유전자 단위 haplotype을 대변하는 토마토 Tag-SNP 선발 및 웹 데이터베이스 구축

정혜리 · 이보미 · 이봉우 · 오재은 · 이정희 · 김지은 · 조성환

Tag-SNP selection and online database construction for haplotype-based marker development in tomato

Hye-ri Jeong · Bo-Mi Lee · Bong-Woo Lee · Jae-Eun Oh · Jeong-Hee Lee · Ji-Eun Kim · Sung-Hwan Jo

Received: 23 June 2020 / Revised: 9 July 2020 / Accepted: 22 July 2020
© Korean Society for Plant Biotechnology

Abstract This report describes methods for selecting informative single nucleotide polymorphisms (SNPs), and the development of an online Solanaceae genome database, using 234 tomato resequencing data entries deposited in the NCBI SRA database. The 126 accessions of *Solanum lycopersicum*, 68 accessions of *Solanum lycopersicum* var. *cerasiforme*, and 33 accessions of *Solanum pimpinellifolium*, which are frequently used for breeding, and some wild-species tomato accessions were included in the analysis. To select tag-SNPs, we identified 29,504,960 SNPs in 234 tomatoes and then separated the SNPs in the genic and intergenic regions according to gene annotation. All tag-SNP were selected from non-synonymous SNPs among the SNPs present in the gene region and, as a result, we obtained tag-SNP from 13,845 genes. When there were no non-synonymous SNPs in the gene, the genes were selected from synonymous SNPs. The total number of tag-SNPs selected was 27,539. To increase the usefulness of the information, a Solanaceae genome database website, TGsol (<http://tgsol.seeders.co.kr/>), was constructed to allow users to search for detailed information on resources, SNPs, haplotype, and tag-SNPs. The user can search the tag-SNP and flanking sequences for each gene by searching for a gene name or gene position through the genome browser. This website can be used to efficiently search for genes related to traits or to develop molecular markers.

Keywords tag-SNP, Haplotype, Molecular marker, NGS, tomato

서 언

효율적인 작물 육종을 위한 분자 마커(molecular marker)의 수요는 꾸준히 증가하고 있으며 분자 마커 개발은 차세대 염기서열 생산 장비(Next Generation Sequencing, NGS)의 발달로 더욱 가속화되어 왔다. 유전체 내 존재하는 다양한 구조적 변이는 분자 마커로 개발되어 품종을 구분하거나 주요한 형질에 연관된 유전자를 찾고 활용하는데 매우 중요한 정보이다. 2007년 이후 NGS 기술의 등장으로 대규모 염기서열 데이터의 생산이 가능해졌고, 이를 통해 유전체 내에 존재하는 구조변이를 확인할 수 있는 분자 마커를 빠르게 대규모로 탐지할 수 있게 되어 변이 탐지에서 양적인 제약이 해결되었다.

SNP는 유전체 내에 가장 빈번하게 발생하는 구조변이의 한 종류로, 목표 유전자 혹은 염색체 위치에서도 변이정보를 확보하기 용이하여 가장 널리 활용되는 분자 마커이다(Collard and Mackill 2008). 그러나 NGS에 의해 발굴된 과도하게 많은 SNP를 모두 확인하는데 소요되는 실험적 비용을 감당할 수 없기 때문에 유전자 유형을 구별하거나 형질과 연결 지을 수 있는 정보력이 높은 분자 마커를 선발하고 최소한의 마커 세트를 구성하고자 하는 시도가 계속되고 있다(Pook et al. 2019). 인간 유전체 분야에서 처음 연구가 시작된 tag-SNP의 개념은 haplotype을 대표하는 마커로서 Hapmap project에서 활발히 연구되었다(Willer et al. 2006). 유전체 내에서 haplotype block을 선발하는 방법은 LD (linkage disequilibrium) block based methods와 genome-wide approaches의 크게 2가지의 카테고리로 나누어 개발되고 있다(Choe et al.

H.-R. Jeong · B.-M. Lee · B.-W. Lee · J.-E. Oh · J.-H. Lee · J.-E. Kim · S.-H. Jo (✉)
(주)씨더스 농업회사법인
(SEEDERS Inc., Daejeon, 34912, Korea)
e-mail: shjo@seeders.co.kr

2018). LD block 기반의 방식은 미리 정의된 haplotype block structure에 의존한다. 염색체 교차 빈도를 기준으로 블록을 구분하게 되면 블록 내에는 교차가 매우 적은 수준에서 발생되므로 블록 내에 haplotype은 아주 적은 변이를 가지게 된다. Shirasawa 등의 보고에 의하면 663개의 토마토 accession을 이용해 LD를 분석했을 때, 총 123 LD Block이 탐색되었고, 염색체 부위별로 LD block의 위치를 조사하였다(Shirasawa et al. 2013). LD block의 평균 길이는 3.2 Mb였으며, block의 크기는 최소 256 bp에서 최대 58.3 Mbp의 LD block이 확인되어 블록 간 크기 차이가 매우 큼을 확인하였다. 123개의 LD block에서 총 437 haplotype을 확인하였고 한 LD block 당 3.6 haplotypes을 보고하였다. de Haas 등은 토마토 60개 RILs의 low coverage sequencing으로 SNP를 탐색하여 genome-wide approaches 방식의 haplotype 분석을 수행하였다(de Haas et al. 2017). Haplotype block 탐색은 마커 수와 유전체 내에서의 거리를 기준으로 분석하였다. Haplotype block은 sliding window approach method를 적용하였으며, 최소 9개의 마커와 최소 거리 10 kbp 이상의 조건에서 시작하여 window size 최대 50개 마커와 250 kbp 이내의 조건에서 탐색하였다. 1,445개의 recombination 발생 결과를 60 RILs 집단에서 haplotype을 적용하여 효과적으로 탐지하였다.

범용성 있는 haplotype을 개발하기 위해서는 모든 토마토에서 공통적으로 적용할 수 있는 통일된 기준을 적용하는 것이 필요하다. LD block의 경우 집단의 특징에 따라 매우 유동적으로, LD block의 위치나 크기가 상당히 큰 차이를 보인다. 또한 sliding window approach를 이용한 genome-wide haplotype 분석 방식도 비교하는 계통 간의 SNP발생유무, SNP수 및 분포에 따라 haplotype block의 크기와 위치가 변화됨을 확인하였다.

따라서 본 연구에서는 유전자를 기본 단위로 하여 haplotype

을 분석하고 유전자의 haplotype을 대변하는 정보력 높은 tag-SNP를 선발하여 분자 마커 개발에 활용하도록 웹 데이터베이스를 구축하여 제공하고자 한다.

재료 및 방법

토마토 표준 유전체 및 SRA data 수집

토마토 표준 유전체(reference genome)는 SGN (ftp://ftp.solgenomics.net/)에서 수집한 *Solanum lycopersicum* Heinz 1706 (ITAG2.4 버전)을 사용하였다(Bombarely et al. 2011). 최근 ITAG4.0의 개선된 표준 유전체 정보가 공개되었으나 기존에 분석되어진 계통과 정확한 비교를 위해 동일한 ITAG2.4 버전을 사용하였다. 표준 유전체는 12개 염색체(chromosome)와 12개 염색체 중 어느 하나의 염색체로 자리 잡지 못한 scaffold 염기서열들을 묶은 1개의 기타 그룹으로 구성된 총 13개의 염색체, 서열 총 길이 781,666,411 bp로 구성되었다. NCBI의 SRA (Sequence Read Archive) 데이터베이스에 축적된 약 800여개 토마토 SRA 정보를 수집하였으며 염기서열 길이는 총 7.8 Tbp이다. 이 중 연구 및 분석에 사용 가능한 형태의 염기서열 데이터로 토마토 234개 계통의 염기서열 데이터를 선정하였다(Lin et al. 2014). 수집된 토마토 SRA 중 재배종으로 가장 많이 이용되고 있는 *Solanum lycopersicum* 126 계통을 선택하였고, 현 재배종과 유사하다고 알려져 있는 *Solanum lycopersicum* var. *cerasiforme* 68계통, 육종과정에서 빈번히 활용되는 *Solanum pimpinellifolium* 33계통을 기본으로 하여 일부의 야생종 토마토 계통을 분석에 포함하였다 (Table 1).

Table 1 Summary of tomato resequencing SRA data

Scientific name	No. of SRA collected	Total length (bp)	No. of SRA analyzed
<i>Solanum lycopersicum</i>	557	5,367,216,027,209	126
<i>Solanum lycopersicum</i> var. <i>cerasiforme</i>	133	818,234,753,929	68
<i>Solanum pimpinellifolium</i>	65	470,510,067,600	33
<i>Solanum habrochaites</i>	17	324,511,688,762	1
<i>Solanum pennellii</i>	14	137,734,433,883	0
<i>Solanum arcanum</i>	12	165,974,167,372	0
<i>Solanum huaylasense</i>	3	101,084,736,600	0
<i>Solanum cheesmaniae</i>	2	68,609,466,800	2
<i>Solanum chilense</i>	2	64,731,490,900	1
<i>Solanum chmielewskii</i>	2	67,986,640,700	0
<i>Solanum neorickii</i>	2	70,608,085,300	1
<i>Solanum peruvianum</i>	2	69,515,482,100	2
<i>Solanum corneliomuelleri</i>	1	32,520,818,500	0
<i>Solanum galapagense</i>	1	30,910,824,200	0
Total	813	7,790,148,683,755	234

Genome-wide SNP 분석

수집한 토마토 234개 계통 별 유전체 염기서열 데이터의 품질을 측정하고, 기준 품질 이상의 염기서열만을 선별(quality filtering) 하는 가공 작업을 SolexaQA package (v. 1.13)를 이용하여 수행하였다(Cox et al. 2010). 품질 측정의 기준 값은 phred score 20 이상, 최소길이 25bp 이상으로 적용하였다. 품질 기준을 통과한 토마토 염기서열들은 BWA (Burrows-Wheeler Aligner; v0.6.1-r104)를 이용하여 토마토 표준 유전체 기준으로 정렬(alignment)을 수행하였다(Li and Durbin 2009). 프로그램 수행을 위한 옵션값으로는 mismatch 최대 2개, seed length 27 bp 설정 외에 기본값(default)을 적용하였다. 토마토 각 계통의 표준 유전체로의 염기서열 정렬 결과를 SAMtools (v0.1.16) 프로그램을 이용하여 유전체 전반적으로 발생한 대량의 SNP를 탐색하였다(Kim et al. 2014; Li et al. 2009). SAMtools 프로그램의 옵션은 대부분 기본 값(default)을 적용하였고, 정확도 높은 SNP를 선별하기 위해 정렬 품질(alignment quality) 값을 기본 값인 25보다 높은 30으로 적용하고, 적어도 3개 이상의 원 서열(raw reads; 염기서열을 구성하는 짧은 단편서열)이 정렬된 위치에서 추출된 SNP만을 선별하였다. 토마토 234개 계통의 각각의 SNP 발생 정보를 통합하여 매트릭스(matrix) 형태로 염색체 번호, 위치 정보, 표준유전체의 염기서열을 포함하여 SNP 정보를 정리하였다.

토마토 유전자 haplotype 분석

유전자 단위의 haplotype을 분석하기 위해 SNP 매트릭스가 제공하는 SNP position 정보와 토마토 표준 유전체의 physical position 정보를 이용해 유전자 내에 발생한 SNP를 분류하였다. SNP 정제는 토마토 계통 간 같은 유형의 SNP 빈도(MAF)가 5% 이상, 원 서열 정렬 개수(read depth)가 5개 이상, 염기서열이 결손된 계통이 30% 이하의 선별 기준을 만족하는 SNP를 선별하였다. 정제된 SNP로 234계통의 SNP를 유전자 별로 구분하여 multiple alignment와 phylogenetic tree 분석을 통하여 haplotype 분석을 진행하였다. Haplotype은 multiple alignment의 결과의 거리 차를 이용해 구분하였으며 표준유전체와 같은 그룹으로 묶인 경우를 'type 1'으로 분류하였다. 그룹 내 묶인 accession 수가 많은 순서로 이름을 명명하였다.

Tag-SNP 선별

유전자 단위로 SNP를 분류하고 유전자 내 SNP가 존재하지 않는 유전자는 분석에서 제외되었다. 유전자 단위의 haplotype을 구분할 수 있는 SNP 중에서 향후 활용도가 높은 SNP를 우선 선별하기 위하여 non-synonymous SNP를 탐색하여 tag-SNP의 우선대상으로 선별하였다. 유전자 내에 non-synonymous SNP가 없는 경우는 haplotype 간 구분이 가능한 synonymous

SNP를 다음 대상으로 선별하였다. 웹 데이터베이스를 구축할 때 non-synonymous SNP가 미치는 아미노산의 변화 정보를 제공하였다.

토마토 계통 간 LD 분석

토마토 계통 간 SNP matrix를 이용하여 염색체 별로 LD 분석을 실시하였다. LD 분석에는 Broad Institute의 Barrett Jeffrey 팀에서 개발한 Haploview 프로그램을 도입하여 이용하였다(Barrett 2009). 프로그램이 가지고 있는 사용가능한 마커 수의 제한을 극복하기 위하여 *Solanum lycopersicum* 그룹과 *Solanum pimpinellifolium* 그룹 간 LD 분포 유형을 염색체 수준에서 조사하되, LD 분석 시 MAF 20% 이상을 만족하는 SNP를 1kb에서 10kb의 window size 당 1개 혹은 2개를 선별하였다.

Web database 구축

SRA 234계통 유래의 SNP는 MySQL을 통해 데이터베이스로 구축하였고, web coding에는 php 언어를 활용하였다. 활용도를 높이기 위해 기존에 구축하여 공개한 TGsol (<http://tgsol.seeders.co.kr/>) 웹 사이트와 연계하였다(Choe et al. 2018). 사이트 내 SNP, haplotype 및 tag-SNP는 genome browser에서 탐색할 수 있도록 구현하였다. 필요에 따른 분석 대상 선택을 위해 234계통은 선택 옵션을 두었고, 선택된 계통을 이용하여 multiple alignment 및 phylogenetic tree를 분석할 수 있게 구현하였다.

결과 및 고찰

Genome-wide SNP

수집한 토마토 234계통 SRA의 총길이는 1.58 Tbp였으며, 유전체 단편 염기서열(short read)의 품질을 측정하고 기준 품질 이하의 염기서열을 제거하여 원 데이터(raw data) 대비 79%인 1.24 Tbp 길이의 염기서열을 확보하였다. 토마토 계통 별로 정제된 염기서열들을 토마토 표준 유전체(*S. lycopersicum* 2.4v)에 정렬한 결과, 전처리된 단편 서열 중 89.97%가 표준 유전체에 정렬되어 변이정보를 탐색할 수 있었다.

토마토 234계통의 각각의 SNP 발생 정보를 통합하여 정리한 결과, 통합 29,504,960개의 SNP 정보를 확보하였다. 토마토 234개 계통 중 주요 3종별 SNP 발생 경향을 조사한 결과, 현재 표준유전체와 재배종이 포함되어 있는 *Solanum lycopersicum*에 속하는 126개 계통은 1개 계통(SRR1572666)을 제외하고 대부분 1만 ~ 10만여 개의 SNP가 탐색되었다. *Solanum lycopersicum* var. *cerasiforme*에 속하는 68개 계통은 적게는

Fig. 1 SNPs from the r1 have removed the hyphen for consistency. esequenced Please check if this should be “genes”. Genomes cannot be in accessions. Genome is correct.genomes in the 234 tomato accessions and phylogenetic analysis. A: Difference in the SNPs from 68 accessions of *Solanum lycopersicum* var. *cerasiforme*. B: Difference in the SNPs from 126 accessions of *Solanum lycopersicum*. C: Difference in the SNPs from 33 accessions of *Solanum pimpinellifolium*. D: phylogenetic analysis of 234 tomato accessions

10만여 개에서 많게는 250만 여개에 이르는 등 *Solanum lycopersicum*에 비해 계통 간 발생한 SNP 개수의 범위가 넓었다. *Solanum pimpinellifolium*에 속하는 33개 계통은 표준 유전체 대비 250만에서 3백만여개의 SNP가 추출되어 품종별 현격한 차이가 있음을 확인하였다(Fig. 1).

추출한 SNP 정보를 기반으로 토마토 234개의 계통수 분석을 실시하였다. SNP 개수와 상응하게 *Solanum lycopersicum*, *Solanum lycopersicum* var. *cerasiforme*, *Solanum pimpinellifolium* 그룹을 명확하게 구분할 수 있었으며 종내 변이의 폭도 확인할 수 있었다. *Solanum lycopersicum*의 경우 대부분 변이의 폭이 적게 나타난 것에 비해 *Solanum lycopersicum* var. *cerasiforme*, *Solanum pimpinellifolium*에서는 계통간 변이의 폭이 크게 나타났다(Fig. 1).

유전자 단위의 haplotype 분석

유전자별 SNP를 분류하기 위해 토마토 표준 유전체 내 유전자 위치(physical position) 정보를 이용해 34,727개의 유전자 내에 발생한 SNP를 선별하여 유전자 별로 SNP를 분류하였다. 유전자별 SNP의 발생정도를 조사한 결과 약 6.2%에 해당하는 2,160개의 유전자 내에는 SNP가 발생하지 않는 것으로 조사되었다. 나머지 32,567개의 유전자에서 $MAF \leq 0.2$ 조건에 해당하는 SNP를 제거하여 27,630개의 유전자에서 175,287개의 정제된 SNP를 확보하여 향후 분석에 사용하였다. 따라서, 유전자 내에 SNP가 존재하지 않는 2,160개의 유전자와 $MAF \leq 0.2$ 필터조건을 만족하지 못하는 소수의 SNP를 가진 유전자가 제외되면서 4,930개의 유전자가 haplotype 분석에서 제외되었다(Table 2).

유전자 당 SNP 발생 빈도를 살펴보면, 1개에서 10개 이내의 SNP가 발생하는 경우가 가장 높은 비율을 차지했고, 50개 이상의 SNP가 발생한 유전자도 2,189개 탐색되었다(Fig. 2).

선발된 SNP의 유전자 haplotype 분석은 multiple alignment를 통해 유전자별로 크게 2개의 단상형으로 구분할 수 있음을 확인하였다. Phylogenetic tree 분석을 통해 살펴보면 *Solanum lycopersicum*과 *Solanum lycopersicum* var. *cerasiforme*가 한 종류의 haplotype으로 묶이고 *Solanum pimpinellifolium*이 또 다른 haplotype으로 분류되는 경향을 보였으나 반드시 일치하지는 않았다. 표준유전체와 동일한 유형을 ‘type 1’으로 분류하고 다음으로 많이 나오는 유형을 ‘type 2’로 분류하였다. SNP 중 30% 이상 차이를 보이는 경우에는 ‘other’로 분류하였다. *Solanum lycopersicum* 각 계통의 유전자를 타입별로 분류해보면 대부분의 염색체가 ‘type 1’으로 구성되어 있으나 *Solanum lycopersicum* var. *cerasiforme*의 경우는 계통별로 ‘type 1’과 ‘type 2’ 구성의 변이차가 큰 것을 확인하였다. *Solanum pimpinellifolium*의 경우는 대부분의 계통이 ‘type 2’의 구성으로 되어 있으나 ‘other’ 구성이 계통에 따라 차이가 있었다(Fig. 2).

유전자 단위의 tag-SNP 분석 (non-synonymous SNP 탐색 및 tag-SNP 선별)

각 유전자의 haplotype은 여러 개의 SNP로 구성되어 있는 block의 기본단위로 볼 수 있다. Haplotype을 실험적으로 활용하기 위해서 각 유전자의 haplotype을 대변할 수 있는 tag-SNP를 선별하였다. 유전자의 haplotype을 구별할 수 있는 SNP 중에서 정보력이 높은 SNP를 tag-SNP로 선별하기 위하여 아미노산 변화를 야기하여 유전자 기능에 영향을 미칠 가능성이 높은 non-synonymous SNP를 우선 선별하였다. 그 결과, 13,845개의 유전자가 선별되었고, 해당 유전자에서 발생한 non-synonymous SNP를 tag-SNP로 선정하였다. 유전자 내에 존재하는 SNP 중 non-synonymous SNP가 없는 경우, synonymous SNP 중에서 haplotype을 구분할 수 있는 SNP를 선별하여

Table 2 SNPs used in gene-based haplotype analysis

Chromosome	No. of input SNPs	No. of input SNPs ($MAF \leq 0.2$)	No. of input genes	No. of input genes ($MAF \leq 0.2$)
Chr0	6,288	1,698	480	370
Chr1	74,835	21,048	4,091	3,493
Chr2	52,467	18,255	3,174	2,771
Chr3	55,128	15,581	3,167	2,607
Chr4	49,706	20,587	2,624	2,387
Chr5	39,644	14,309	2,303	2,062
Chr6	56,530	12,386	2,685	2,211
Chr7	45,452	11,247	2,399	1,954
Chr8	46,664	13,379	2,357	2,020
Chr9	55,581	11,040	2,379	1,902
Chr10	36,516	10,379	2,376	1,961
Chr11	38,389	13,907	2,217	1,944
Chr12	36,131	11,451	2,315	1,948
Total	563,331	175,287	32,567	27,630

Fig. 2 Gene-based haplotypes. **A:** Distribution of SNPs per gene. Most genes contained one to nine SNPs. **B:** Classification of gene-based haplotypes using multiple alignment analysis. Gene SNPs were mainly divided into two groups. **C:** Comparison of haplotype compositions in three *Solanum* species. Accessions in *S. lycopersicum* were mainly composed of type 1 (blue); accessions in *S. pimpinellifolium* were mainly composed of type 2 (red) and other types (green). Composition of accessions in *S. lycopersicum* var. *cerasiforme* was mixed, and the gap between individuals was large

13,694개의 유전자를 대변할 수 있는 tag-SNP로 선발하였다 (Table 3). 따라서 총 27,539개 유전자를 대변하는 tag-SNP를 선발하여 웹 데이터베이스에 공개하였다. 이는 전체 32,567개의 유전자들 중 85%에 해당하는 haplotype block 확장을 위한 기본단위로 활용될 수 있을 것이다.

토마토 계통 간 LD (Linkage Disequilibrium) 분석

LD block은 분자 마커 수를 효과적으로 줄일 수 있는 방법이다. 재배종이 포함되어 있는 *Solanum lycopersicum* 계통들을 대상으로 LD 분석을 실시한 결과, 1번, 2번, 4번, 6번, 10번, 11번, 12번 염색체에서 큰 범위의 LD block이 관찰되었다(Fig. 3).

Table 3 Selection of tag-SNPs that can distinguish gene-based haplotypes

Contents	No. of SNPs
SNPs in 234 tomato accessions	29,504,960
SNPs in gene regions	593,331
SNPs filtered by $MAF \leq 0.2$	175,287
Tag-SNPs from non-synonymous SNPs	13,845
Tag-SNPs from synonymous SNPs	13,694
Total tag-SNPs	27,539

염색체의 진정염색질(euchromatic) 혹은 이질염색질(heterochromatic) 영역별 LD의 분포 경향이 나뉘기보다는 염색체 별 혹은 특정 부위에 LD가 크게 형성되어 있음을 확인할 수 있었다. 이러한 현상은 육종 과정 중에서 인위적인 선발을 거쳐온 *Solanum lycopersicum* 그룹의 경우 육종가들의 목적에 따라 유용 형질을 포함하고 있는 유전자군을 인위적으로 선발했기 때문에 일어나는 영향으로 추측할 수 있다. 유전체 내에서 임의로 선발되는 SNP를 이용하는 기술에 비하여 단위가 규정되어 있는 유전자 단위의 tag-SNP와 함께 LD

block의 tag-SNP를 활용하면 교배조합이 바뀌는 다양한 유전집단에 대하여 표준화된 tag-SNP를 구축하여 haplotype block을 데이터베이스화 할 수 있을 것으로 예상된다.

Web database 구축

정보 활용을 증대하기 위해 기존에 구축된 가지과(Solanaceae) 작물 유전체 데이터베이스인 TGSol (<http://tgsol.seeders.co.kr/>) 웹 사이트 내에 사용된 자료의 상세 정보, SNP, haplotype, tag-SNP 정보를 탐색할 수 있도록 구현하였다. 유전자별 haplotype은 genome browser를 통해 탐색하고자 하는 유전자를 유전자 명 혹은 위치 정보를 검색하여 web interface 환경하에 해당 유전자로 접근하도록 하였다. 검색하고자 하는 유전자를 선택하면, 유전자의 상세정보가 제공되며 확인하고자 하는 계통을 선발할 수 있도록 선택 옵션창을 구성하였다. 모든 계통을 비교하고자 할 경우에는 ‘All’ 버튼을 선택하면 된다. SNP는 MAF filtering을 적용한 경우와 그렇지 않은 경우를 선택하여 확인할 수 있다. 이 과정은 ‘Re-align’ 버튼을 통해 실행시킬 수 있으며, 실행 후 하단에 해당 유전자

Fig. 3 LD blocks of *Solanum lycopersicum*. Red blocks of chromosomes 1, 2, 4, 6, 10, 11, and 12 represent LD blocks

Fig. 4 Online database of haplotypes and tag-SNPs in the genome browser. A: Tomato genome browser is TGsol. By clicking on a gene box, the user can see detailed information on the gene. B: Accession selection checkbox. The user can select the accession by selecting this box. C: SNPs in the gene of selected accessions. Clustering of similar haplotypes through multiple alignment. D: Phylogenetic tree visualization. Phylogenetic trees can be constructed using multiple sequence alignment. The user can choose between circular and phylogram phylogenetic trees by selecting the available options

에 속한 haplotype 결과를 확인할 수 있으며 ‘Tree view’를 통해 phylogenetic tree를 분석할 수 있다. Phylogenetic tree는 옵션을 변경하여 그림의 크기, tree의 type 등을 결정하여 사용하기에 적합한 목적에 맞는 이미지가 생성될 수 있도록 구현하였다(Fig. 4). 분자 마커 개발을 위해 필요한 염기서열 정보는 유전자 단위로 다운로드 하거나 혹은 SNP 주변 서열을 이용하여 사용하고자 하는 장비에 맞도록 분자 마커를 개발하

는데 사용할 수 있다.

적요

유전체 정보가 공공 데이터베이스 내에 빠르게 축적이 되면서 유전체 데이터의 활용도를 높이기 위한 재가공 기술과 공

유 기술이 지속적으로 중요해지고 있다. 특히 분자육종을 가속화하기 위해서 다양한 목적에 맞는 분자 마커 개발이 중요하다. 본 연구는 이러한 요구를 해소하기 위해 유전자 단위에서 haplotype을 기본단위로 구분하고 해당 유전자의 haplotype을 대변하는 tag-SNP를 선발하여 분자 마커 등을 개발하는데 사용할 수 있도록 관련 정보를 웹 사이트를 통해서 제공하고자 웹 데이터베이스를 구축하였다. 본 연구를 통해 선발된 각 tag-SNP는 하나의 유전자를 대변할 수 있고, 각 유전자의 haplotype을 구분할 수 있으며, 해당 유전자의 염색체 내 위치 정보, non-synonymous SNP의 정보를 담고 있다. 따라서 기존 무작위 방식으로 선발되어 사용되던 SNP에 비하여 정보력이 높은 tag-SNP를 활용해서 haplotype block을 확장할 수 있을 것이다. Haplotype의 기본 단위를 유전자로 설정함으로써 집단이 바뀌에 따라 발생하는 SNP의 유무, LD block의 크기 등이 변하는 문제점을 극복하고, 표준화된 haplotype library 작성이 가능할 것이며 이는 또한 분자육종을 위한 분자 마커를 선발하는데 활용될 수 있을 것으로 기대된다.

사 사

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

References

- Barrett JC (2009) Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc* 2009(10):pdb.ip71. doi:10.1101/pdb.ip71
- Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA. (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.* 39 (Database issue): D1149–55. doi: 10.1093/nar/gkq866.
- Choe J, Kim JE, Lee BW, Lee JH, Nam M, Park YI, Jo SH (2018) A comparative synteny analysis tool for target-gene SNP marker discovery: connecting genomics data to breeding in Solanaceae. *Database (Oxford)* bay047, doi:10.1093/database/bay047
- 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(1491): 557–572. doi:10.1098/rstb.2007.2170
- Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485. doi:10.1186/1471-2105-11-485
- 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.* 1;24(6):549–558. doi: 10.1093/dnares/dsx024
- Kim JE, Oh SK, Lee JH, Lee BM, Jo SH (2014) Genome-wide SNP calling using next generation sequencing data in tomato. *Mol Cells* 37(1):36–42. doi:10.14348/molcells.2014.2241
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25(14):1754–1760. doi:10.1093/bioinformatics/btp324
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. doi:10.1093/bioinformatics/btp352
- Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Stadler T, Li J, Ye Z, Du Y, Huang S (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat Genet.* 46(11):1220–6. doi: 10.1038/ng.3117
- Pook T, Schlather M, de Los Campos G, Mayer M, Schoen CC, Simianer H (2019) HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. *Genetics.* 212(4): 1045–1061. doi: 10.1534/genetics.119.302283
- Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I, Hirakawa H, Isobe S, Tabata S (2013) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. *DNA Res.* 20(6):593–603. doi: 10.1093/dnares/dst033
- Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF, Kinnunen L, Mohlke KL, Valle TT, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol.* 30(2):180–90. doi: 10.1002/gepi.20131