45	72	69
40~49	56	55	25	25	37	37
50~59	27	27	11	11	17	17
60~69	31	31	9	9	16	16
70~79	12	12	4	4	9	9
80~89	14	14	4	4	10	10
90~99	7	7	3	3	4	4
100~109	6	6	1	1	1	1
110~119	1	1	0	0	0	0
120~129	4	4	0	0	1	1
130~139	2	2	0	0	0	0
150~159	1	1	0	0	0	0
160~169	2	2	0	0	0	0
170~179	1	1	1	1	1	1
180~189	2	2	1	1	1	1
200~209	1	1	0	0	1	1
210~219	2	2	0	0	0	0
360~369	1	1	1	1	1	1
10~369	825	653	402	348	577	478
Total	2,769	1,682	1,691	1,172	2,250	1,439

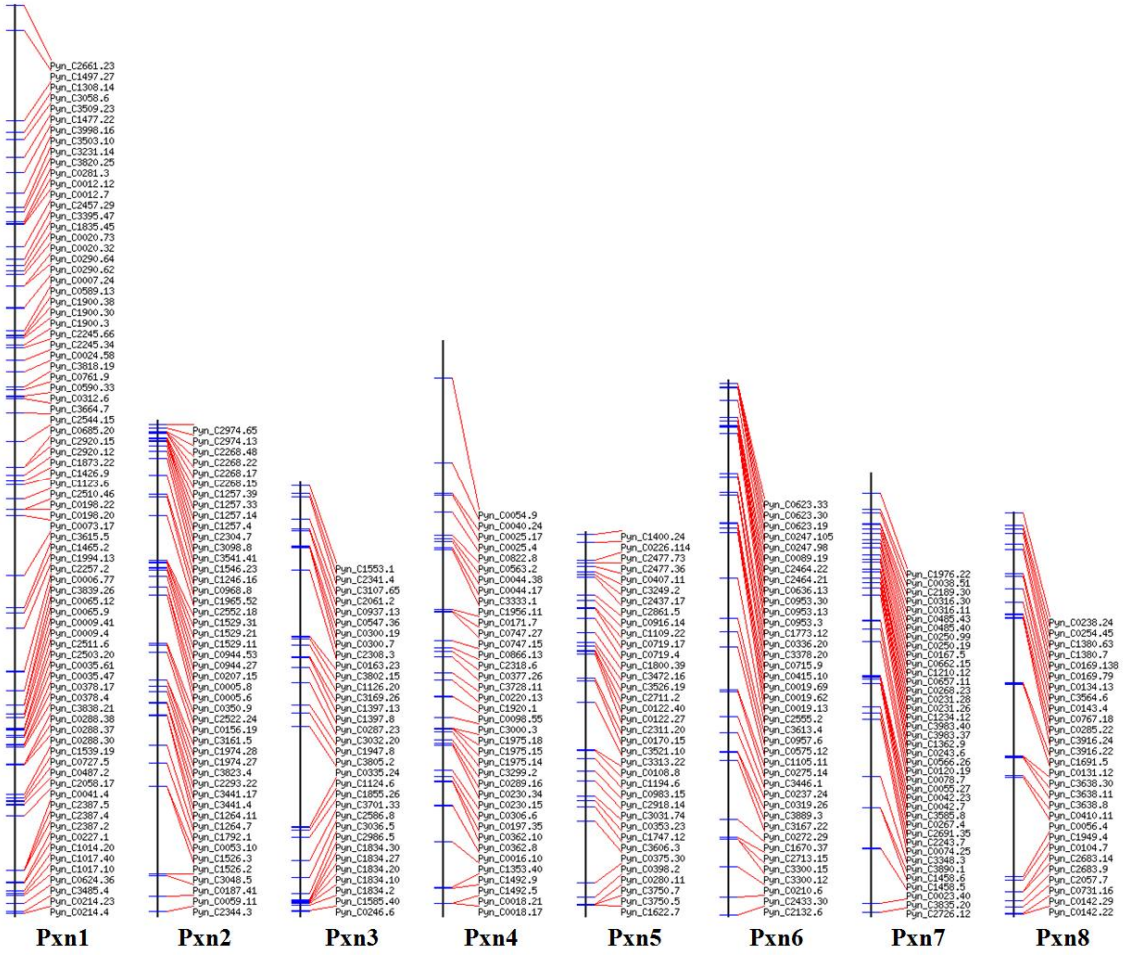


Figure 24. Locations of 339 candidate species-diagnostic COS genes in the *P.* x *nudiflora* chromosomes.

Blue lines represent positions of COS genes mapped to each chromosome.

제 4 장 결 론

본 연구에서는 제주도 자생 왕벚나무 기념목 지정 개체 Pxn-Jeju2의 전체 유전체를 NGS 기술로 해독하고 장거리 서열 PacBio subread를 이용하여 전체 유전체를 조립한 후 유전자 모델 예측, 염색체 수준의 조립 서열 정렬, 부계와 모계 유래 유전자의 구분과 이들의 조직 특이적 발현 패턴 분석, 근연종 벚나류 및 소메이 요시노와 유전체 변이 분석, 장미과 비교 유전체 분석 등 수행하여 자생 왕벚나무 유전체의 특성과 기원을 분석하였다.

Illumina short read 서열에 대한 K-mer 분석을 수행한 결과에서 자생 왕벚나무 유전체는 이형접합성을 나타내는 두 개의 곡선 분포를 보였고, 이형접합 서열 분포의 정점에 해당하는 K-mer 빈도는 동형접합 서열 분포의 정점에 해당하는 K-mer 빈도보다 2배 높았다. 이러한 K-mer 빈도 분포 양상은 자생 왕벚나무 4개체, 미국과 일본에서 채집한 왕벚나무 4개체에서도 동일하게 나타났다. 따라서 제주도 자생 왕벚나무뿐만 아니라 미국 및 일본 왕벚나무는 모두 이형접합성이 높은 잡종 유전체의 특성을 가지고 있다.

이형접합도가 높은 자생 왕벚나무 전체 유전체 조립을 위해 PacBio subread 서열을 FALCON 어셈블러를 사용하여 N50 길이 198.9 kb으로 구성된 전체 유전체 컨티그 서열 323.7 Mb를 작성하였다. *De bruijn* graph 알고리즘을 적용한 ALLPATH-LG 조립 서열과 overlap layout consensus 알고리즘을 적용한 Cellera 어셈블러의 hybrid 조립 서열을 추가 제작하고 FALCON 조립 서열과 비교하여 전체 유전체 완성도를 비교하였다. FALCON 조립 서열은 ALLPATH-LG와 Cellera 어셈블러 조립 서열을 서열 일치도 약 88.6%~91.2% 수준에서 커버하였고, 조립 서열의 N50~N90 평균 길이도 이들의 2.7~13.2배 수준으로 조립 서열 연속성과 완성도가 높았다. 벚나무속 복숭아, 체리, 매실 유전체와 자생 왕벚나무 조립 서열을 재배열하여 비교 분석한 결과에서 각 유전체의 염색체별로 약 49%~71% 수준에서 정렬되었고, synteny 영역의 오솔로그 유전자는 동일 선형을 나타냈다. 비교 유전체 분석 결과에서 복숭아와 매실의 서열 일치성은 유사하였고, 자생 왕벚나무와 체리 유전체 서열 일치성은 벚나무속 중에서 가장 높았다. 장미과 오솔로그 유전자를 이용한 Ks 분석 결과, 체리 오솔로그 유전자의

Ks 값은 자생 왕벚나무 파랄로그 유전자의 Ks 값과 함께 가장 작았고, 매실과 복숭아의 Ks 값은 동일하여 복숭아와 매실, 체리, 벚나무가 순차적으로 종 분화했다고 사료된다. 또한 분자 연대 측정을 통해 벚나무속은 66백만 년 전에 분기되어 종 분화가 일어났으며, 44백만 년 전에 복숭아와 매실, 36백만 년 전에 벚나무와 체리가 분기된 것으로 계산되었다.

근연종 벚나무류와 미국 및 일본 소메이 요시노 그리고 제주도 자생 왕벚나무에서 유전체 서열 변이 분석을 통해 왕벚나무와 근연종 벚나무 분류군의 유연관계를 확인하였다. 벚나무 유연관계에 있는 유전체 서열을 매핑하여 SNP 정보 분석을 통해 양친의 유전자를 분류하였다. 자생 왕벚나무 유전체는 양친 모두에서 유래한 유전자는 59.2%, 모계 유래 유전자 19.4%, 부계 유래 유전자 21.4%로 구성되어 있고, 부계 유전자 개수와 모계 유래 유전자 개수가 유사하여 벚나무와 올벚나무의 1세대 잡종임이 밝혀졌다. 분류군간 SNP 데이터에서 계산한 MDS 값을 사용하여 주성분 분석을 수행한 결과, 모계 올벚나무 개체와 부계 벚나무 개체는 각각 하나의 집단을 형성하였고 모계와 부계 집단 사이에 제주도 자생 왕벚나무 개체들이 집단으로 위치하였다. 한편, 소메이 요시노는 자생 왕벚나무와 독립적인 다른 집단을 형성하였다. 따라서 자생 왕벚나무와 소메이 요시노는 서로 다른 유전자형을 가진 벚나무속 별개 분류군임을 확인하였다.

벚나무의 배우체 자가 불화합성 특성을 결정하는 반수체 S-locus의 구조를 분석한 결과, 왕벚나무 Pxn-Jeju3과 Pxn-Jeju4는 모계 올벚나무 Ppa-3과 부계 벚나무 Pjj-1 또는 Pjj-2의 서로 다른 S-locus haplotype 각각을 갖고 있음을 확인하였다. 따라서 제주도 자생 왕벚나무의 지역적 분포 특성과 벚나무속 자생 왕벚나무와 벚나무류에서 유전체 변이 그룹과 배우체 자가 불화합성 S-locus의 구조를 종합적으로 고려할 때, 제주도 자생 왕벚나무는 부계 벚나무와 모계 올벚나무의 종간 교잡에 의해 생성된 F1 잡종으로 사료된다.

참 고 문 헌

Aguiar, B., Vieira, J., Cunha, A. E., Fonseca, N. A., Iezzoni, A., van Nocker, S., and Vieira, C. P. (2015). Convergent evolution at the gametophytic self-incompatibility system in *Malus* and *Prunus*. PLoS One 10, e0126138.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12, R18.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol 215, 403-410.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Baek, S., Choi, K., Kim, G. B., Yu, H. J., Cho, A., Jang, H., Kim, C., Kim, H. J., Chang, K. S., Kim, J. H., and Mun, J. H. (2018). Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. Genome Biol 19, 127.

Bailey, L. H. and Bailey, E. Z.. (1976). Prunus L. Hortus Third; A Concise Dictionary of Plants Cultivated in the United States and Canada. Macmillan Publishing Company, New York.

Bao, Z., and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12, 1269-1276.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6, 11.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter,

N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33, 623-630.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81, 1084-1097.

Chagne, D., Crowhurst, R. N., Pindo, M., Thrimawithana, A., Deng, C., Ireland, H., Fiers, M., Dzierzon, H., Cestaro, A., Fontana, P., Bianco, L., Lu, A., Storey, R., Knabel, M., Saeed, M., Montanari, S., Kim, Y. K., Nicolini, D., Larger, S., Stefani, E., Allan, A. C., Bowen, J., Harvey, I., Johnston, J., Malnoy, M., Troggio, M., Perchepied, L., Sawyer, G., Wiedow, C., Won, K., Viola, R., Hellens, R. P., Brewer, L., Bus, V. G., Schaffer, R. J., Gardiner, S. E., and Velasco, R. (2014). The draft genome sequence of European pear (*Pyrus communis* L. 'Bartlett'). PLoS One 9, e92644.

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238.

Cheng, S.. Mcbride, J. R. and Fukunar, K.. (2000). The urban forest of Tokyo. Arboricultural Journal 23: 379-392.

Chew, W., Hrmova, M., and Lopato, S. (2013). Role of Homeodomain leucine zipper (HD-Zip) IV transcription factors in plant development and plant protection from deleterious environmental factors. Int J Mol Sci 14, 8122-8147.

Chin, S. W., Shaw, J., Haberle, R., Wen, J., and Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - molecular systematics and biogeographic history of *Prunus* (Rosaceae). Mol Phylogenet Evol 76, 34-48.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A.,

Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13, 1050-1054.

Cho, M. S., Kim, C. S., Kim, S. H., Kim, T. O., Heo, K. I., Jun, J., and Kim, S. C. (2014). Molecular and morphological data reveal hybrid origin of wild *Prunus yedoensis* (Rosaceae) from Jeju Island, Korea: implications for the origin of the flowering cherry. Am J Bot 101, 1976-1986.

Cho, A., Baek, S., Kim, G.-B., Shin, C.-H., Kim, C.-S., Choi, K., Kang, Y., Yu, H.-J., Kim, J.-H., and Mun, J.-H. (2017). Genomic clues to the parental origin of the wild flowering cherry *Prunus yedoensis* var. *nudiflora* (Rosaceae). Plant Biotechnol Rep 11, 449-459.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (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, 80-92.

Claros, M. G., Bautista, R., Guerrero-Fernandez, D., Benzerki, H., Seoane, P., and Fernandez-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. Biology (Basel) 1, 439-459.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Genomes Project Analysis, G. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry A 51, 127-128.

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7, 214.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7, e47768.

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam,

D., Roos, D. S., and Stoeckert, C. J., Jr. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6, Unit 6 12 1-19.

Fujii, S., Kubo, K., and Takayama, S. (2016). Non-self- and self-recognition models in plant self-incompatibility. Nat Plants 2, 16130.

Fuji n o, K. (1900). Species of flowering cherry in Ueno Park. Jap. Hort. Mag., 92:1.

Gao, S., Bertrand, D., Chia, B. K., and Nagarajan, N. (2016). OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome Biol 17, 102.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108, 1513-1518.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-652.

Guerra-Assuncao, J. A., and Enright, A. J. (2010). MapMi: automated mapping of microRNA loci. BMC Bioinformatics 11, 133.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., and White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9, R7.

Hibrand Saint-Oyant, L., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N. N., Bourke, P. M., Daccord, N., Leus, L., Schulz, D., Van de Geest, H., Hesselink, T., Van Laere, K., Debray, K., Balzergue, S., Thouroude, T., Chastellier, A., Jeauffre, J., Voisine, L., Gaillard, S., Borm, T. J. A., Arens,

P., Voorrips, R. E., Maliepaard, C., Neu, E., Linde, M., Le Paslier, M. C., Berard, A., Bounon, R., Clotault, J., Choisne, N., Quesneville, H., Kawamura, K., Aubourg, S., Sakr, S., Smulders, M. J. M., Schijlen, E., Bucher, E., Debener, T., De Riek, J., and Foucher, F. (2018). A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nat Plants 4, 473-484.

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., Kohara, M., Watanabe, A., Kishida, Y., Fujishiro, T., Tsuruoka, H., Minami, C., Sasamoto, S., Kato, M., Nanri, K., Komaki, A., Yanagi, T., Guoxin, Q., Maeda, F., Ishikawa, M., Kuhara, S., Sato, S., Tabata, S., and Isobe, S. N. (2014). Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species. DNA Res 21, 169-181.

Hiscock, S. J., and Kues, U. (1999). Cellular and molecular mechanisms of sexual incompatibility in plants and fungi. Int Rev Cytol 193, 165-295.

Hiscock, S. J., and Tabah, D. A. (2003). The different mechanisms of sporophytic self-incompatibility. Philos Trans R Soc Lond B Biol Sci 358, 1037-1045.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32, 767-769.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57.

Jiao, W. B., and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36, 64-70.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24, 1384-1395.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., and Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46, D335-D342.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.

Kim, S.-Y., Kim, M.-H., and Kim, J. (2012). The chromosome index of Korean

native plants. National Institute of Biological Resources, Korea.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Knight, R. (1969). "Abstract bibliography of fruit breeding and genetics to 1965; *Prunus*," Commonwealth Agricultural Bureaux, Farnham Royal.

Koehne, V. E. (1912). 95 *Prunus yedoensis* var. *nudiflora*, nov. var. Von E. Koehne. Repertorium Specierum Novarum Regni Vegetabilis 10, 507.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics 5, 59.

Korlach, J. Understanding Accuracy in SMRT Sequencing [https://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf].

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68-73.

Krizek, B. A., and Eaddy, M. (2012). AINTEGUMENTA-LIKE6 regulates cellular differentiation in flowers. Plant Mol Biol 78, 199-209.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33, 1870-1874.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. Genome Biol 5, R12.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40, D1202-1210.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., and Caccamo, M. (2014). NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics 30, 566-568.

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

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., et al. (2010). The sequence and *de novo* assembly of the giant panda genome. Nature 463, 311-317.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and *de-bruijn*-graph. Brief Funct Genomics 11, 25-37.

Li, H. Z., Gao, X., Li, X. Y., Chen, Q. J., Dong, J., and Zhao, W. C. (2013). Evaluation of assembly strategies using RNA-seq data associated with grain development of wheat (*Triticum aestivum* L.). PLoS One 8, e83530.

Li, Y., Pi, M., Gao, Q., Liu, Z., and Kang, C. (2019). Updated annotation of the wild strawberry *Fragaria vesca* V4 genome. Hortic Res 6, 61.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D., Yiu, S., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T., and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

Ma, H., Olsen, R., Pooler, M., and Kramer, M. (2009). Evaluation of flowering cherry species, hybrids, and cultivars using simple sequence repeat markers. J Amer Soc Hort Sci 134, 435-444.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinformatics 20, 2878-2879.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764-770.

Martinez, M. (2011). Plant protein-coding gene families: emerging bioinformatics approaches. Trends Plant Sci 16, 558-567.

Matsumura, J.. (1901). Cerasi Japanicae duae species novae. The Botanical Magazine Tokyo 15: 99-101.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S. A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y. J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirske, W., Mooijman, P., Klein Lankhorst, R., Rose, M., Hauf, J., Kotter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysshaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Doggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M. A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blocker, H., Scharfe, M., Grimm, M., Lohnert, T. H., Dose, S., de Haan, M., Maarse, A., Schafer, M., et al. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature 402, 769-777.

McClure, B. A., and Franklin-Tong, V. (2006). Gametophytic self-incompatibility: understanding the cellular mechanisms involved in "self" pollen tube inhibition. Planta 224, 233-245.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.

Meyers, L. A., and Levin, D. A. (2006). On the abundance of polyploids in flowering plants. Evolution 60, 1198-1206.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24, 2818-2824.

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., and Soltis, D. E. (2010). Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci U S A 107, 4623-4628.

Nakai, T.. (1916). Flora Sylvatica Koreana. Vol. 5. Government of Chsen, Forestal Experiment Station, Seoul. (in Japanese).

Nakamura, N., Hirakawa, H., Sato, S., Otagaki, S., Matsumoto, S., Tabata, S., and Tanaka, Y. (2018). Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. DNA Res 25, 113-121.

Nakamura, T., Yamada, K. D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34, 2490-2492.

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933-2935.

Noh, B., Lee, S. H., Kim, H. J., Yi, G., Shin, E. A., Lee, M., Jung, K. J., Doyle, M. R., Amasino, R. M., and Noh, Y. S. (2004). Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of *Arabidopsis* flowering time. Plant Cell 16, 2601-2613.

Picard [http://broadinstitute.github.io/picard].

Potter, D., Eriksson, T., Evans, R., Oh, S., Smedmark, J., Morgan, D., Kerr, M., Robertson, K., Arsenault, M., Dickinson, T., and Campbell, C. (2007). Phylogeny and classification of Rosaceae. Pl Syst Evol 266, 5-43.

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-1358.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43, D974-981.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for

comparing genomic features. Bioinformatics 26, 841-842.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol 67, 901-904.

Rambaut, A. FigTree [http://tree.bio.ed.ac.uk/software/figtree].

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., Carrere, S., Caissard, J. C., Couloux, A., Cottret, L., Aury, J. M., Szecsi, J., Latrasse, D., Madoui, M. A., Francois, L., Fu, X., Yang, S. H., Dubois, A., Piola, F., Larrieu, A., Perez, M., Labadie, K., Perrier, L., Govetto, B., Labrousse, Y., Villand, P., Bardoux, C., Boltz, V., Lopez-Roques, C., Heitzler, P., Vernoux, T., Vandenbussche, M., Quesneville, H., Boualem, A., Bendahmane, A., Liu, C., Le Bris, M., Salse, J., Baudino, S., Benhamed, M., Wincker, P., and Bendahmane, M. (2018). The *Rosa* genome provides new insights into the domestication of modern roses. Nat Genet 50, 772-777.

Roh, M., Cheong, E., Choi, I.-Y., and Young, Y. (2007). Characterization of wild *Prunus yedoensis* analyzed by inter-simple sequence repeat and chloroplast DNA. Sci Hortic 114, 121-128.

Sassa, H., Kakui, H., and Minamikawa, M. (2010). Pollen-expressed F-box gene family and mechanism of S-RNase-based gametophytic self-incompatibility (GSI) in Rosaceae. Sex Plant Reprod 23, 39-43.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J 8, 289-317.

Seo, E., Lee, H., Jeon, J., Park, H., Kim, J., Noh, Y. S., and Lee, I. (2009). Crosstalk between cold response and flowering in *Arabidopsis* is mediated through the flowering-time gene SOC1 and its upstream negative regulator FLC. Plant Cell 21, 3185-3197.

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H., and Isobe, S. (2017). The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. DNA Res 24, 499-508.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J. M., Rees, D. J., Williams, K. P., Holt, S. H.,

Ruiz Rojas, J. J., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T. L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Jr., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Lopez Girona, E., Zdepski, A., Wang, W., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43, 109-116.

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

Smit, A., Hubley, R, and Green, P. RepeatMasker 4.0.5 [http://www.repeatmasker.org]

Smit, A., Hubley, R. RepeatModeler 1.0.8 [http://www.repeatmasker.org].

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32, W309-312.

Sun, B., Looi, L. S., Guo, S., He, Z., Gan, E. S., Huang, J., Xu, Y., Wee, W. Y., and Ito, T. (2014). Timing mechanism dependent on cell division is invoked by Polycomb eviction in plant stem cells. Science 343, 1248559.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34, W609-612.

Takenaka, Y. (1963). The origin of the Yoshino cherry tree. J Hered 54, 207-211.

Tan, S. C., and Yiap, B. C. (2009). DNA, RNA, and protein extraction: the past and the present. J Biomed Biotechnol 2009, 574398.

The International Peach Genome Initiative (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45, 487-494.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-515.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46-53.

Ushijima, K., Sassa, H., Dandekar, A. M., Gradziel, T. M., Tao, R., and Hirano, H. (2003). Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. Plant Cell 15, 771-781.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal, R., A, Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V., King, S., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A., Bus, V., Chagné, D., Crowhurst, R., Gleave, A., Lavezzo, E., Fawcett, J., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R., Durel, C., Gutin, A., Bumgarner, R., Gardiner, S., Skolnick, M., Egholm, M., Van, d., Peer, Y, Salamini, F., and Viola, R. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42, 833-839.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang,

Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. Science 291, 1304-1351.

Verelst, W., Twell, D., de Folter, S., Immink, R., Saedler, H., and Munster, T. (2007). MADS-complexes regulate transcriptome dynamics during pollen maturation. Genome Biol 8, R249.

Vieira, J., Santos, R. A., Habu, T., Tao, R., and Vieira, C. P. (2008). The *Prunus* self-incompatibility locus (S locus) is seldom rearranged. J Hered 99, 657-660.

Vitting-Seerup, K., Porse, B. T., Sandelin, A., and Waage, J. (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics 15, 81.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., and Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40, e49.

Wilson, E. H.. (1916). The Cherries of Japan. Univ. Press, Cambridge, MA, USA.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859-1875.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M. A., Tao, S., Korban, S. S., Wang, H., Chen, N. J., Nishio, T., Xu, X., Cong, L., Qi, K., Huang, X., Wang, Y., Zhao, X., Wu, J., Deng, C., Gou, C., Zhou, W., Yin, H., Qin, G., Sha, Y., Tao, Y., Chen, H., Yang, Y., Song, Y., Zhan, D., Wang, J., Li, L., Dai, M., Gu, C., Wang, Y., Shi, D., Wang, X., Zhang, H., Zeng, L., Zheng, D., Wang, C., Chen, M., Wang, G., Xie, L., Sovero, V., Sha, S., Huang, W., Zhang, S., Zhang, M., Sun, J., Xu, L., Li, Y., Liu, X., Li, Q., Shen, J., Wang, J., Paull, R. E., Bennetzen, J. L., Wang, J., and Zhang, S. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23, 396-408.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35, W265-268.

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., and Chen, S. (2012). FastUniq: a fast de novo duplicates removal tool for paired short

reads. PLoS One 7, e52249.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591.

Young, N. D., Debelle, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., Mayer, K. F., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S., Mudge, J., Bharti, A. K., Murray, J. D., Naoumkina, M. A., Rosen, B., Silverstein, K. A., Tang, H., Rombauts, S., Zhao, P. X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X., Doyle, J. J., Dudez, A. M., Farmer, A. D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A. J., Green, P. J., Hallab, A., Hartog, M., Hua, A., Humphray, S. J., Jeong, D. H., Jing, Y., Jocker, A., Kenton, S. M., Kim, D. J., Klee, K., Lai, H., Lang, C., Lin, S., Macmil, S. L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E. L., Mun, J. H., Najar, F. Z., Nicholson, C., Noirot, C., O'Bleness, M., Paule, C. R., Poulain, J., Prion, F., Qin, B., Qu, C., Retzel, E. F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A. J., Sherrier, D. J., Shi, R., Sims, S., Singer, S. R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B. B., et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480, 520-524.

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., Xing, Z., Han, C., Pan, H., Zhong, X., Shi, W., Liang, X., Du, D., Sun, F., Xu, Z., Hao, R., Lv, T., Lv, Y., Zheng, Z., Sun, M., Luo, L., Cai, M., Gao, Y., Wang, J., Yin, Y., Xu, X., Cheng, T., and Wang, J. (2012). The genome of *Prunus mume*. Nat Commun 3, 1318.

부 록

약 어

bp: Base pair

CDS: Coding sequence

CTAB: Cetyl trimethylammonium bromide

DEG: Differentially expressed gene

DBG: *de bruijn* graph

FPKM: Fragments per kilobase million

Fv: *Fragaria vesca*

Gb: Gigabase

gDNA: Genomic DNA

GSI: Gametophytic self-incompatibility

HPD: Higher posterior density

InDel: Insertion or deletion

kb: Kilobase

K-mer: Substring of length K

Ks: Synonymous substitution rate

Mb: Megabase

MDS: Multidimensional scaling

ML: Maximum likelihood

MP: Mate-paired

mRNA-seq: Messenger RNA sequencing

Mt: *Medicago truncatula*

Mxd: *Malus* x *domestica*

Mya: Million years ago

NCBI: National Center for Biotechnology Information

NGS: Next-generation sequencing

OLC: Overlap layout consensus

PCR: Polymerase chain reaction

PE: Paired end

Pjj: *P. jamasakura* var. *jamasakura*

Pjq: *P. jamasakura* var. *quelpaertensis*

Ppa: *P. pendula* f. *ascendens*

Psa: *P. sargentii*

Pxn: *P.* x *nudiflora*

Pxy: *P.* x *yedoensis*

SFB: S haplotype-specific F-box protein

SNP: Single nucleotide polymorphism

SSI: Sporophytic self-incompatibility

S-RNase: S-locus ribonuclease

Ti: Transition

Tv: Transversion

Draft Genome Sequence of *Prunus* x *nudiflora*, a Natural Hybrid Flowering Cherry

Baek Seunghoon
Department of Biological Sciences
Graduate School, Myongji University
Directed by Professor Mun Jeong-hwan

Hybridization is an important evolutionary process that results in increased plant diversity. Flowering *Prunus* includes popular cherry species that are appreciated worldwide for their flowers. The ornamental characteristics were acquired both naturally and through artificially hybridizing species with heterozygous genomes. Therefore, the genome of hybrid flowering *Prunus* presents important challenges both in plant genomics and evolutionary biology.

Prunus yedoensis Matsumura is one of the popular ornamental flowering cherry trees native to northeastern Asia. The natural populations of its close relative wild taxon (*P.* x *nudiflora*, Pxn) have only been found on Jeju Island, Korea. Previous studies suggested that Pxn is a hybrid taxon and closely related to Yoshino cherry (*P.* x *yedoensis*), however, there are no solid evidences on its exact parental origin, genomic organization, and genetic boundary against its related taxa. To solve these problems, I sequenced and assembled the genome of Pxn and compared it with those of closely related *Prunus* species, including its candidate parental species and Yoshino cherry.

I used long reads to sequence and analyze the highly heterozygous genome of Pxn. The genome assembly covered more than 93% of the gene space, and annotation identified 41,294 protein-coding genes. Comparative analysis of the

genome with 16 accessions of 6 related taxa and phasing of genes based on short read sequence mapping and single nucleotide polymorphism analysis showed that 41% of the genes were assigned into the maternal or paternal state. This indicates that Pxn is an F1 hybrid originating from a cross between maternal *P. pendula* f. *ascendens* and paternal *P. jamasakura*, and it can be clearly distinguished from its confusing taxon, Yoshino cherry. A focused analysis of the S-locus haplotypes of closely related taxa distributed in a sympatric natural habitat suggested that reduced restriction of interspecific hybridization due to strong gametophytic self-incompatibility is likely to promote complex hybridization of wild *Prunus* species and the development of a hybrid swarm.

In conclusion, I report the draft genome assembly of a natural hybrid *Prunus* species using long-read sequencing and sequence phasing. Based on a comprehensive comparative genome analysis with related taxa, it appears that cross-species hybridization in sympatric habitats is an ongoing process that facilitates the diversification of flowering *Prunus*. Overall, this study makes a significant contribution to address issues of the origin, taxonomic delimitation, nomenclature, and genetic relationship of Pxn with other *Prunus* species.

Keyword

Prunus x *nudiflora*, hybrid genome, long-read sequencing, sequence phase, S-locus haplotype, genome assembly, Rosaceae