Tópicos em Genética e Melhoramento de Plantas

Maiara de Oliveira
Patricia Braga

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Computer vision and machine learning for robust phenotyping in genome-wide studies

Received: 05 October 2016
Accepted: 02 February 2017
Published: 08 March 2017

Jiaoping Zhang[1], Hsiang Sing Naik[2], Teshale Assefa[1], Soumik Sarkar[2], R. V. Chowda Reddy[1], Arti Singh[1], Baskar Ganapathysubramanian[2] & Asheesh K. Singh[1]

Traditional evaluation of crop biotic and abiotic stresses are time-consuming and labor-intensive

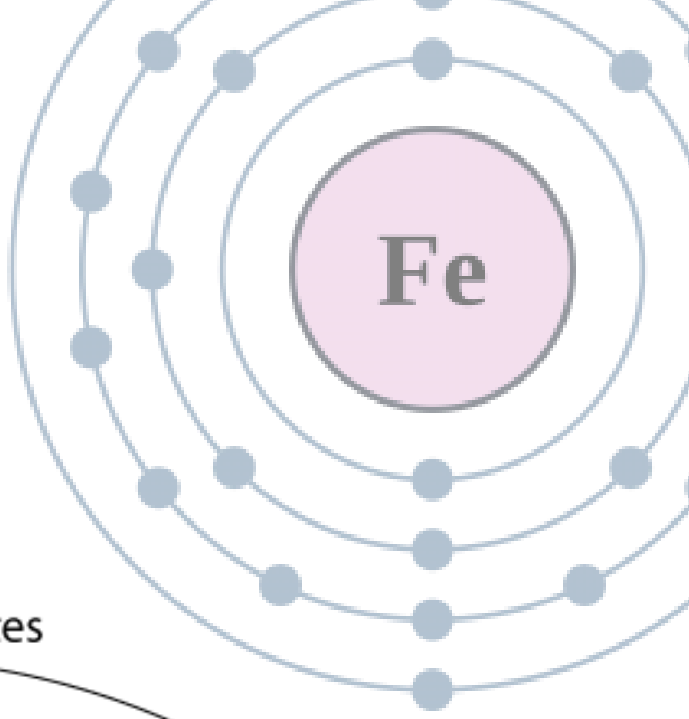
Introdução

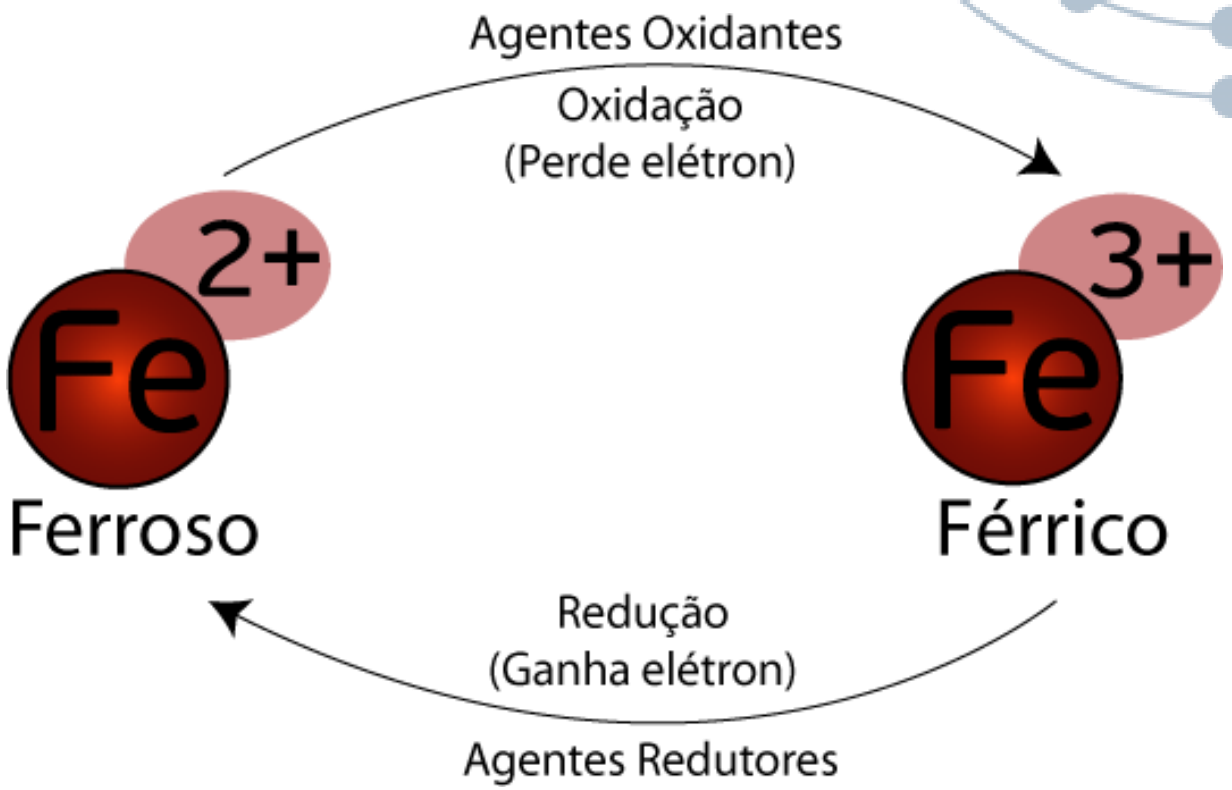
- Nutriente essencial:

  fotossíntese e respiração;

pH 7,4 - 8,5

Solos ricos e Ca





Clorose por deficiência de ferro (IDC)

- Limitante da produção de grãos;
- Clorose internerval;
- $260 milhões em perdas somente em 2012.

- Locus de maior efeito Gm03 (R^2 está em 25-80%);

- loci com menor efeito em 10 dos 20 cromossomos.

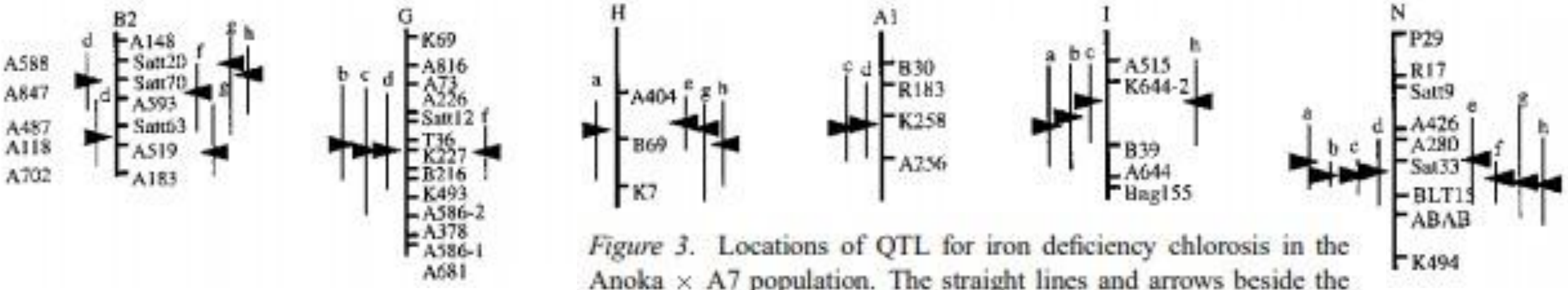


Figure 3. Locations of QTL for iron deficiency chlorosis in the Anoka × A7 population. The straight lines and arrows beside the linkage groups show the confidence intervals and locations of the detected QTL, respectively. Symbols: a, visual scores at V4 stage in 1993; b, visual scores at V2 stage in 1994; c, visual scores at V4 stage in 1994; d, combined visual scores at V4 stages in 1993 & 1994; e, chlorophyll concentrations at V4 stage in 1993; f, chlorophyll concentrations at V2 stage in 1994; g, chlorophyll concentrations at V4 stage in 1994; h, combined chlorophyll concentrations at V4 stages in 1993 & 1994.

Lin, S., Cianzio, S. & Shoemaker, R. Mapping genetic loci for iron deficiency chlorosis in soybean. Molecular Breeding 3, 219–229 (1997)

- A análise GWAS tem sido utilizada para descobrir novas regiões de interesse;

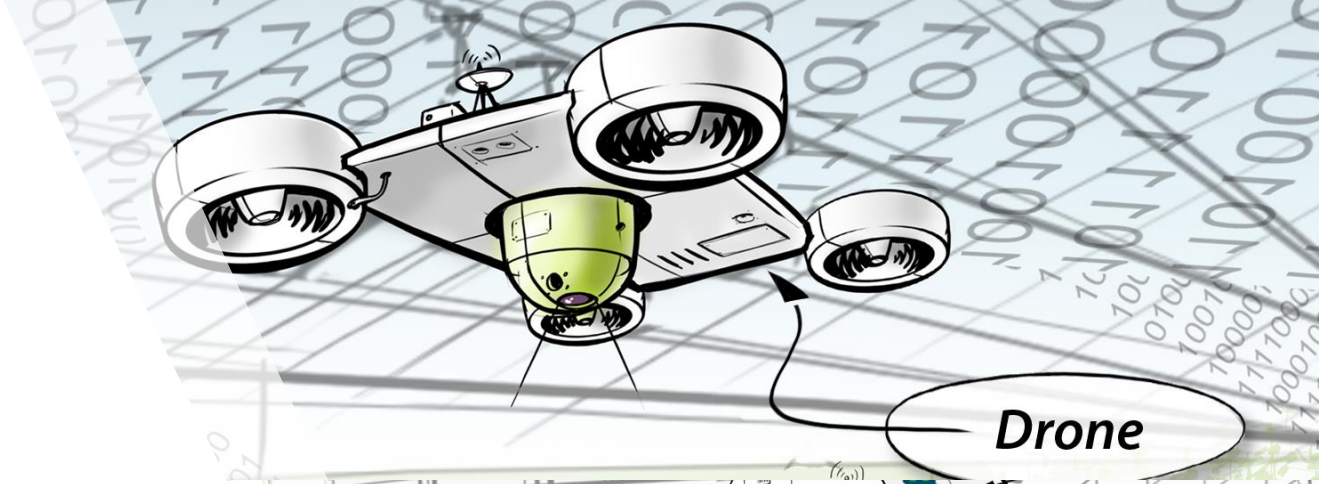
- A tolerância à IDC da coleção de germoplasma de soja do USDA permanecia inexplorada;

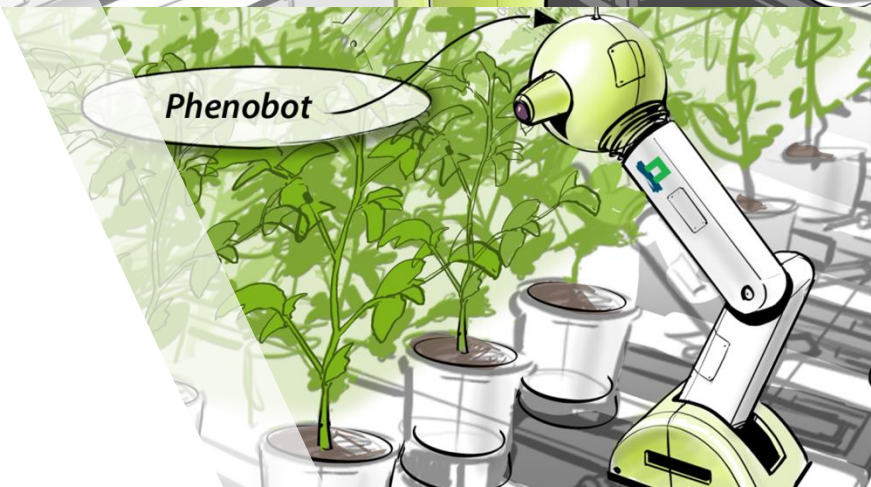
Desafios da fenotipagem

- Notas visuais;
- Leituras com SPAD;

Highthrouput phenotiping

- Sensoriamento remoto;
- Sensores portáteis e;
- Imagens digitais.





- Big-data;
- Machine Learning;
- Acelera o ganho genético.

Objetivo

Realizar análises seleção genômica ampla (GWAS) e predição genômica (GP) da tolerância a clorose por deficiência de ferro em soja usando um grande painel de germoplasma diversificado através de um pipeline de fenotipagem de imagem baseado em machine learning (ML).

Métodos

- **Material vegetal e fenotipagem**

  478 linhas de soja
  - 473 linhas de introdução de plantas (PI)
  - três testemunhas de maturidade
  - e as testemunhas resistente (Clark) e suscetível (Iso-Clark)

Métodos

- **Material vegetal e fenotipagem**

  Linhagens oriundas da coleção de germoplasma de soja do USDA;

  Diferentes grupos de maturação (MG) I (31%), II (36%) e III (33%);

  Delineamento de blocos completos casualizados com quatro repetições (cinco planta/parcela);

Métodos

- **Material vegetal e fenotipagem**

As avaliações visuais de campo (FVR) e imagens digitais das plantas foram coletadas em V2-V3, em V5-V6 e duas semanas mais tarde;

SPAD e coletas de solo apenas em V2-V3 e V5-V6;

Métodos

- **Aquisição de imagens**

As imagens foram tiradas usando uma câmera Canon EOS REBEL T5i com o modelo Scene Intelligent Auto.

Todas as imagens foram armazenadas em qualidade de imagem RAW com uma resolução de 5184 × 3456 (18 M).

Aquisição de imagens

- **Recomendações:**

  Certifique-se de que a iluminação não muda entre uma foto e outra;

  Certifique-se de não haver outros objetos na imagem;

  Certifique-se de que o fundo da imagem é uma cor plana (preferencialmente preto);

  Tirar foto de todo o dossel;

  Evitar uso do flash;

  Tirar fotos de cima para baixo.

Métodos

- **Pré-processamento de imagem**
  - **Segmentação**

  As imagens foram convertidas para RGB;

  148 imagens aleatórias foram utilizadas para definir um limite de valores de matiz verde e amarelo;

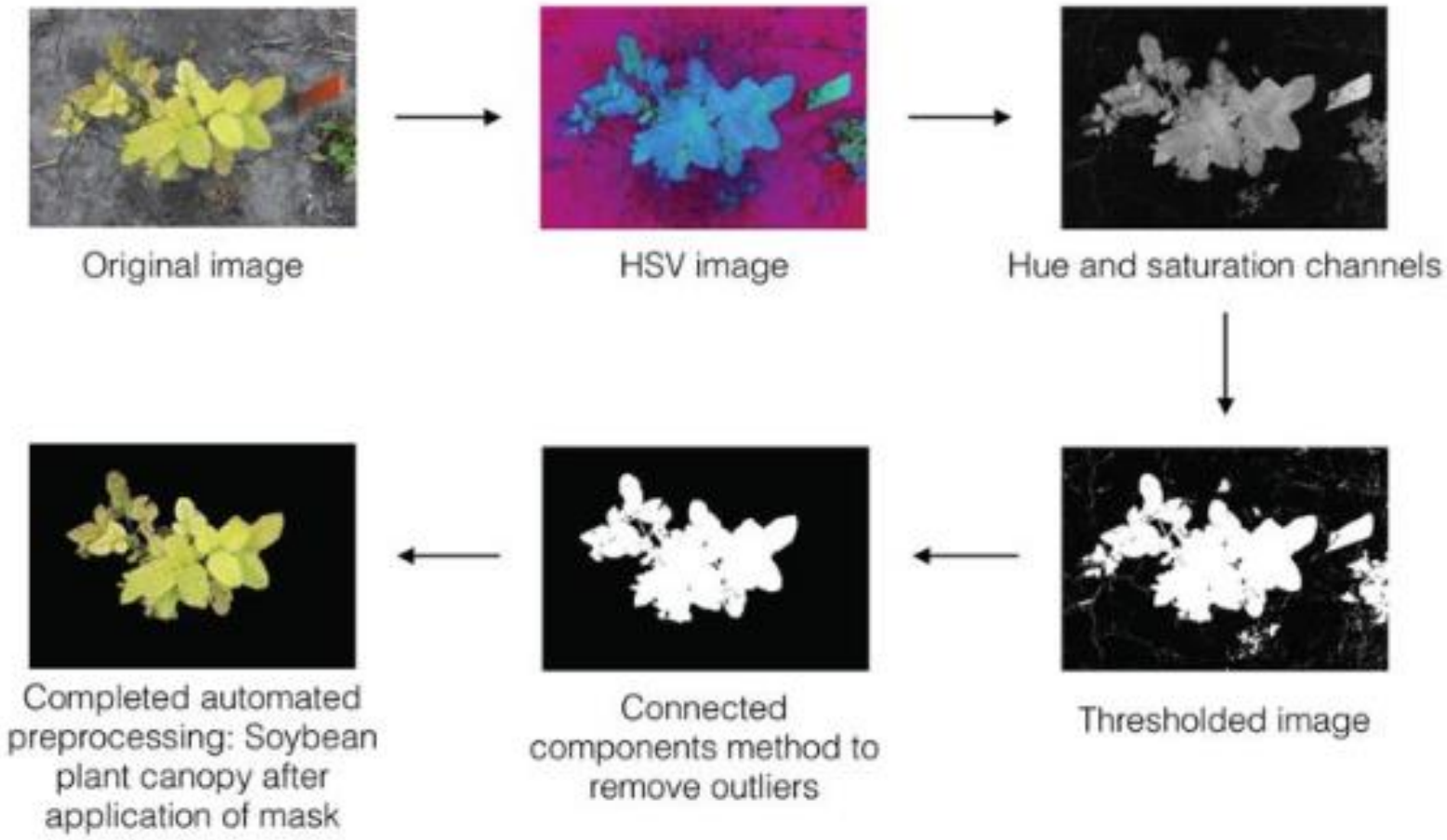
  O limite foi obtido removendo das imagens os pixels que não eram nem verdes nem marrons.

Métodos

- **Pré-processamento de imagem**
  - **Remoção e limpeza de outliers**

Utilizando o método connected-component

Figure 5: Image preprocessing overview on iron deficiency chlorosis-impacted plant canopies in soybean.



Métodos

- **Pré-processamento de imagem**

  Todas as análises foram realizadas usando o Matlab.

Métodos

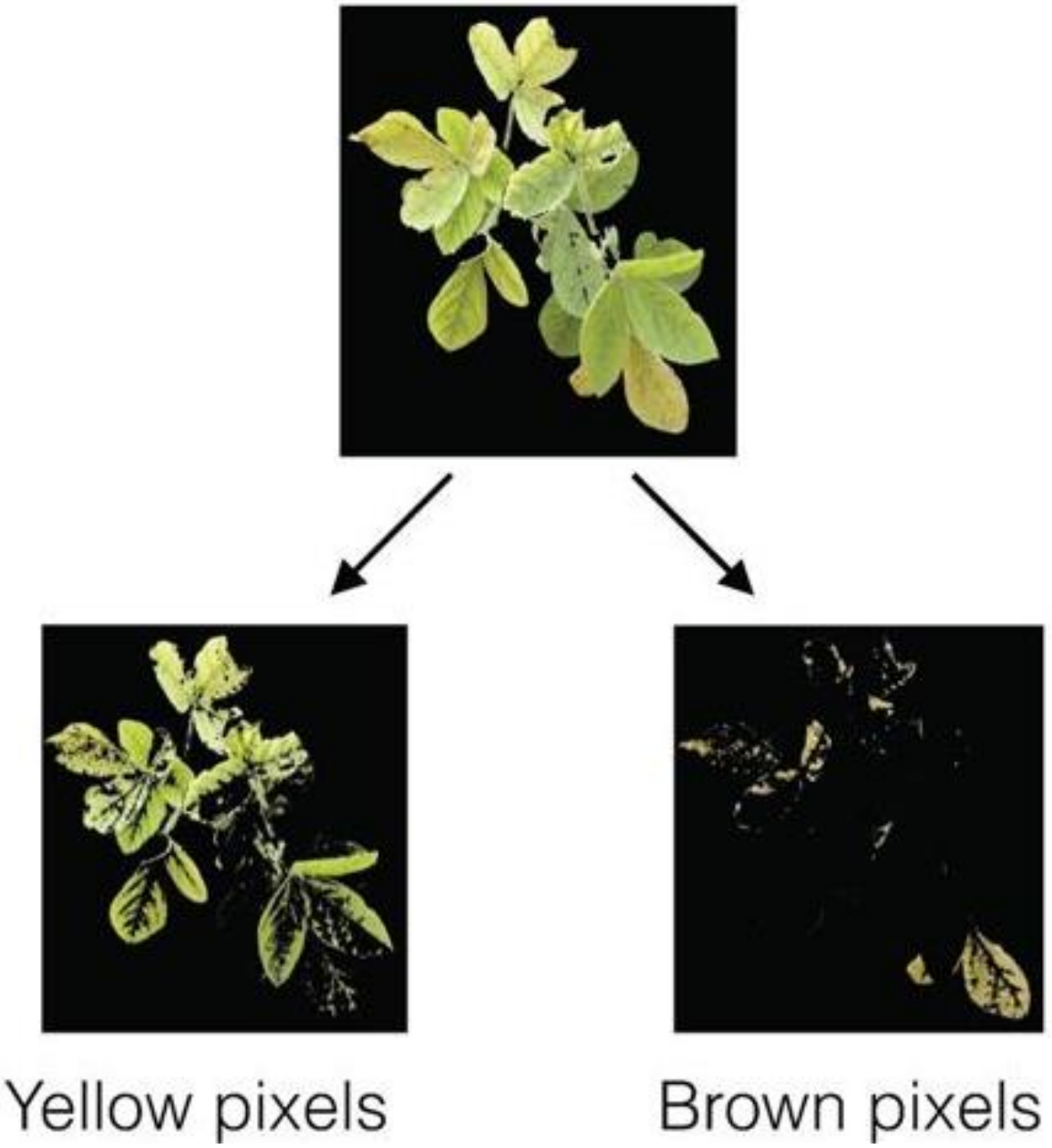
- **Extração de *Features***

A FVR é baseada no amarelamento e necrose das plantas;

Métodos

- **Extração de *Features***

  Os valores de matiz foram usados para identificar os pixels amarelos e marrons e apresentados com a porcentagem das áreas amarelas e marrons do dossel;



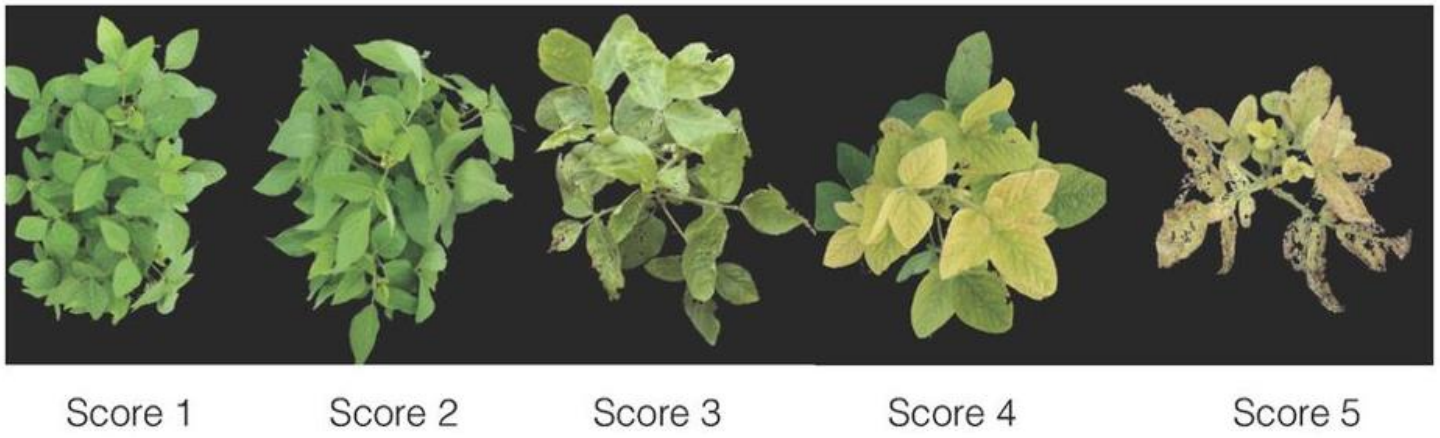
Métodos

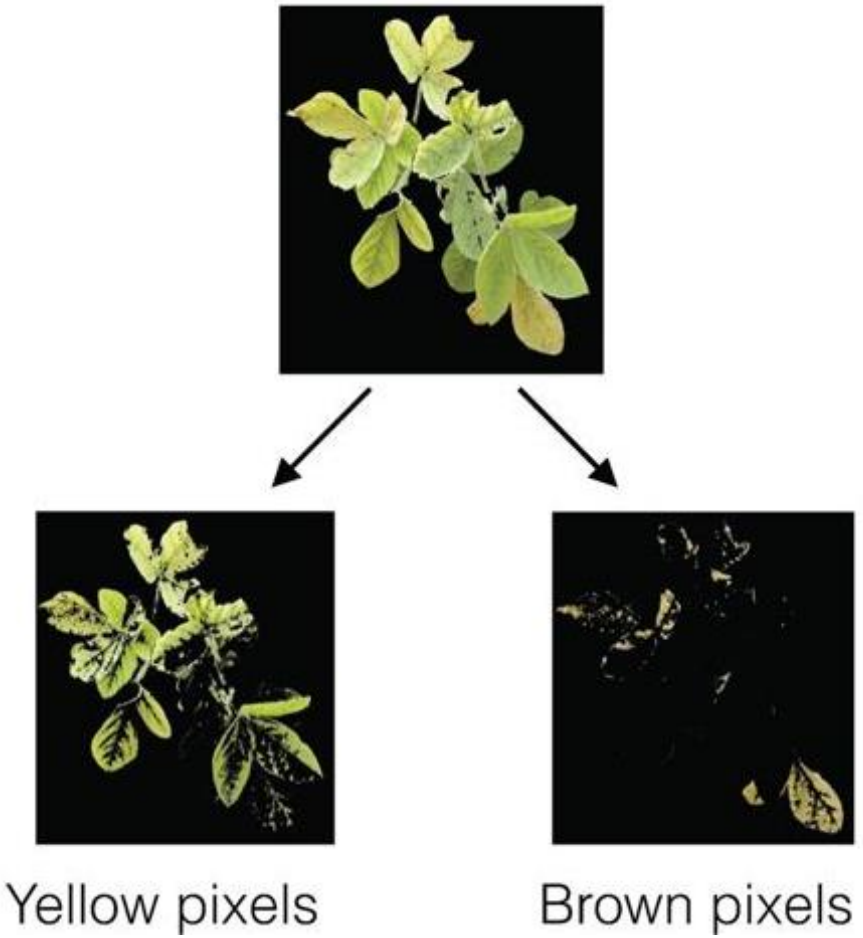
- **Classificação**

75% treinamento

25% teste

Score ML





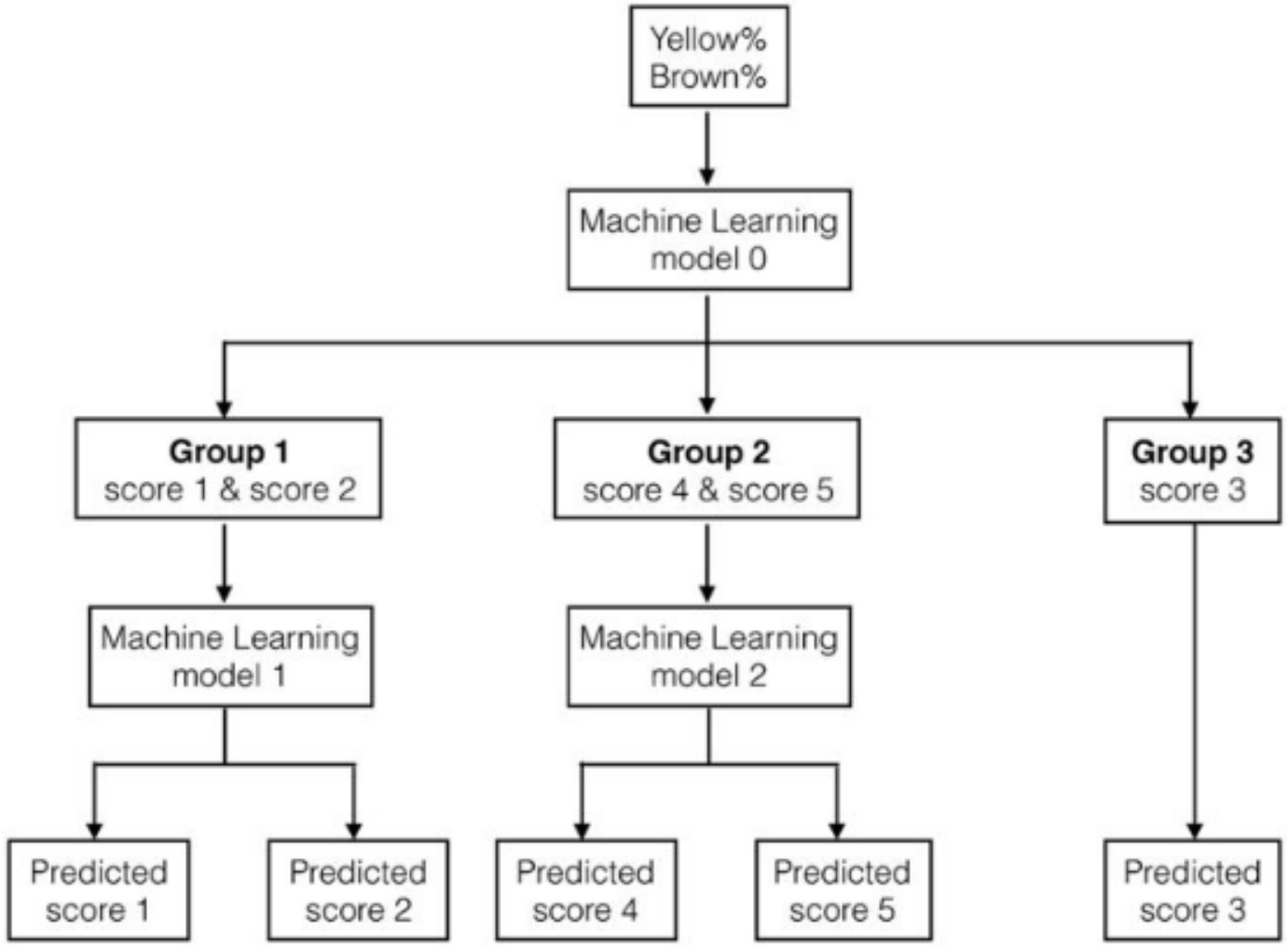
Severity ML

Métodos

- **Classificação**

  Vários algoritmos foram testados, incluindo random forests, decision trees e **support vector machines;**

Figure 8: Flowchart of the hierarchical classifier that was built to map the extracted features (Yellow %, Brown %) to the IDC score.



Métodos

- **Classificação**

Depois disso, é construída uma função de gravidade quantitativa que mapeia o vetor (B%, Y%) para um único número entre 0 e 100.

Table S2. Functions for the development of ML-severity of iron deficiency chlorosis in soybean.

Algorithm†	Method‡	Resubstitution Errors (%)
HM	1	3.09
	2	3.87
	3	9.75
LDA	1	1.16*
	2	1.08
	3	3.02
QDA	1	1.16
	2	0.40
	3	4.03

HM: Hierarchical Model, LDA: Linear Discriminant Analysis, QDA: Quadratic Discriminant Analysis

† These 3 algorithms were used because they performed the best in the task of classifying features to rating and did not require hyperparameter tuning (e. g. k in kNN).

‡ Method 1: $\text{Severity} = (w_1 * Y\%) + (w_2 * B\%)$ (1)

Method 2: $\text{Severity} = (w_1 \text{^} Y\%) + (w_2 \text{^} B\%)$ (2)

Method 3: $\text{Severity} = (w_1 * Y\%) * (w_2 * B\%)$ (3)

Métodos

- **Genotipagem e controle de qualidade**

A genotipagem foi feita utilizando Illumina Infinium SoySNP50K BeadChip;

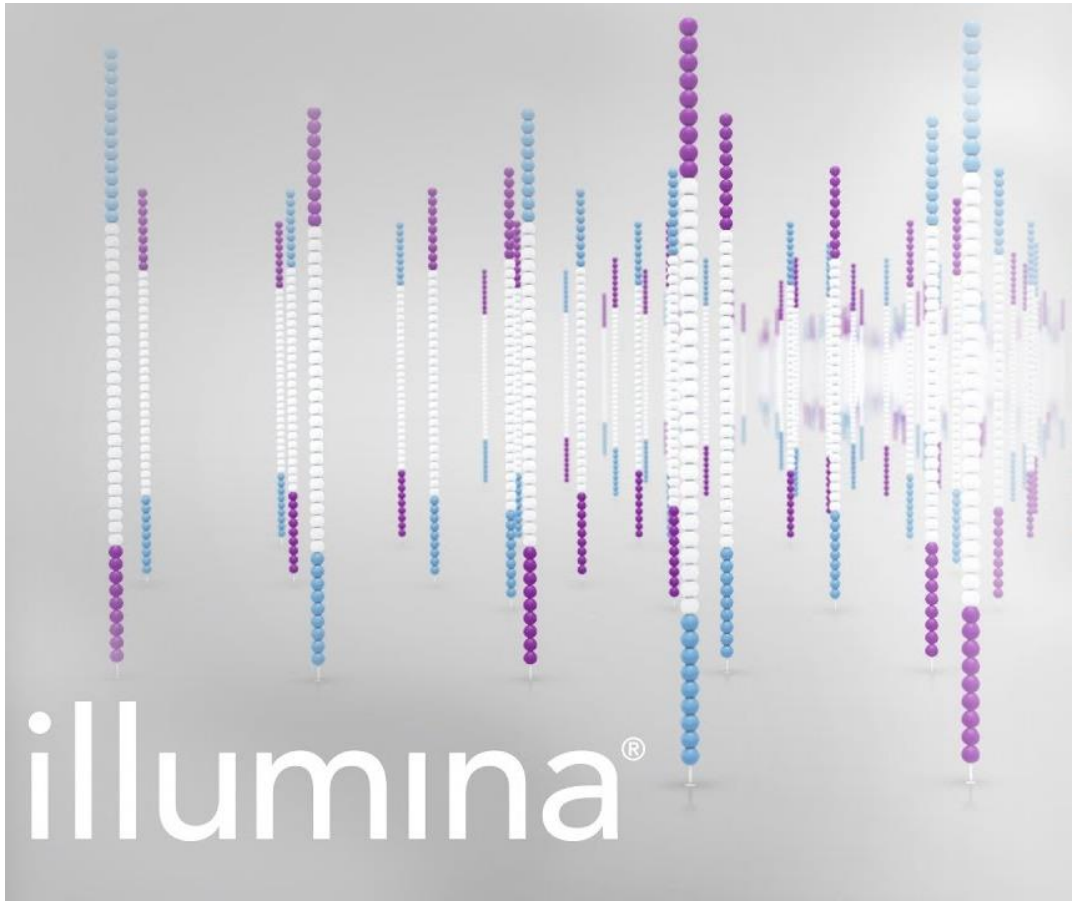
Um total de 42.509 SNPs foram identificados

60 SNPs foram excluídos;

0,5 dados faltantes imputados usando Beagle 3.3.1;

Os SNPs com uma frequência alélica menor (MAF) <5% foram também excluídos ;

36.139 SNPs foram usados para análises GWAS e GP.



Métodos

- **Análise de associação genômica ampla (GWAS)**

Score ML, Severity ML, medidas de SPAD e pH do solo coletados no estádio V5-V6 foram utilizados no GWAS;

Foram obtidos os BLUPs de cada genótipos usando o pacote R lme4;

As análises foram implementadas usando a ferramenta GAPIT;

$$y = \mu + X\alpha + P\beta + Zu + e$$

GAPIT

Genome Association and Prediction Integrated Tool 3.0

Métodos

- **Predição genômica (GP)**

As predições genômicas foram obtidas modelando as características como efeitos fixos e os SNPS como tendo efeitos aleatórios com a melhor predição linear não viesada (RR-BLUP).

$$y = X\beta + Zu + e$$

Resultados

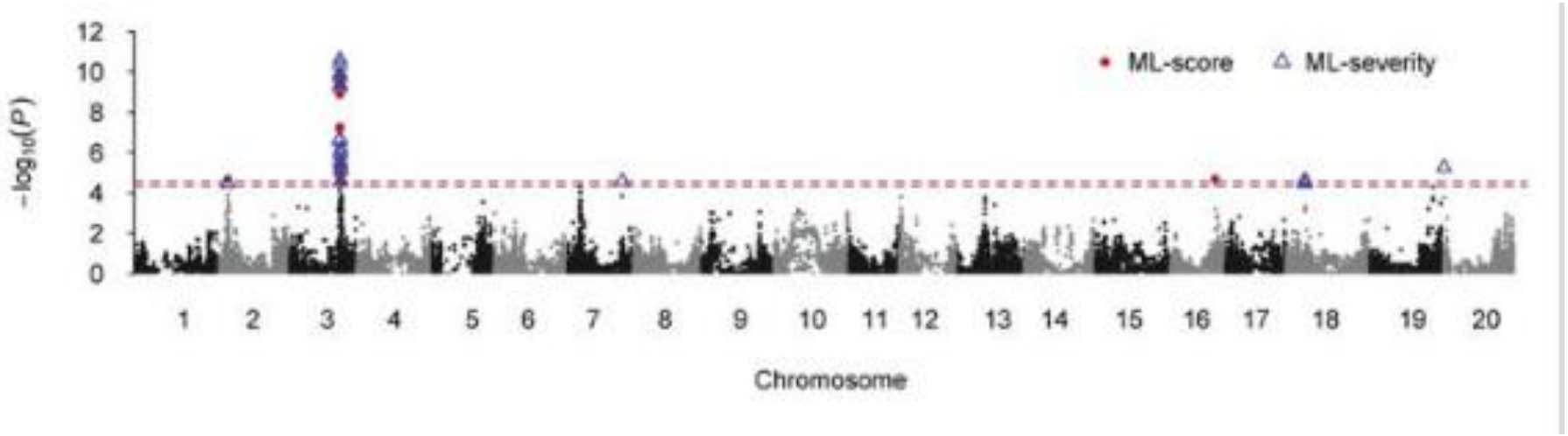
- **Aquisição de imagens, processamento de imagens e fenotipagem usando ML**
  - Um total de 5.916 imagens RGB foram obtidas
  - Após o processamento → Obtenção de 4.366 imagens
  - O ML foi utilizado para classificaçao hierarquica dos genotipos .
    - Uso do SVM, que obteve alta acurácia: 99,4% e 95,9%

Resultados

- **Aquisição de imagens, processamento de imagens e fenotipagem usando ML**
  - Para verificaçao da estrutura. Foram recolhidas 124 imagens de Pis que não faziam parte do painel de associação, seguindo o mesmo protocolo.
    - Acurácia de 96,0% e 90,6%, respectivamente.
  - Por último a função de severidade foi então projetada para associar um único valor (de 0 a 100) para cada genótipo.

Resultados

- **Identificação dos principais QTL ligados em repulsão para o IDC**
  - Análises GWAs identificaram 19 e 27 SNPs associados com o score ML e severity ML, respectivamente.
    - 17 SNPs eram comuns e marcados em uma região de 847 kb em Gm03.



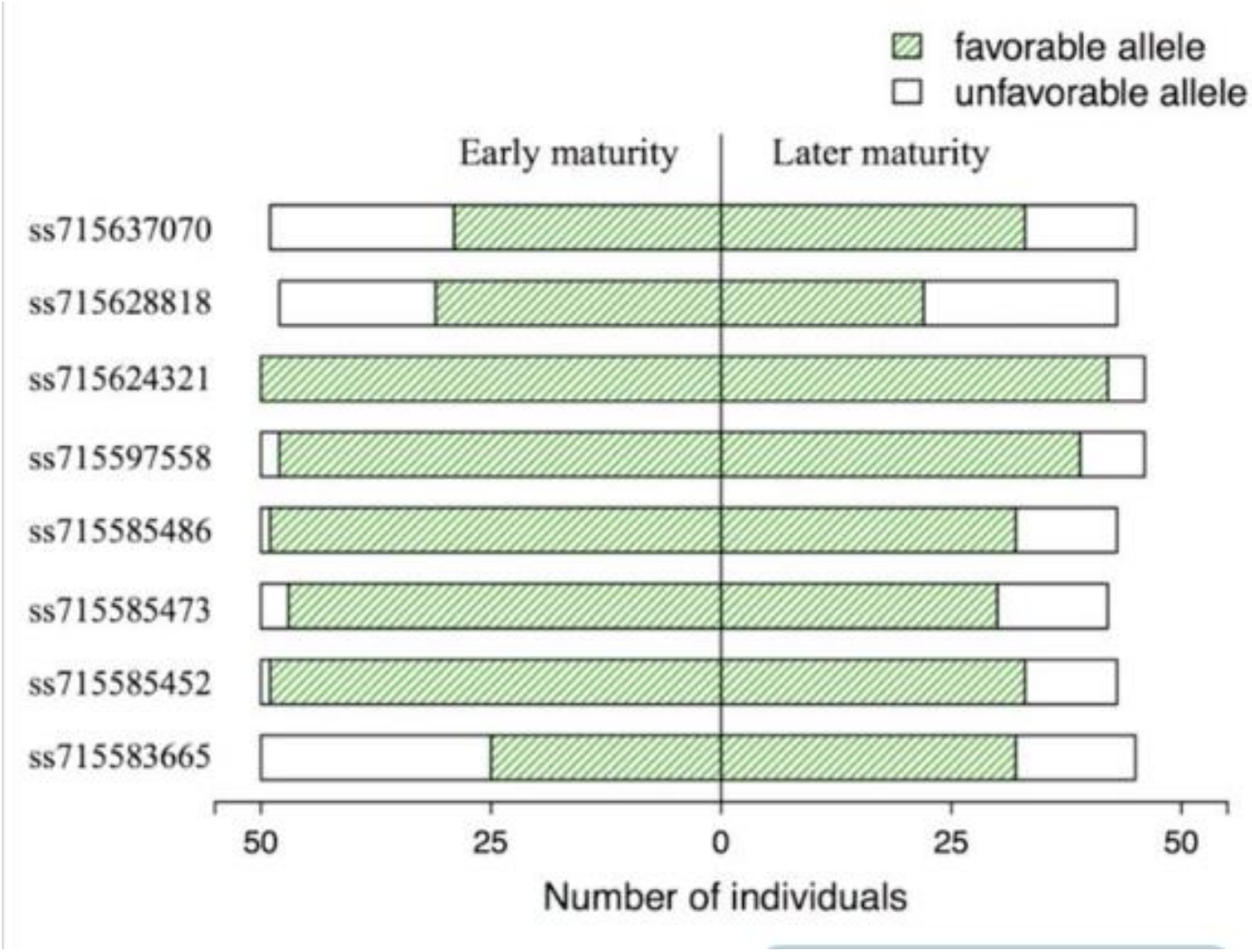
Resultados

- **Identificação dos principais QTL ligados em repulsão para o IDC**

Trait	SNP	Chr	Position	MAF	FDR	Allelic effect	R^2 (%)	Candidate gene	Annotation
ML-score	*ss715583665*	2	5855107	0.22	0.04	0.10	4.4		
	ss715585473	3	34547382	0.43	0.04	−0.08	3.5	*Glyma.03g130400 Glyma.03g130600*	bHLH[6] bHLH[6]
	*ss715585486**	3	34612476	0.23	0.00	0.15	8.0	*Glyma.03g131200 Glyma.03g131400 Glyma.03g131500*	2-oxoglutarate -Fe(II) oxygenase superfamily
	ss715624321	16	30708368	0.11	0.04	0.12	3.5		
ML-severity	*ss715583665*	2	5855107	0.22	0.04	0.97	3.4		
	*ss715585452**	3	34403919	0.31	0.00	1.43	9.1	*Glyma.03g128300 Glyma.03g129200*	Ferredoxin-dependent GLUS Cytochrome P450
	ss715585473	3	34547382	0.43	0.04	−0.77	3.4		
	ss715597558	7	37177741	0.20	0.04	0.90	3.5		
	ss715628818	18	13099563	0.37	0.04	0.87	3.5	*Glyma.18g111000*	AtF6'H1[25,26]
	ss715637070	20	248491	0.07	0.01	1.47	4.1		

Resultados

- **Estudo alélico no painel de cultivares de soja**
  - Foram utilizadas 96 cultivares elite
    - Status alélico dos locos identificados, especialmente os dois QTLs ligados, *ss715585452* e *ss715585473* em Gm03
  - Em cultivares precoces é maior a presença de ligação de atração com alelos favoráveis nos dois QTL ligados.

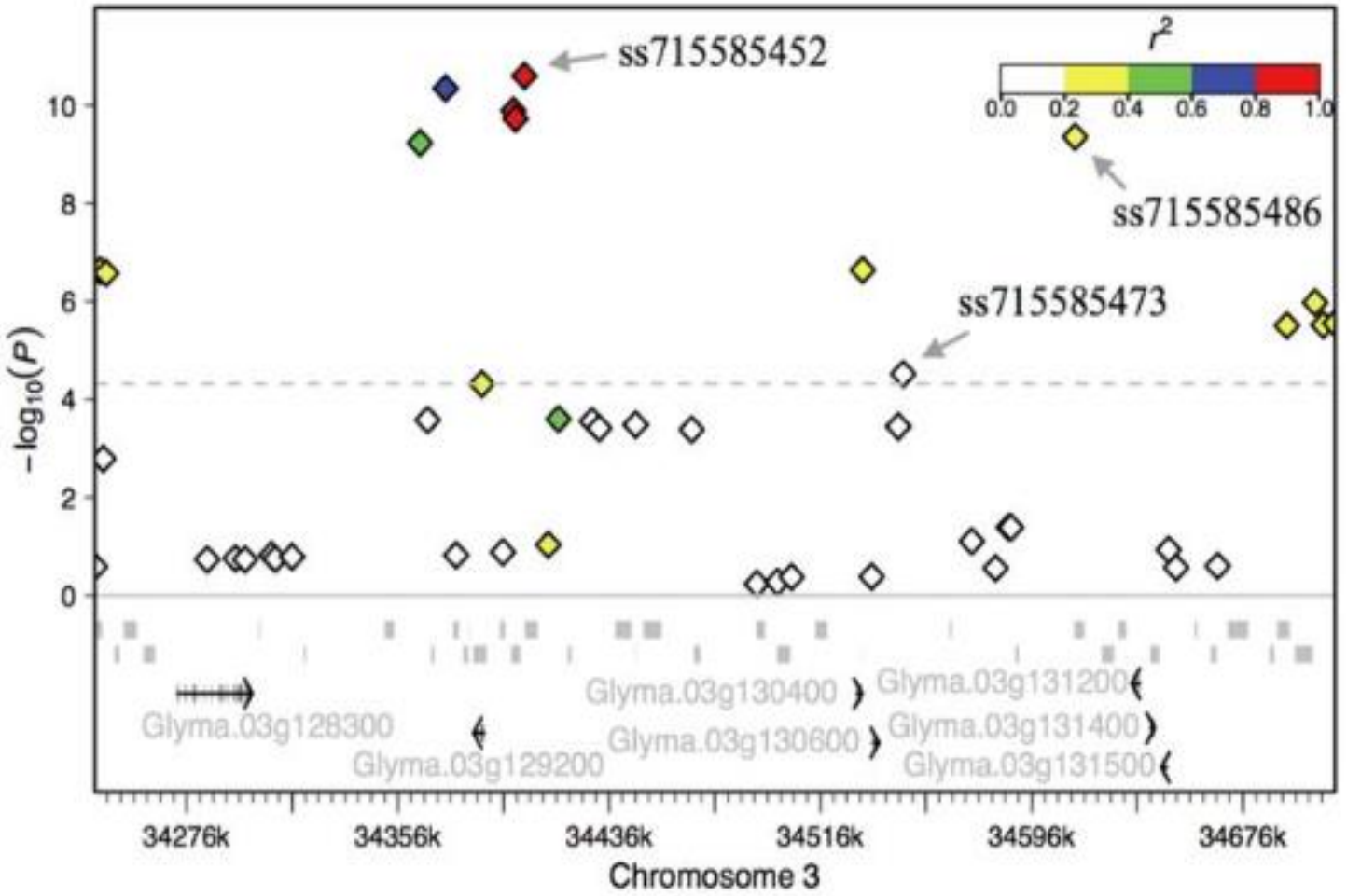


Resultados

- **Genes Candidatos**

Dentro da região de 847 kb em Gm03 associada aos marcadores observados em *Score* ML e *Severity* ML, foram identificados sete candidatos nas proximidades

Figure 3: The chromosomal region for loci associated with IDC tolerance on Gm03.



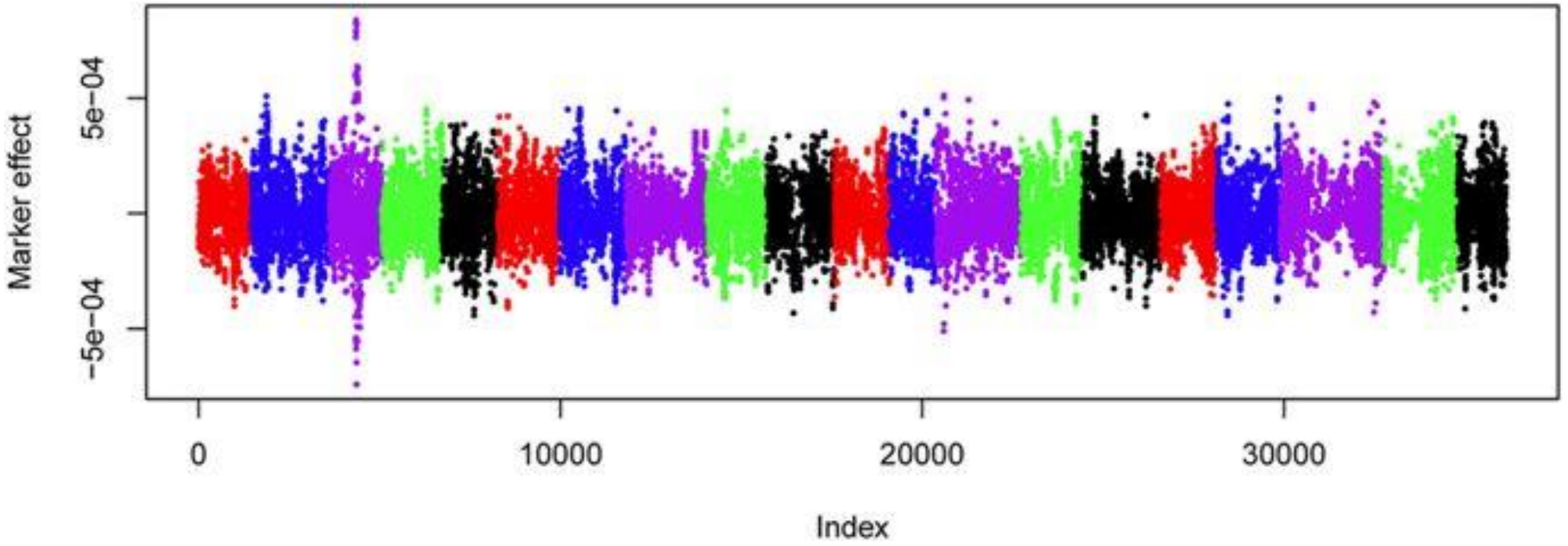
Resultados

- **Predição genômica**

Assumindo os SNPs como efeito aleatório

Modelo	RR-BLUP
Pontuação do ML	0,44
Gravidade do ML	0,37

Figure 4: Estimates of the marker effect in genomic prediction with rrBLUP.



Resultados

- **Predição genômica**

Table 2: Genomic prediction accuracies of different models developed for iron deficiency chlorosis in soybean using an association mapping panel.

Model	RR-BLUP	RR-BLUP with the major QTL and SPAD values as fixed effects ss715585452	SPAD	SPAD + ss715585452
ML-score	0.44	0.51	0.55	0.58
ML-severity	0.37	0.46	0.51	0.56

Discussão

- **Uso de Machine Learning**

  Automação do processo;

  - **Fluxo de trabalho integra:**

  Uso de imagens digitais do dossel das plantas;

  Pré-processamento automatizado das imagens;

  Obtenção de características dos sintomas da IDC;

  Classificação e geração de função de severidade baseada em ML.

Discussão

- **Uso de Machine Learning**

  Alta precisão de Predição > 0,95;

  - **GWAS valida o uso do ML na fenotipagem:**

  Identificação do principal QTL no Gm03, já relatado em vários trabalhos;

  Identificação de novo QTL intimamente ligado ao anterior;

  - **Complexidade genética da região envolvida**

  Independência entre os locos dos dois SNPs

  baixo LD (r^2 = 0,11);

Confirmado pela predição genômica

Efeito do principal QTL superestimado

Discussão

É desejável integrar a Highthrouput phenotiping e componentes ML no pipeline GP para maximizar o ganho genético

- **Uso de Machine Learning**

  Vantagens:

  - Maneira rápida e eficiente de obter avaliação do estresse abiótico;
  - Diferentes saídas: a classificação do IDC e função que resulta índice de severidade contínuo;
  - Diferentes abordagens são capazes de fornecer um conjunto mais rico de locos do que os descritores baseados em classes convencionais;

Conclusão

- O estudo confirma o poder do uso de técnicas de highthrouput phenotiping e machine learning para análises GWAS e GP para IDC em soja;
- Permite a observação da menor variância fenotípica da característica;
- Complexidade da base genética da resistência da IDC em soja;
- A pipeline de fenotipagem de imagens ativadas por ML apresentada neste estudo pode ser estendida a outras características e culturas;
- Promessa para acelerar o ganho genético do melhoramento de desenvolvimento de cultivares.

Referências

- ZHANG, Jiaoping et al. Computer vision and machine learning for robust phenotyping in genome-wide studies. **Scientific reports**, v. 7, p. 44048, 2017.

- Lin, S., Cianzio, S. & Shoemaker, R. Mapping genetic loci for iron deficiency chlorosis in soybean. Molecular Breeding 3, 219–229 (1997)

- Google Imagens.

Obrigada!!