염색체 도식화와 imputation에 의한 GBS 기반 여교잡 회복률 계산 정확도 증진 방법

정혜리¹ · 최준경¹ · 이봉우¹ · 이보미¹ · 강윤주¹ · 이정희¹ · 김지은¹ · 남문² · 박영훈³ · 박민우⁴ · 박기림³ · 조성환^{1*}

¹주식회사 씨더스, ²지노타입 주식회사, ³부산대학교, ⁴농업회사법인 현대종묘(주)

Method for Improving Marker Selection Efficiency through a Graphical Representation of Molecular Markers

Hye-ri Jeong¹, Junkyuoung Choe¹, Bong-Woo Lee¹, Bo-Mi Lee¹, Yun-Joo Kang¹, Jeong-Hee Lee¹, Ji-Eun Kim¹, Moon Nam², Young-Hoon Park³, Minwoo Park⁴, Girim Park³, and Sung-Hwan Jo^{1*}

¹SEEDERS Inc., Daejeon, 34912, Republic of Korea

²Xenotype Co., LTD, Daejeon, 34912, Republic of Korea

³Dept. Horticultural Bioscience, Pusan National University, Miryang, 50464, Republic of Korea

⁴Hyundai Seed Co., LTD., Yeosu, 12660, Republic of Korea

Abstract Marker-assisted backcrossing is a powerful method for developing new cultivars. To develop genomic-wide markers, genotyping-by-sequencing (GBS) can be an efficient method. However, unrefined low-quality markers and missing data between markers can contribute to hampering the marker selection process, particularly in multi-way crosses. In this study, we aimed to calculate the recovery rate of offspring individuals and minimize errors that occur among a large number of markers. Initially, missing data were imputed by comparing samples using the *k*-nearest neighbor (*k*-NN) algorithm. Thereafter, low-quality single-nucleotide polymorphisms (SNPs) were corrected by applying the graphical representation method based on the *k*-NN algorithm in order of the SNPs in a chromosome designed for a multi-parental population. Four-way cross and double-backcrossed tomato BC₁F₁ (230 lines) and BC₂F₁ (96 lines) populations were genotyped by GBS. The genotype of samples of the BC₁F₁ and BC₂F₁ populations was determined based on the parental haplotype. Thus, the method of visualizing the genotype of offspring individuals, generated via crosses of multiple parents, not only improves estimation of the recovery rate but also facilitates easier selection in breeding programs.

Keywords tomato, a graphical representation, marker selection, marker-assisted backcrossing, marker-assisted selection

Received on September 22, 2020. Revised on September 28, 2020. Accepted on October 26, 2020.

* Corresponding Author (E-mail: shjo@seeders.co.kr, Tel: +82-42-710-4035, Fax: +82-42-710-4036)

서 언

분자마커를 활용하는 분자육종은 육종프로그램 가운데 중요성이 지속적으로 증가하고 있다(Edwards & Batley 2010). 단일유전자 혹은 주동유전자에 의해 형질이 결정되는 경우 형질에 연관된 분자마커를 이용하여 대상 개체를 선발하는 기술(Marker-assisted Selection, MAS) 뿐 만 아니라 여러 유전자에 의해 결정되는 양적형질을 선발하는 분자마커 개발도 최근 유전체 선발(genomic selection) 등으로 빠르게 발전하고 있다(Collard & Mackill 2008). 특히, 여교배 개체의 선발을 위해서 다수의 마커를 활용하여 회복친의 회복률을 마커 기반으로 탐색하고 선발하는 기술(Marker-assisted backcrossing, MAB) 또한

분자육종에 널리 활용되고 있다(Kim et al. 2013). 이와 같은 분자육종 프로그램의 활성화는 차세대염기서열 해독 장비(Next-Generation DNA Sequencing technologies)의 빠른 발달로 게놈 전반에 존재하는 SNP를 비교적 신속하고 저렴하게 확보하게 되었기 때문에 더욱 촉진되고 있다(Hyten et al. 2010). 특히 Genotyping-by-sequencing (GBS)은 유전체 전체를 대상으로 하는 NGS를 활용하면서 제한효소로 절단된 주변 서열만 부분적으로 해독함으로써 데이터 생산량을 적게 하여 유전체의 복잡도는 낮추면서도 SNP 당 read depth를 확보하고 유전체 전체를 확인할 수 있는 유전적 변이를 집단수준에서 탐색 가능한 기술이다(Elshire et al. 2011). GBS는 수집된 유전자원 집단이나 교배에 의해 작성된 교배집단에서 빠르게 분자마커를 개발하

나 유전자원집단의 구조적 특징, 유전지도 작성 등의 연구를 위해 폭넓게 시도되고 있다(Crossa et al. 2013). GBS 데이터는 multiplexing을 통해 수 십 개 혹은 수백개의 시료를 동시에 분석하면서 수 천개 이상의 SNP를 확보할 수 있는 장점을 가지고 있다. 그러나 염기서열이 확보되지 않는 결측 좌(missing locus)가 다수 발생하거나 시퀀싱양이 낮은 커버리지(coverage)를 갖는 좌가 발생하여 genotyping의 정확도를 낮추는 원인을 제공할 수 있다(Crossa et al. 2013).

결측 좌를 해결하기 위해 imputation 기술이 다양하게 개발되어 왔다. 결측 값 대체를 통한 genotyping 결과의 품질적인 개선은 genomic selection과 계통 전체 연관 연구의 결과를 개선하는데 효율적으로 적용되었다(Mao et al. 2016). 결측 좌를 예측하기 위한 방법으로는 크게 generic imputation method와 genotype-specific method를 사용하였다. Generic imputation method를 적용한 kNN이나 LD-kNN 프로그램은 k -nearest neighbors imputation 혹은 linkage disequilibrium k -nearest neighbors imputation 알고리즘을 적용하여 개발하였다(Troyanskaya et al. 2001, Money et al. 2015). Genotype-specific method를 적용하기 위해서는 phased haplotype information이 필요하며 BEAGLE (Pook et al. 2020), fastPHASE (Scheet & Stephens. 2006)와 같은 프로그램이 개발되었다(Browning & Browning 2007). Genotype-specific method의 경우 연구가 많이 진행된 모델 생물을 중심으로 개발되었고 다배체 생물이나 비모델 생물의 경우는 generic imputation method를 적용하였다(Su et al. 2008, Money et al. 2015).

결측 좌뿐만 아니라 genotyping error는 회복친의 회복률을 계산할 경우 문제를 야기할 수 있으나 현재까지는 low quality sequence를 필터링하거나 서열 생산량을 증가시키는 방법 이외에 특별히 개발된 사례는 매우 제한적이다. 이러한 문제점을 해결하기 위하여 본 연구에서는 recombination pattern and frequency를 기반으로 k -NN 알고리즘을 적용하여 genotype error를 개선하여 최소화하고 염색체 도식화(genotype graphical representation) 기술을 통해 결과를 검증하였다. 재조합 패턴과 빈도(recombination pattern and frequency)를 토마토 RILs (recombinant inbred lines)에서 조사한 결과에 의하면, 60개의 RILs 계통에서 1,445 recombination site를 발견했다고 보고하였다(de Haas et al. 2017). 이는 계통당 24번의 recombination site가 발견된 것이며 염색체 수준에서는 2번 발견되는 것을 의미한다. 교배의 횟수에 따라 recombination 발생 빈도는 바뀔 수 있을 것으로 예측해볼 때, RILs에 비해 교배가 줄어드는

F_2 집단이나 BC_1 집단 등에서는 보다 적은 빈도의 recombination 이 발생될 수 있어 GBS 등의 genotyping 결과 해석과 genotyping error 교정에 참조할 수 있다. 염색체 도식화 기술은 마커 기반의 개체 선발을 돕기 위해 개발된 가시화 도구이다. 즉 계통 전체의 구성을 그래픽적으로 표현함으로써 새로 육성되는 계통의 유전적 구성을 직관적으로 판단하게 하는 방법론이다(van Berloo R 2008). 본 연구의 목적은 다량의 마커를 활용하는 경우 빈번히 발생하는 데이터 결측과 genotype error를 효과적으로 개선함으로써 복교잡을 실시하는 여교잡 육종에서 개체 선발 효율을 증진하고자 하였다.

재료 및 방법

공시재료

연구 소재로는 Fig. 1의 복교배와 여교배 과정을 통해 육종회사(현대종묘, 여주)로부터 작성된 BC_1F_1 , BC_2F_2 집단과 이들 집단 작성에 활용된 부모친본 4계통(TGA2, SV17, MS15, HA1)

Fig. 1. The breeding procedure with the four-way cross method. Four founder parents (TGA2, SV17, MS15, HA1) were used. After four-way crossing, two back crosses with recurrent parent and one selfing were performed. MAS and MAB selection were done at each step.

을 사용하였다(Fig. 1). 부모친 4계통은 모두 무한생육형의 분홍색 대과종 토마토로써 HA1, TGA2 그리고 MS15는 일반적인 동양계 분홍색 대과종 토마토인 반면, SV17은 동양계 분홍토마토와 유럽계 레드토마토의 교잡으로 육성된 고정도 계통이다. HA1과 SV17은 복합내병계이며, TGA2는 단위결과성, MS15는 응성불임성 계통이다. 확보된 부모친 계통 및 BC₁F₁, BC₂F₁ 집단에 대해 각각 230개, 96개 개체로부터 어린잎을 채취하고 핵 DNA를 추출하여 GBS 분석에 사용하였다. BC₁F₁ 두 집단(SV17-BC₁F₁, HA1-BC₁F₁)은 복교배 후대(DF₁) 중 한 개체를 선발하여 두 회복친(SV17, HA1)에 교배하여 얻었으며, BC₂F₁ 두 집단(SV17-BC₂F₁, HA1-BC₂F₁)은 각 BC₁F₁ 집단에서 선발된 한 개체씩을 각 회복친에 다시 여교배하여 얻었다. BC₂F₁ 두 집단(SV17-BC₂F₁, HA1-BC₂F₁)은 각 BC₂F₁ 집단에서 선발된 후대 개체들을 자가수정하여 얻었다.

GBS 라이브러리 작성 및 시퀀싱

GBS 라이브러리(library)는 Elshire et al. (2011)에서 수정된 프로토콜을 사용하여 제한 효소 *ApeKI* (GCWGC)를 사용하여 제작하였다. DNA 샘플(100 ng/μl)을 각각 다른 바코드 어댑터를 갖는 96개의 DNA 샘플을 하나의 실험 세트로 pooling하였다. PCR 과정을 통하여 200-500 bp의 fragments가 enrichment되도록 시퀀싱용 GBS 라이브러리를 제작하였다. NGS sequencing은 HiSeq 2500 PE101 (Illumina, CA, 미국)을 이용하여 library 당 1 lane (50 Gbp 이상)을 생산하여, 미국 총 2 library, 2 lane (총 100 Gbp 이상) 생산하였다.

SNP 탐색 및 매트릭스 작성

생산된 염기서열은 barcode sequence를 이용하여 demultiplexing을 수행하고, adapter sequence 제거 및 sequence quality trimming을 수행하였다. Adapter trimming은 cutadapt (version 1.8.3) 프로그램을 사용하고, sequence quality trimming은 SolexaQA (version 1.13) package의 DynamicTrim과 LengthSort 프로그램을 사용하였다(Cox et al. 2010). DynamicTrim의 phred score ≥ 20 , LengthSort의 short read length ≥ 25 bp 옵션을 적용하였다. 전처리 과정을 통과한 cleaned reads를 BWA (0.6.1-r104) 프로그램을 사용하여 토마토 표준유전체(*Solanum lycopersicum* SL3.0)에 mapping을 수행하였다(Li & Durbin. 2009). Mapping 조건은 seed length (-l) = 30, maximum differences in the seed (-k) = 1, number of threads (-t) = 16, mismatch penalty (-M) = 6, gap open penalty (-O) = 15, gap extension

penalty (-E) = 8을 사용하였고 나머지 옵션은 기본값을 사용하였다. SNP를 탐색하는 조건은 minimum mapping quality for SNPs (-Q) = 30, minimum mapping quality for gaps (-q) = 15, minimum read depth (-d) = 3, maximum read depth (-D) = 202, min indel score for nearby SNP filtering (-G) = 30, SNP within INT bp around a gap to be filtered (-w) = 15, window size for filtering을 만족하도록 하였다. 분석대상 간의 SNP 비교분석을 수행하기 위해 샘플간 통합 SNP 매트릭스(matrix)를 작성하였다. 매트릭스 내 SNP 필터 조건은 SNP 좌가 biallelic 유형이며, 최소 read depth 3개 이상을 만족하고, minor allele frequency (MAF)가 5% 이상을 만족하고, SNP 좌에서 missing data가 30% 이내를 만족하는 SNP만을 선발하였다. 각 SNP 좌는 aligned read의 구성에 기반하여 3가지 SNP 유형으로 구분하였다. Read depth $\geq 90\%$ 이상 동일한 경우 homozygous-type SNP, $40\% \leq \text{read depth} \leq 60\%$ 인 경우 heterozygous-type SNP, 두 유형으로 구분되지 않는 나머지 경우는 기타 SNP로 구분하였다.

Haplotype 기반의 genotyping

BC₁F₁, BC₂F₁ 집단의 각 샘플의 유전자형은 부모의 haplotype을 이용하여 수행하였다. 부모 샘플의 유전자형은 'TGA2=A, SV17=B, MS15=C, HA1=D'로 정의하고, BC₂F₁ 집단이 heterozygous type일 경우 'H'로 정의하였다. 각 샘플마다 염색체 별 SNP를 position 순으로 정렬하고, haplotype 분석을 위한 window size는 각 SNP 좌마다 앞뒤로 10개씩 총 21개의 SNP를 하나의 단위로 지정하여 부모 4개 샘플(TGA2, SV17, MS15, HA1)의 염기서열과 비교하여 해당 SNP 좌를 가장 가까운 부모의 유전자형(ABCD) 또는 'H' 유전자형으로 결정하였다. 만약 가장 가까운 부모의 유전자형으로 결정하지 못하는 경우, 앞뒤로 10개씩의 SNP를 계속 확장해 가면서 해당 SNP 좌를 가장 가까운 부모의 유전자형(A, B, C, D) 또는 'H' 유전자형으로 결정하였다.

Imputation과 genotype error correction

Imputation은 근접이웃(k -nearest neighbors, k -NN) 알고리즘을 적용하였다(Laaksonen & Oja 1996, Money et al. 2015). Missing data를 채우기 위해서 두 단계를 적용하였는데, 첫째 가장 근접이웃을 샘플들 가운데서 찾고 두번째로 해당부위의 genotype을 추정하는 단계로 진행하였다. Genotype error correction을 위해서도 k -NN 알고리즘을 적용하였으나 근접이웃을 염색체 위치 상에서의 거리로 측정하여 적용하였다. 가장 근접한 샘플의 거리를 측정하는 수식은 아래와 같다.

$$d_n(s_1, s_2) = \frac{1}{n} \sum_{p \in P} |g(s_1, p) - g(s_2, p)|$$

거리측정 $d_n(s_1, s_2)$ 은 시료 s_1 과 s_2 와의 거리이며, P 는 모든 SNP 세트를 뜻하고 $g(s, p)$ 는 s 샘플의 p 위치를 의미한다. N 은 SNP 수를 의미한다.

Imputation 할 genotype $g(s_i, s_j)$ 을 수식으로 표현하면 아래와 같다.

$$g_i(s_i, s_j) = \arg \max_{a \in \{0, 1, 2\}} \sum_{s \in N} \frac{1}{d_n(s_i, s)} I(g(s, P_j) = a)$$

N 은 s_i 샘플과 가장 가까운 k 샘플의 세트를 의미한다. 이때 p_j 의 SNP는 알고 있고 있다. $I[g(s, p_j) = a]$ 는 지시 함수를 의미한다. 즉 $g(s, p_j) = a$ 이면 1을 취하고 그렇지 않으면 0을 취한다.

결과 및 고찰

GBS 분석

여교배 집단의 회복된 염색체 회복률을 조사하기 위하여 BC₁F₁ 230개체와 BC₂F₁ 96개체에 대해 GBS 분석을 실시하였다. BC₁F₁의 경우 최대한 많은 개체를 검토하여 회복률이 높은 개체를 선발하는 것이 필요하다고 판단하여 multiplex index를 증가하여 첫번째 pooling은 총 112개체를 포함하여 55,953,617,210 bp를 확보하였고 두번째 pooling은 118개체를 포함하여 55,184,871,466 bp의 서열을 확보하였다. 총 230개체의 BC₁F₁ 자손에서 111,138,488,676 bp를 확보하여 분석하였다. 첫번째 pool의 결과를 살펴보면 샘플 당 평균 raw reads의 수는 3,565,014개가 확보되었다. Adaptor와 low quality sequence를 제거하면 read의 평균 길이는 83.5 bp를 확보하였고 이를 참조유전체에 mapping하였을 때 85.78%의 reads가 mapping되었다. GBS 분석이 얼마나 많은 영역을 조사하는 지를 확인하기 위하여 mapping 영역을 조사하였을 때 평균 83,832 mapping 영역이 조사되었다. Mapping 영역에서 조사되는 read depth는 약 10.98로 조사되었다. Read depth per SNP는 homozygous-type SNP 또는 heterozygous-type SNP를 구분하거나 genotype의 정확도를 결정하는 지표로 사용하였다. 두번째 pool의 결과를 살펴보면 샘플 당 평균 raw reads의 수는 4,098,496개가 확보되어 첫번째 pool에 비해 샘플수가 많음에도 불구하고 샘플 별 확보된 raw

read 수가 증가된 것을 확인하였다. Adaptor와 low quality sequence를 제거하면 read의 평균 길이는 80.7 bp를 확보하였고 이를 참조 유전체에 mapping하였을 때 84.02%의 reads가 mapping되어 첫번째 pool에 비해 소폭 감소하였다. Mapping 영역을 조사하였을 때 평균 48,491 mapping 영역이 조사되어 첫번째 pool에 비해 상당히 적은 영역에서 sequencing 되었음을 확인하였다. Mapping 영역에서 조사되는 read depth는 약 15.98로 첫번째 pool보다 높게 조사되었는데 이는 mapping 영역이 줄어들면서 read depth가 증가된 것으로 판단하였다. BC₂F₁의 경우 총 96개의 샘플(SV17-BC₂F₁ 71개, HA1-BC₂F₁ 25개)을 1개의 pool로 만들어 GBS를 실시하였다. 총 96개체의 BC₂F₁ 자손에서 106,989,031,432 bp의 염기서열을 생산하였다. 샘플 당 평균 raw reads의 수는 6,928,803개가 확보되어 BC₁F₁에 비해 2배 정도 증가하였는데 이는 샘플의 수는 절반 이하로 줄었지만 생산된 염기서열의 총량은 유사했기 때문이다. Adaptor와 low quality sequence를 제거하면 read의 평균길이는 116.59 bp로 증가되었다. 반면 이를 참조유전체에 mapping하였을 때 76.01%의 reads가 mapping되어 매핑률(mapping rate)이 감소되었다. Mapping 영역을 조사하였을 때 평균 70,261 mapping 영역이 조사되었다. Mapping 영역에서 조사되는 read depth는 24.37로 가장 높게 나타났다.

GBS 분석 데이터 기반의 SNP 선별

BC₁F₁ 분석에 사용한 230개 샘플을 하나로 통합하여 SNP를 탐색하였을 때 총 223,998개 SNPs가 확보되었다. MAF 5% 이하인 SNP를 제거하면 33,350개 SNPs 좌가 확인되었고, SNP 좌 별로 30% 이상의 샘플에서 결측(missing data)이 일어난 좌를 제거했을 때는 42,572 SNPs 좌가 확인되었다. MAF와 결측치 기준을 동시에 적용했을 때는 2,231개 좌가 확보되었다. GBS에서 일반적으로 사용하던 96개 단위의 multiplex index에 비해 개체별로 시퀀싱 원 데이터(raw reads) 생산량이 고르지 않고 불균형적으로 생산된 것을 확인하였다. Raw read의 생산량이 1,000,000개 이하의 샘플은 첫번째 pool에서 16개, 두번째 pool에서 32개 샘플로 확인되었고, 1,000,000개 reads 이하로 생산된 48개 샘플을 제외한 182개 샘플을 이용하여 SNP를 선발하였을 때 총 219,812 SNPs가 확보되었다(Table 1). MAF 5% 이하 SNP를 제거하면 33,132 SNPs가 확인되었고, SNP 좌별로 30% 이상의 샘플에서 결측이 일어난 좌를 제거했을 때 64,124 좌가 확인되었다. MAF와 결측치 기준을 동시에 적용했을 때는 3,897개 SNPs 좌가 확보되었다. SNP 좌의 수를 더 확보하기

Table 1. Statistics of number of markers after filtration from GBS results of BC₁F₁ and BC₂F₁ tomato lines.

Chromosome	BC ₁ F ₁ population (182ea) before filtered		BC ₁ F ₁ population (182ea) after filtered		BC ₂ F ₁ population (94ea) before filtered		BC ₂ F ₁ population (94ea) after filtered	
	No. of SNPs	Avg. missing rate	No. of SNPs	Avg. missing rate	No. of SNPs	Avg. missing rate	No. of SNPs	Avg. missing rate
SL3.0ch00	26,054	37.74%	1,373	14.27%	9,095	13.46%	829	7.20%
SL3.0ch01	21,218	49.64%	181	18.07%	6,146	12.99%	277	10.26%
SL3.0ch02	16,553	49.92%	115	19.58%	4,281	16.48%	101	10.10%
SL3.0ch03	17,317	47.57%	180	15.56%	5,209	12.81%	243	9.89%
SL3.0ch04	15,624	52.83%	238	17.58%	4,157	12.36%	373	10.64%
SL3.0ch05	11,607	51.24%	91	20.37%	3,421	11.21%	211	11.02%
SL3.0ch06	21,208	50.49%	669	17.45%	5,979	10.16%	1,027	9.05%
SL3.0ch07	12,953	50.04%	118	18.99%	4,074	12.86%	183	11.03%
SL3.0ch08	13,199	48.09%	85	18.47%	3,651	16.35%	97	10.70%
SL3.0ch09	21,821	68.65%	366	18.23%	3,121	10.73%	665	10.54%
SL3.0ch10	11,847	52.15%	116	19.66%	3,575	12.91%	220	10.99%
SL3.0ch11	16,602	52.68%	260	1.10%	4,460	12.67%	335	10.19%
SL3.0ch12	13,809	55.08%	268	17.34%	4,045	10.58%	345	10.80%
Total	219,812	51.24%	4,060	16.67%	61,214	12.74%	4,906	10.19%

위하여 MAF 5% 이하, 결측치 50% 이하 조건으로 완화하여 총 4,060 SNPs 좌를 확보하였다(Table 1).

BC₂F₁ 분석에 사용한 96개 샘플을 이용하여 SNP를 분석하였을 때는 총 110,961 SNPs를 확보하였다. Raw reads 생산량이 1,000,000 reads 이하인 샘플이 2개 샘플로 확인되어 분석에서 제외하고 94개의 샘플을 분석하였을 때는 총 61,214개 SNPs가 확보되었다(Table 1). MAF 5% 이하이면서 SNP 좌별로 30% 이상의 샘플에서 결측이 일어난 좌를 제거하여 총 4,906 개의 SNPs가 확보하여 이후 분석에 사용하였다(Table 1).

Haplotype 기반의 genotyping and correction

본 실험에서 사용한 SNP는 biallelic type을 사용하였기 때문에 4개 부모의 유전형을 구분할 수 없다. 따라서 20개의 SNP를 haplotype block으로 묶어서 4개의 부모간 유전형을 분석하고자 하였다(Fig. 2A). 유래부모의 결정은 *k*-NN 알고리즘을 적용하여 4개의 부모와 비교하여 평균값을 적용하여 적정 유래부모를 판정하였다. 결과는 기존에 유전자 기반 haplotype 분석과 동일하게 2개의 유전형으로 구분되는 것을 확인하였다(Jeong et al. 2020). 유형으로 나누어 보면, 두개의 부모씩 유전형이 동일한 경우 즉, A와 B의 유전형이 같고 C와 D 유전형이 같은 경우, 혹은 3개의 부모 유전형은 같고 나머지 1개의 부모 유전형이 다른 경우, 예를 들어 A, B, C 가 동일한 유전형을 가지고 있고

Fig. 2. Snap shot of manual curation and enlarged graphical genotyping results (A). BC₂F₁ which is backcross with SV17 (B). Two lines backcross with HA1 (C). ①: sample name, ②: nucleotide sequence from the sample, ③: nucleotide sequence from the parent TGA2, ④: nucleotide sequence from the parent SV17, ⑤: nucleotide sequence from the parent MS15, ⑥: nucleotide sequence from the parent HA1, ⑦: graphical representation, ⑧: position of the SNP (Mbp). The X axis represents lines and the Y axis represents markers along chromosome.

D는 다른 유전형을 가진 유형으로 구분되었다. 따라서 이러한 문제는 회복친에 가중치를 주어 회복률을 계산하도록 하였다. 예를 들어 SV17를 이용해 여교잡을 실시하는 T4 line은 SNP nucleotide를 염색체별 위치별로 정렬 후 4개의 부모친의 해당 좌 nucleotide와 비교하였다(Fig. 2B). 그 결과 T4 line의 chromosome 10을 구성하도록 하는 유래 부모는 SV17와 MS15가 가능한데, 이 중 SV17로부터 유래되었다고 판단하였다. 이는 *k*-NN 알고리즘에 가장 적합하기도 하고 회복친으로 사용된 SV17을 제공친으로 우선권을 부여하였기 때문이다. T5 line의 경우 T5의 SNP가 heterozygous type이기 때문에 H (green color)로 선정하고 *k*-NN 알고리즘에 따라 heterozygous type으로 결정하였다. 이와 마찬가지로 B19와 B20 line은 HA1를 회복친으로 사용하는 HA1-BC₂F₁ 집단에 속한다(Fig. 2C). Haplotype 방법이나 *k*-NN 알고리즘에 의한 유래친을 판단해 보면 TGA2와 HA1은 해당 부위가 동일한 유전형을 가지고 있는 것으로 관찰된다. 그러나 *k*-NN 평균값과 HA1이 회복친인 것을 적용하여 B19의 chromosome 10의 해당부위는 HA1으로부터 유래되었다고 판단하였다. B20 line의 경우 4개의 불일치 되는 nucleotide가 존재함에도 불구하고 *k*-NN 알고리즘에 의해 heterozygous type으로 판정된 사례이다. 이와 같이 *k*-NN 알고리즘에 의해 유래부모를 판정하고 확인할 수 있는 manual curation tool을 통해 판정의 적합성을 재확인하였다.

Graphical representation 방법에 의한 chromosome visualization

여교잡 집단의 각 계통은 부모의 유전자형에 따라 염색체의 순서를 따라 haplotype에 의한 genotype을 결정하였다. 부모 샘플의 유전자형은 TGA2=A (red), SV17=B (blue), MS15=C (yellow), HA1=D (black)으로 정의하고, heterozygous type일 경우 'H'(green color)로 설정하였다. 염색체별 recombination 빈도는 기존 보고에서 염색체당 2회임을 고려하여 5회 이내로 발생할 것으로 추정하였다(de Haas et al. 2017). 20 SNPs 단위의 haplotype을 염색체의 순서에 따라 이동하면서 부모 유전형을 추정하고 두번째 단계로 *k*-NN 알고리즘에 의하여 산발적으로 발생하는 genotyping error를 교정하였다. 즉 20개의 SNPs를 각각 4개의 부모와 비교하여 동일한 genotype을 갖는 부모를 결정한 후 평균 이상의 genotype을 갖는 부모로 추정하였다. 평균에 미치지 못하는 경우는 결정하지 않고 다음 부위로 이동하여 동일하게 유래부모를 예측하였다. 이전의 유전형과 동일한 부모인 경우는 결정하지 못한 부위를 전후의 부모형으로 결정하였다. 만약에

다른 부모형으로 결정이 되면 recombination이 발생하는 것으로 결정하였다. 염색체별 재조합 패턴과 빈도를 살펴보면 염색체별 고유의 특징이 있는 것을 확인할 수 있다. BC₂F₁ 집단을 보면 5번 염색체와 8번 염색체는 전체 개체에서 회복률이 상당히 높게 나타나고 있는 반면 3번 염색체나 6번 염색체는 회복률이 비교적 낮게 나타나고 있다(Fig. 3) Recombination이 발생하는 부위는 비교적 선호되는 영역이 존재하는 것으로 보여 기존의 보고와 일치된 결과를 보였다(de Haas et al. 2017). 목표형질이 존재하는 부위를 알고 있는 경우 염색체별 선발결과를 확인하는 것과 개체의 12개 염색체를 모두를 확인하였다(Fig. 4).

회복률 계산

BC₁F₁과 BC₂F₁의 회복친 회복률과 heterozygous type 영역을 계산하였다(Fig. 5). SV17를 회복친으로 하는 SV17-BC₁F₁ 집단에서 가장 낮은 회복률을 보인 계통은 S41로 55.25%의 회복률을 보였고, 가장 높은 계통은 S113으로 93.49%의 회복률을 보였다. BC₁F₁ 집단의 평균 회복률은 79.1%였다. HA1을 회복친으로 하는 HA1-BC₁F₁의 경우는 최저 51.12%에서 최고

Fig. 3. Graphical representation of 94 BC₂F₁ individuals by chromosome. Homozygous genotype from SV17 (blue color), TGA201 (red color), MS15 (yellow color), HA1 (black color), heterozygous genotype (green color). The X axis represents lines and the Y axis represents chromosome position.

Fig. 4. Examples of graphical representation of three BC_2F_1 individuals. Genotype from SV17 (blue color), TGA201 (red color), MS15 (yellow color), HA1 (black color), heterozygous genotype (green color) represented by each color. Each line represents twelve chromosomes.

Fig. 5. Recovery rate in BC_1F_1 (A, B) and BC_2F_1 (C, D) population. Lines with high recovery rates become candidates for selection.

95.92%의 회복률을 보였다. Heterozygous type의 비율을 보면 SV17- BC_2F_1 의 경우 4.21%에서 34.86% 범위로 존재하였다. HA1- BC_2F_1 의 경우는 3.76%에서 27.16%의 heterozygous type으로 구성되었다. BC_2F_1 집단의 평균 회복률은 80.2%였다. 이러한 결과를 통해 BC_1F_1 과 BC_2F_1 단계 모두에서 5% 이내의 heterozygous type를 가진 계통을 선발할 수 있음을 의미한다.

적 요

분자유종이 연구의 단계를 넘어 실용화 단계로 접어들면서 분석과정 중에 발생하는 다양한 문제를 빠르게 해결하는 방법을

개발하는 일이 더욱 더 필요하게 되었다. 본 연구를 통해 GBS 분석 기반으로 확보된 데이터 내 결측치와 genotype error를 해결하고 회복친의 회복률의 계산의 정확도를 높이기 위한 방법을 생물정보학적으로 시도하였다. 또한 염색체 가시화 기술은 분자마커를 활용할 때 사용자가 보다 직관적으로 결과를 쉽게 이해하게 하는 중요한 도구가 될 것이다. 특히 현장 육종 프로그램에서는 다수의 부모친을 사용하는 경우가 빈번히 발생하게 되는데 본 연구를 통해 haplotype을 통해 다수의 부모를 교배친으로 사용한 경우 그리고 부모친들 간 동일한 유전적 구성을 가진 경우 도식화하는 방안을 개발함으로써 현장의 육종 프로그램에 쉽게 적용할 수 있을 것으로 기대한다.

사 사

본 연구는 차세대바이오그린21사업(농생물계농활용연구사업단 과제번호: PJ01313203)의 “원예작물의 유전체 육종 구현을 위한 데이터베이스 구축” 과제의 지원에 의해 이루어진 것이다.

REFERENCES

- Browning SR, Browning BL. 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.
- Collard BC, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 363: 557-572.
- Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.
- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, Campos G. de los, Burgueño J, Windhausen VS, Buckler ES, Jannink JL, Lopez Cruz MA, Babu R. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3-Genes Genom Genet* 3: 1903-1926.
- de Haas LS, Koopmans R, Lelivelt CLC, Ursem R, Dirks R, Velikkakam James G. 2017. Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs. *DNA Res* 24: 549-558.
- Edwards D, Batley J. 2010. Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J* 8: 2-9.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.
- Hyten DL, Song Q, Fickus EW, Quigley CV, Lim JS, Choi IY, Eun-Young Hwang, Pastor-Corrales M, Cregan PB. 2010. High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11: 475.
- Jeong HR, Lee BM, Lee BW, Oh JE, Lee JH, Kim JE, Jo SH. 2020. Tag-SNP selection and web database construction for haplotype-based marker development in tomato. *J Plant Biotechnol* 47: 218-226.
- Kim JE, Lee BW, Kim SM, Lee BM, Lee JH, Jo SH. 2013. Genome-wide SNP database for marker-assisted background selection in tomato. *Korean J Breed Sci* 45: 232-239.
- Laaksonen J, Oja E. 1996. Classification with learning k-nearest neighbors. pp. 1480-1483. *Proc Intl Conf Neural Networks (ICNN'96)*. Washington, dc, USA.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.
- Mao X, Sahana G, Koning DJ, Guldbbrandtsen B. 2016. Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data. *J Anim Sci* 94: 1426-1437.
- Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong GY, Myles S. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3- Genes Genom Genet* 5: 2383-2390.
- Pook T, Mayer M, Geibel J, Weigend S, Cavero D, Schoen CC, Simianer H. 2020. Improving imputation quality in BEAGLE for crop and livestock data. *G3- Genes Genom Genet* 10: 177.
- Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78: 629-644.
- Su SY, White J, Balding DJ, Coin LJM. 2008. Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* 9: 513-513.
- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17: 520-525.
- van Berloo R. 2008. GGT 2.0: versatile software for visualization and analysis of genetic data. *J Hered* 99: 232-236.