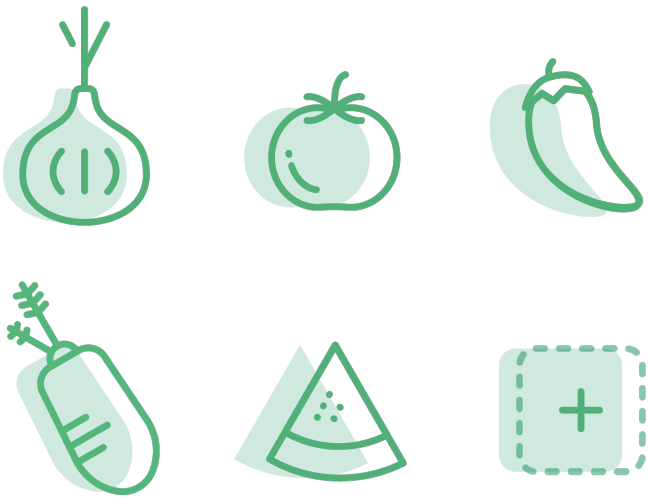
Vol.1

AI를 이용한 디지털 육종 사례집

DIGITAL BREEDING CASE STUDY

양파 구중 개량
고추 신미 판별
밤 중량 예측
도라지꽃색 예측



㈜인실리코젠의 디지털 육종은

생물정보, 빅데이터, AI 기술을 융합하여 *in silico* 상에서
새로운 종자를 찾아 미래 바이오 산업의 근간이 될
가치를 창출합니다.

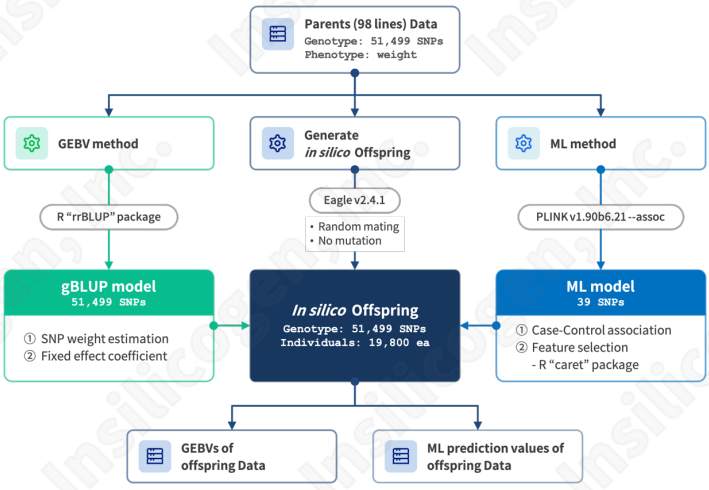
OFFICE 경기도 용인시 기흥구 흥덕1로 13
흥덕IT밸리 타워A동 2901~2904, 2906호
EMAIL info@insilicogen.com
PHONE 031-278-0061
FAX 031-278-0062

㈜ 인실리코젠

양파 구중 개량: 시뮬레이션을 통한 교배조합 작성법

형질	양파의 구중 예측
ML model	RF
모델 정확도	0.88
마커 수	39 SNPs
학습/검증 데이터 수	65 / 30 ea (+19,800 ea)

- 유전자원: 국내 양파 98 계통
- 표현형: 양파 구의 중량
- 유전형: 개체별 1.5 Gb의 GBS 시퀀싱 서열, 51,499 SNPs
- 마커 탐색 방법: GWAS (case-control), feature selection
- 정확도 검증법:
  In silico 자손 19,800 개체의 유전형을 바탕으로 한 육종가와 ML 방식의 형질 예측치의 동일성 비교



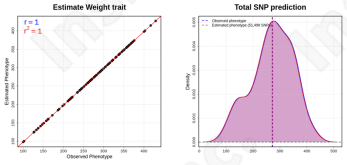
[Fig.1] *In silico* breeding의 모식도

1) GEBV method: 모부본의 유전형을 바탕으로 육종가 산출
2) ML method: 형질 관련 마커 탐색 및 형질 예측 모델링
3) Generate *in silico* offspring: 자손세대 *in silico* 유전형 생성
4) 자손 세대 유전형을 바탕으로 형질 예측 및 고효율 모부본 교배 조합 산출

RESULT

모부본의 51,499 SNPs를 이용한 gBLUP 육종가 모델

- 총 98건의 부모 개체
- gBLUP을 활용한 표현형 예측 모델
- r^2=1.00, 평균오차=2.204(g)

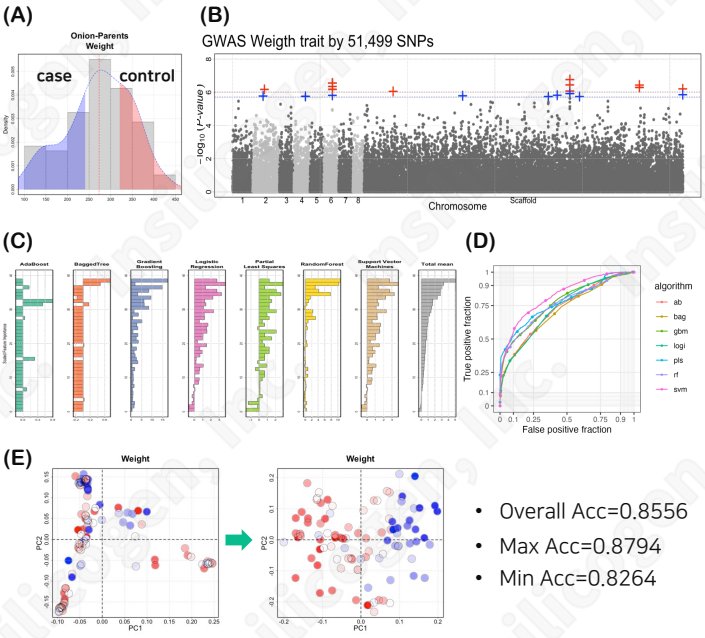


[Fig.2] gBLUP을 이용한 육종가 예측 및 표현형 분포도

RESULT

모부본의 구중 관련 기계학습 모델 구축

- 1차 GWAS를 통한 주요 1,000 SNPs 선발
- 2차 기계학습의 feature selection을 이용한 최종 39 SNPs 선발
- 39 SNPs을 이용한 양파 구중 예측 모델 구축(정확도 : 87.9%)



- Overall Acc=0.8556
- Max Acc=0.8794
- Min Acc=0.8264

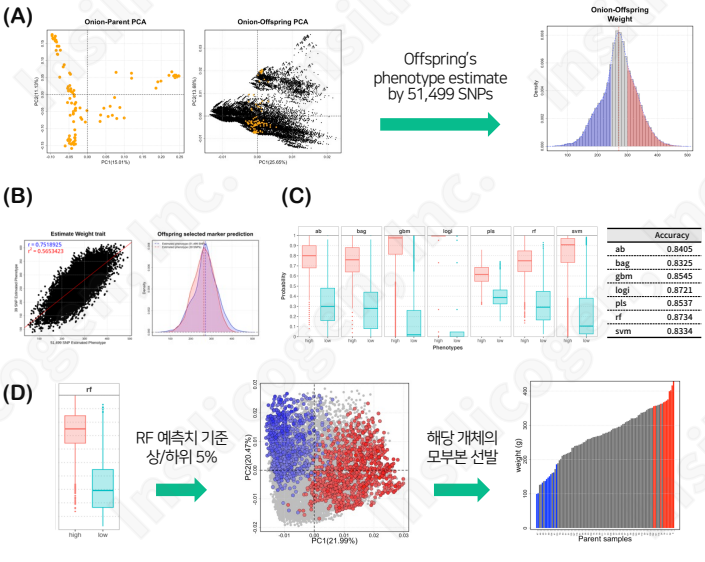
[Fig.3] 구중 관련 마커 탐색 및 예측 모델 구축

(A) 구중 표현형 분포; (B) GWAS 분석 결과 Manhattan plot; (C) Feature selection 결과 마커 중요도 분포; (D) 모델 예측 정확도 (ROC); (E) 전체 51,499 및 39 SNPs 선별 마커를 활용한 PCA 결과 비교

RESULT

교배 조합 작성: 자손 세대 시뮬레이션(예측 모델 적용)

- 부모 개체의 유전형 정보를 활용한 *in silico* 자손 19,800 개체 생산
- *In silico* 자손의 구중 예측치
- 최종 우수 교배 조합 가계



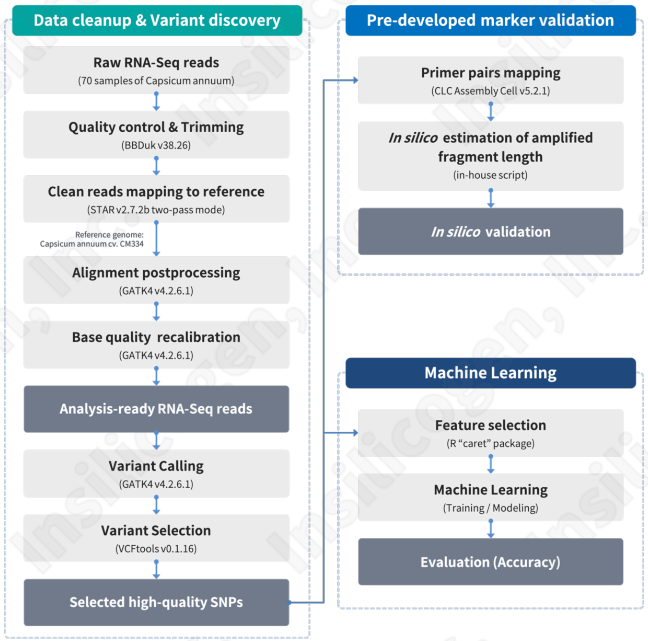
[Fig.4] 교배 조합 선발을 위한 *in silico* 자손 시뮬레이션

(A) *In silico* 자손 유전형 생산; (B) gBLUP 모델을 이용한 자손의 구중 예측치; (C) ML 알고리즘에 의한 자손의 구중 예측치; (D) 우수 알고리즘 기준 상/하위 5% 자손의 유전형과 모부본 표현형 분포

고추 신미 판별: 머신러닝 기반 개체 선발

형질	고추의 신미 판별
ML model	SVM
모델 정확도	0.95
마커 수	23 SNPs
학습/검증 데이터 수	49 / 21 ea

- 유전자원: 고추 70개체
- 표현형: 신미 (고: 10개체 / 중: 10개체 / 저: 50개체)
- 유전형: 개체별 5.89 Gb의 전사체 서열, 48,024 SNPs
- 마커 탐색 방법: Target gene, feature selection
- 정확도 검증법: 진행 중

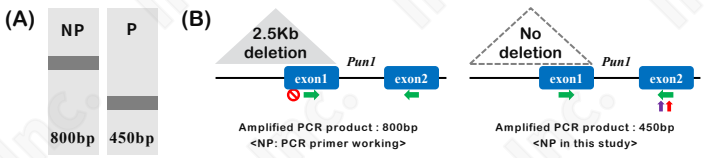


[Fig.1] 머신러닝을 이용한 *in silico* 마커 검증 및 신미 예측 워크플로우

RESULT

기존 분자 마커의 불확실성 발견

- *Pun1* 유전자 다형성: 저신미 자원에서 보고된 결실 없음
- 기존 마커 활용 시 증폭 좌위 예측 결과: 450bp
- 저신미 자원을 신미 자원으로 예측하는 마커의 오작동 사례 확인

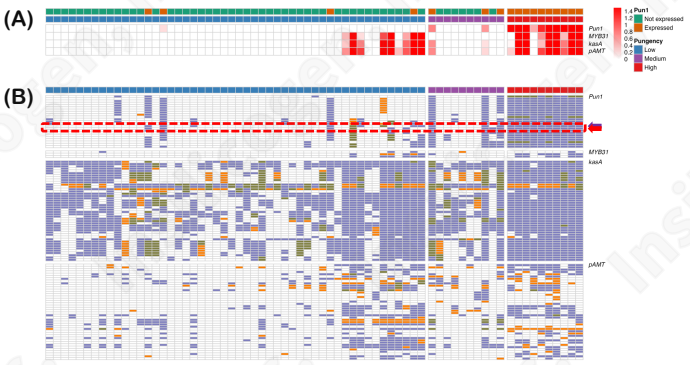


[Fig.2] *Pun1* 유전자 다형성 분석을 통한 기 구축 마커 활용성 검토

RESULT

새로운 변이 23 SNPs 탐색

- *Pun1* 유전자의 발현과 유전변이 양상: 신미 관련 핵심 유전자로 간주함
- *Pun1* 유전자의 발현 개체를 신미 그룹으로 간주(신미 형질이 100% 고정되지 않음)



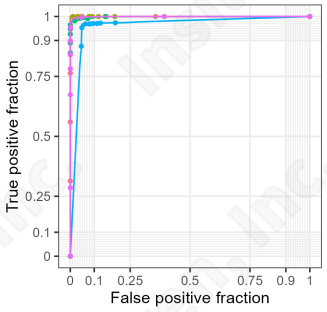
[Fig.3] 주요 캡사이신 생합성 관련 유전자의 발현(A) 및 변이(B) 양상

RESULT

23 SNPs를 이용한 머신러닝 모델 구축

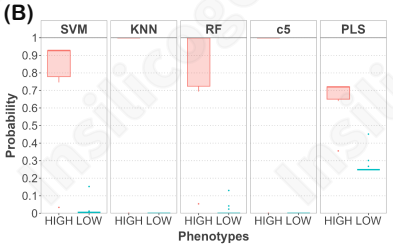
- 선별된 23 SNPs 활용: 48,024 SNPs 분석 필요 없이 신미 판별 가능

(A)



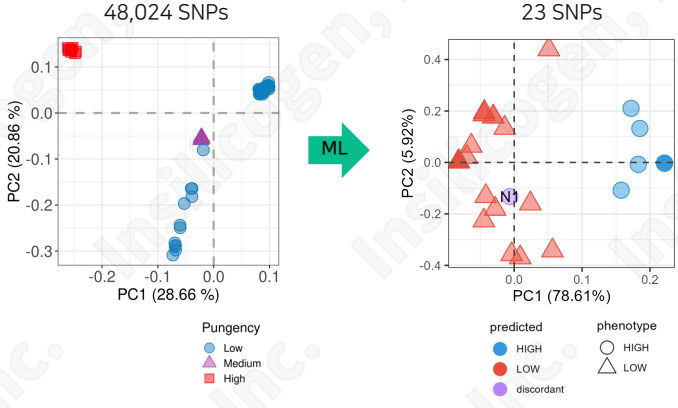
Algorithm	Accuracy
SVM	1.00
KNN	0.99
RF	0.98
c5	0.97
PLS	1.00

(B)



	Sn	Sp	Ac
SVM	0.86	1.00	0.95
KNN	0.86	1.00	0.95
RF	0.86	1.00	0.95
c5	0.86	1.00	0.95
PLS	0.86	1.00	0.95

[Fig.4] 신미 판별을 위한 머신러닝 알고리즘 학습(A) 및 예측 정확도(B)



[Fig.5] 전체 48,024 및 23 SNPs 선별 마커를 활용한 PCA 결과 비교

밤 중량 예측:
밤나무, 산림 자원 육종법

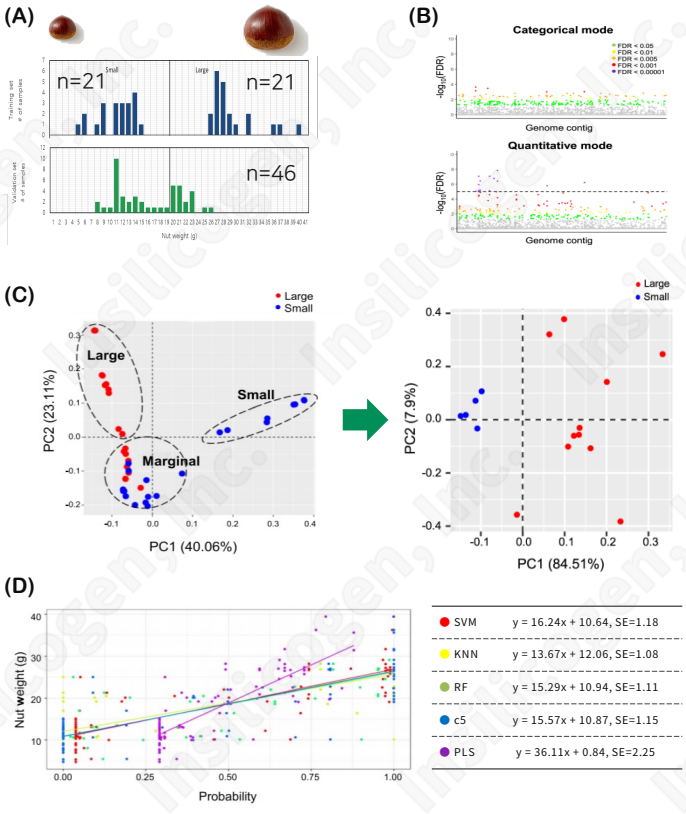
형질	밤 중량 예측
ML model	PLS
모델 정확도	0.70
마커 수	21 SNPs
학습/검증 데이터 수	42 / 46 ea


SCIENTIFIC REPORTS

OPEN Identification of transcriptome-wide, nut weight-associated SNPs in *Castanea crenata*

(Kang et al., 2019)

- 유전자원: 국내 산지의 밤나무 88그루
- 표현형: 밤나무 개체별 알곡의 중량을 측정
- 유전형: 개체별 5.2 Gb의 전사체 서열, 3,271,142 SNPs
- 마커 탐색 방법: GWAS (case-control, quantitative), target gene
- 정확도 검증법: 46개 신규 샘플의 21 SNPs 영역을 시퀀싱하여 유전형을 확인 후 PLS 모델의 예측치와 실측치를 비교



[Fig.1] 밤 알곡 관련 마커 탐색 및 머신러닝 모델 구축

(A) 학습 및 검증 데이터로 이용된 밤 알곡 88개 샘플의 알곡 무게 분포; (B) GWAS 분석을 통한 알곡 크기 연관 마커; (C) 학습 데이터로 활용된 42개 알곡의 전체 유전형으로 확인한 유전적 유사도와 알곡 크기 연관 마커 21 SNPs로 확인한 마커 효율성; (D) 21 SNPs를 활용한 머신러닝 모델의 알곡 중량 예측치

도라지꽃색 예측:
예측 모델 구축 및 실험적 검증

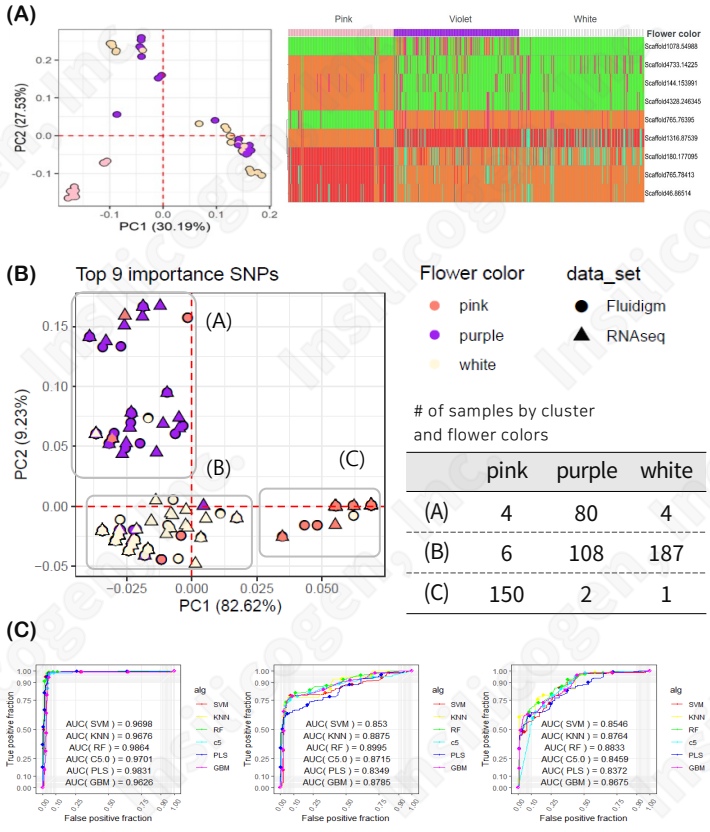
형질	도라지꽃색 예측
ML model	RF
모델 정확도	0.87
마커 수	9 SNPs
학습/검증 데이터 수	60 / 480 ea


scientific reports

OPEN Machine learning, transcriptome, and genotyping chip analyses provide insights into SNP markers identifying flower color in *Platycodon grandiflorus*

(Yu et al., 2021)

- 유전자원: 국내 품종 40종, 북미 품종 20종
- 표현형: 도라지꽃색
- 유전형: 개체별 6.8Gb의 전사체 서열, 76,629 SNPs
- 마커 탐색 방법: GWAS (case-control), feature selection
- 정확도 검증법: 방사선 조사를 통해 유전적 변형이 유발된 480개체의 유전형을 fluidigm chip으로 확인 후 RF 모델의 예측치와 실측치를 비교



	pink	purple	white
(A)	4	80	4
(B)	6	108	187
(C)	150	2	1

[Fig.1] 도라지꽃색 예측을 위한 머신러닝 모델

(A) 3가지 꽃색 60개 샘플의 전체 유전형으로 확인한 유전적 유사도와 꽃색 연관 마커 9 SNPs 변이 양상; (B) 꽃색 연관 마커로 확인한 검증 샘플의 유전적 분포도; (C) 검증 데이터 480개체를 활용한 각 꽃색 예측 모델의 ROC 커브