Machine learning-based differential network analysis: A study of stress-responsive transcriptomes in *Arabidopsis*

Bruna Orsi

Roberta L. Vidal

Disciplina: Tópicos Especiais em Genética

The Plant Cell, Vol. 26: 520–537, February 2014, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

Machine Learning–Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in *Arabidopsis*[W]

Chuang Ma, Mingming Xin, Kenneth A. Feldmann, and Xiangfeng Wang[1]

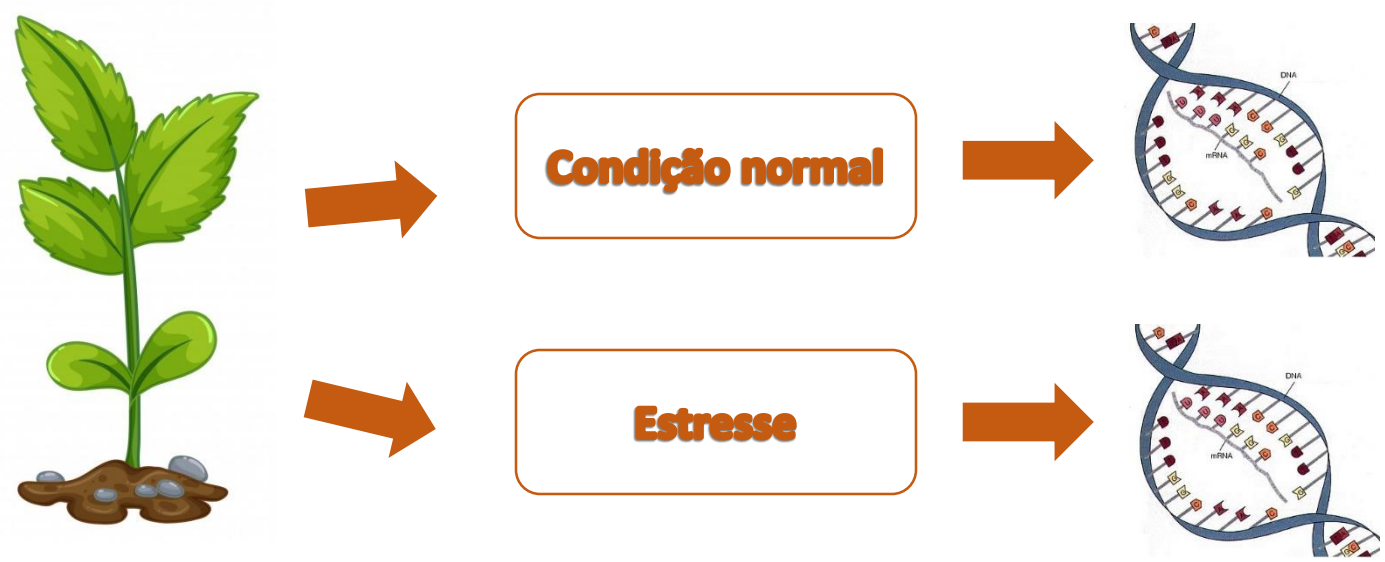
School of Plant Sciences, University of Arizona, Tucson, Arizona 85721-0036

ORCID IDs: 0000-0001-9612-7898 (C.M.); 0000-0002-3306-5594 (M.X.); 0000-0002-6406-5597 (X.W.)

Machine learning (ML) is an intelligent data mining technique that builds a prediction model based on the learning of prior knowledge to recognize patterns in large-scale data sets. We present an ML-based methodology for transcriptome analysis via comparison of gene coexpression networks, implemented as an R package called machine learning–based differential network analysis (mlDNA) and apply this method to reanalyze a set of abiotic stress expression data in *Arabidopsis thaliana*. The mlDNA first used a ML-based filtering process to remove nonexpressed, constitutively expressed, or non-stress-responsive "noninformative" genes prior to network construction, through learning the patterns of 32 expression characteristics of known stress-related genes. The retained "informative" genes were subsequently analyzed by ML-based network comparison to predict candidate stress-related genes showing expression and network differences between control and stress networks, based on 33 network topological characteristics. Comparative evaluation of the network-centric and gene-centric analytic methods showed that mlDNA substantially outperformed traditional statistical testing–based differential expression analysis at identifying stress-related genes, with markedly improved prediction accuracy. To experimentally validate the mlDNA predictions, we selected 89 candidates out of the 1784 predicted salt stress–related genes with available SALK T-DNA mutagenesis lines for phenotypic screening and identified two previously unreported genes, mutants of which showed salt-sensitive phenotypes.

Introdução

- Estudos do transcriptoma possibilitam identificar quais genes estão associados à determinadas condições e inferir sobre como eles se relacionam entre si



Introdução – Análise DE

- Abordagem tradicional: Análise de expressão gênica diferencial (DE)
- Abordagem analítica centrada no gene individualmente
- Testes de hipóteses (*t* test, F-teste ou ANOVA)

- Reduz o conjunto gênico a uma pequena lista de genes candidatos relacionados a condição

Introdução - Análise DE

- Limitações:

- Natureza da expressão gênica e consideração técnicas dos testes estatísticos ➡ A extensão na qual genes considerados importantes estão relacionados às condições permanece em aberto
- Fatores técnicos na metodologia de análise levam a resultados variados

Introdução – Análise DN

- Análise de *network* diferencial (DN):
- Abordagem analítica centrada em *networks*
- Detecta mudanças nas associações entre os genes pela comparação entre networks construídos em diferentes condições experimentais
- Análise complementar a DE
- Eficaz em detectar genes que possuem mudanças em sua expressão menos acentuadas

Introdução – Análise DN

- Muitos softwares foram desenvolvidos para análises DN, mas muitos problemas técnicos permanecem não resolvidos
- GCN – *Gene coexpression network*:

A conexão entre dois genes é baseada no coeficiente de correlação dos perfis de expressão → Reflexo de associação na função destes genes

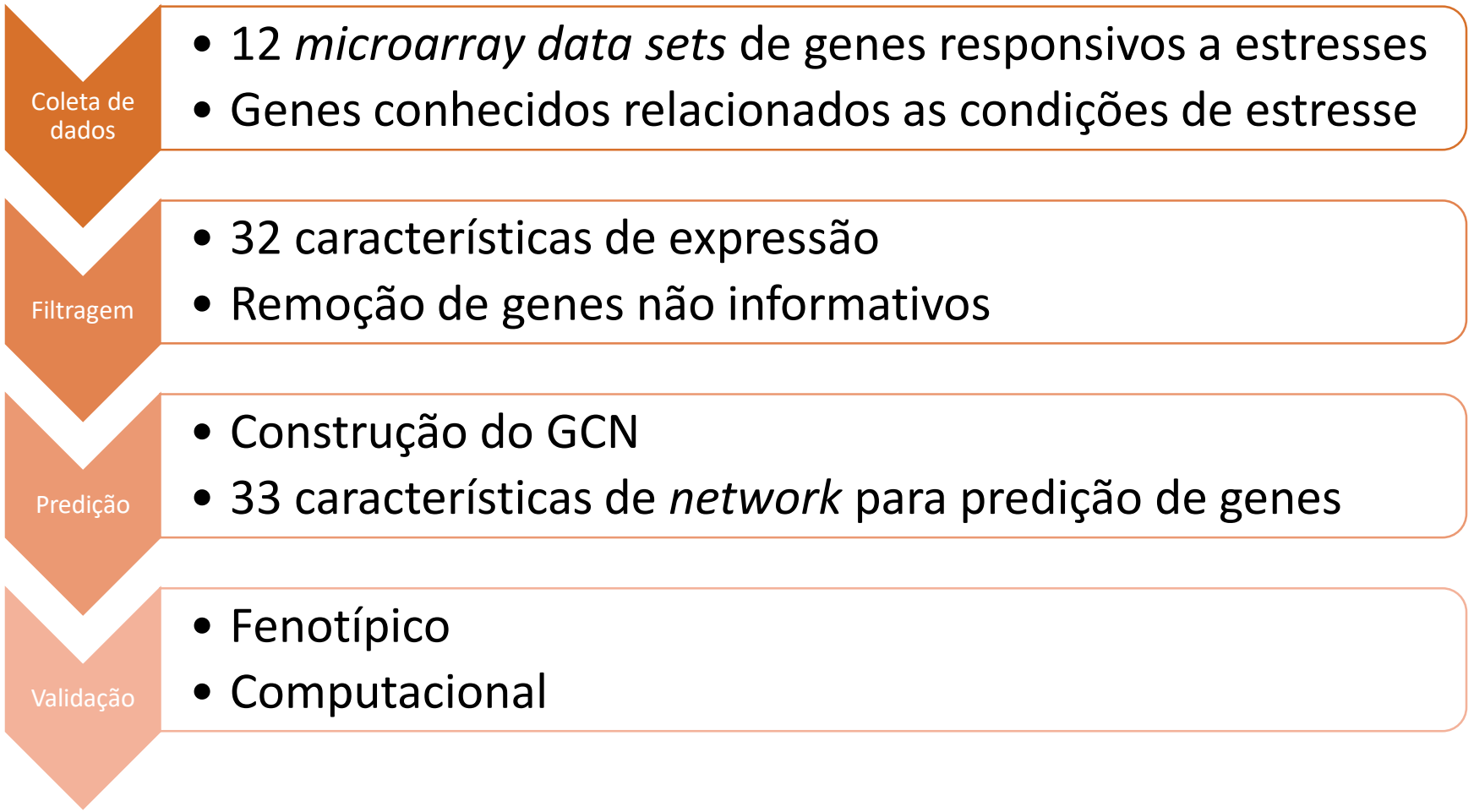
- Análises DN geralmente utilizam apenas uma característica de *network* para identificação de genes de interesse

Introdução

- mlDNA: Sistema computacional baseado em *machine learning* de análise transcriptônica centrada em network para identificação de genes relacionados à estresses

- *Machine learning* aplicado em dois momentos:
- Filtragem do conjunto de dados
- Predição de genes relacionados a estresses baseando-se em análises DN

Introdução


Coleta de dados
• 12 *microarray data sets* de genes responsivos a estresses
• Genes conhecidos relacionados as condições de estresse
Filtragem
• 32 características de expressão
• Remoção de genes não informativos
Predição
• Construção do GCN
• 33 características de *network* para predição de genes
Validação
• Fenotípico
• Computacional

Coleta de dados

- Obtenção do *data set*:
- AtGenExpress *database* (Killian et al., 2007)
- 12 *microarray datasets* de brotos e raízes de *Arabidopsis thaliana*
- Níveis expressão de 22.591 genes sobre as condições de controle e estresses salino, por frio, por seca, por ferimento, por calor e genotóxico obtidos em seis tempos (0.5, 1, 3, 6, 12 e 24h)

Coleta de dados

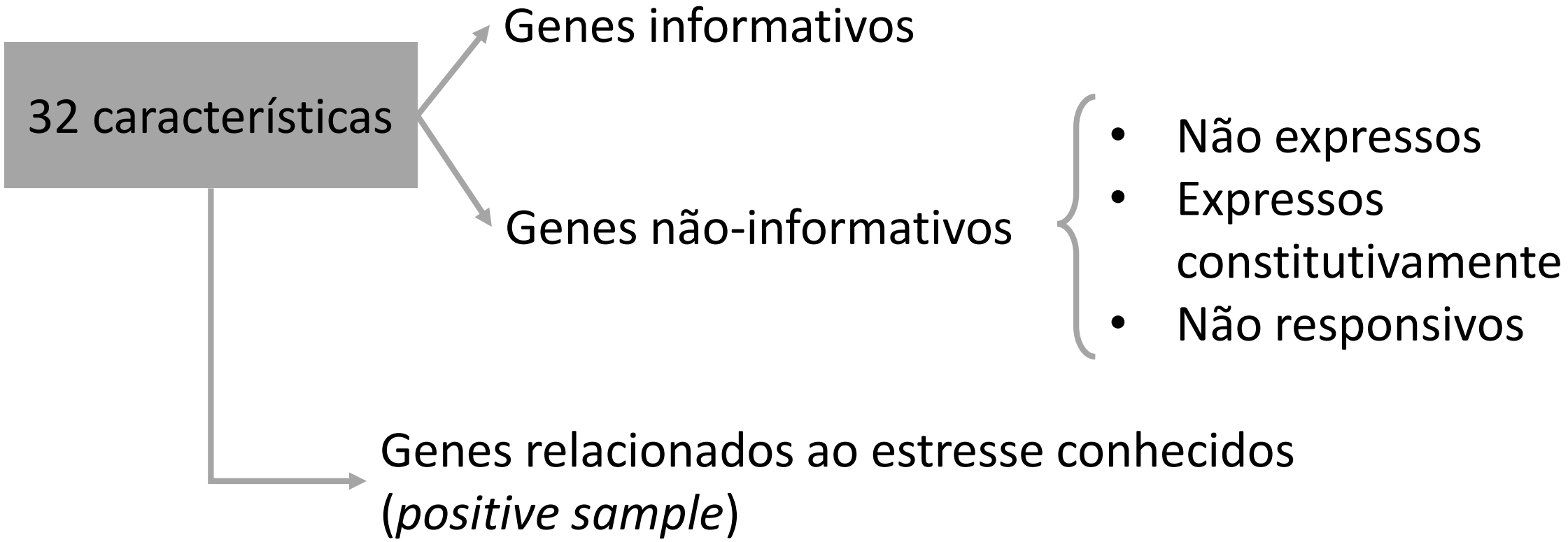
- Amostras positivas:

- Genes conhecidos relacionados a cada estresse compilados do TAIR 10 e DRASTIC *data bases*

Estresse	**Número de genes**
Salino	895
Frio	433
Seca	394
Ferimento	357
Calor	46
Genotóxico	42

Pré-seleção baseada em ML

- Filtragem de genes “informativos” para uso em GCN



Pré-seleção baseada em ML

- Processo de filtragem:

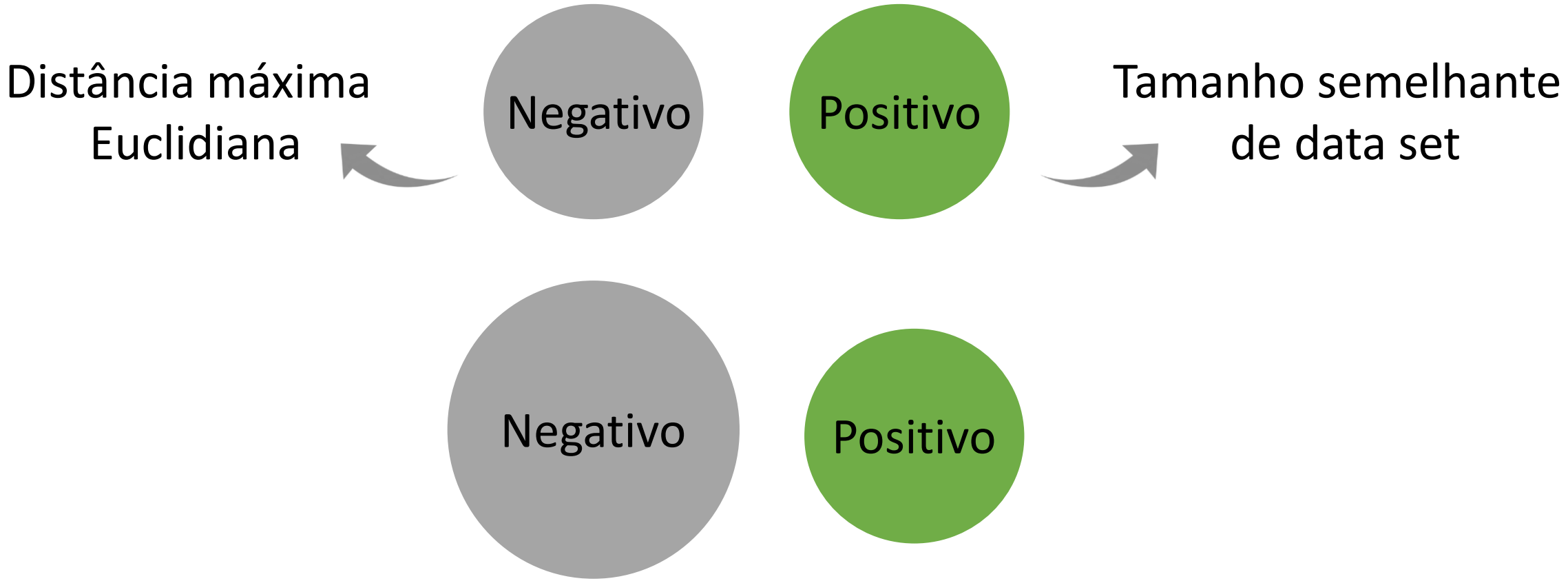
- Classificador: Random forest (RF)
- Algoritmo: Positive Sample-only Learning (PSOL) (Wang et al., 2006)

Matriz de características

- 12 características de expressão absoluta
- 12 características medidas em z-scores
- 6 características de *fold-changes*
- 2 características de coeficiente de variação

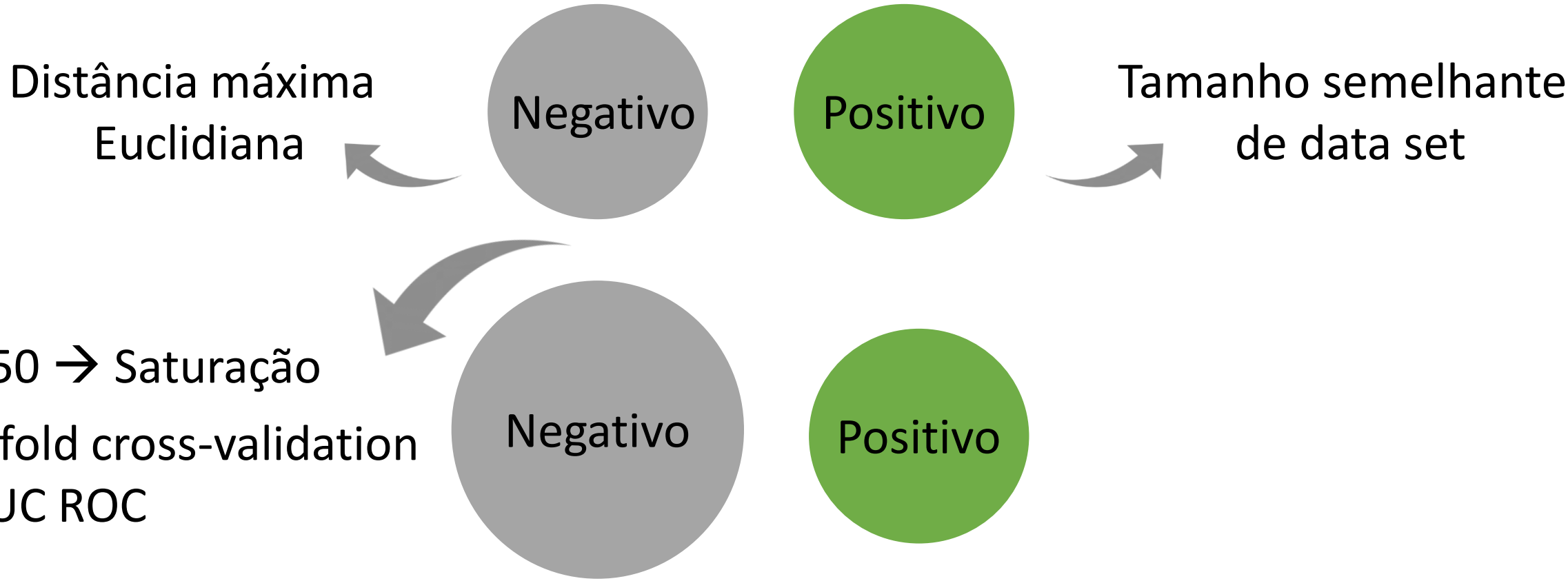
Pré-seleção baseada em ML

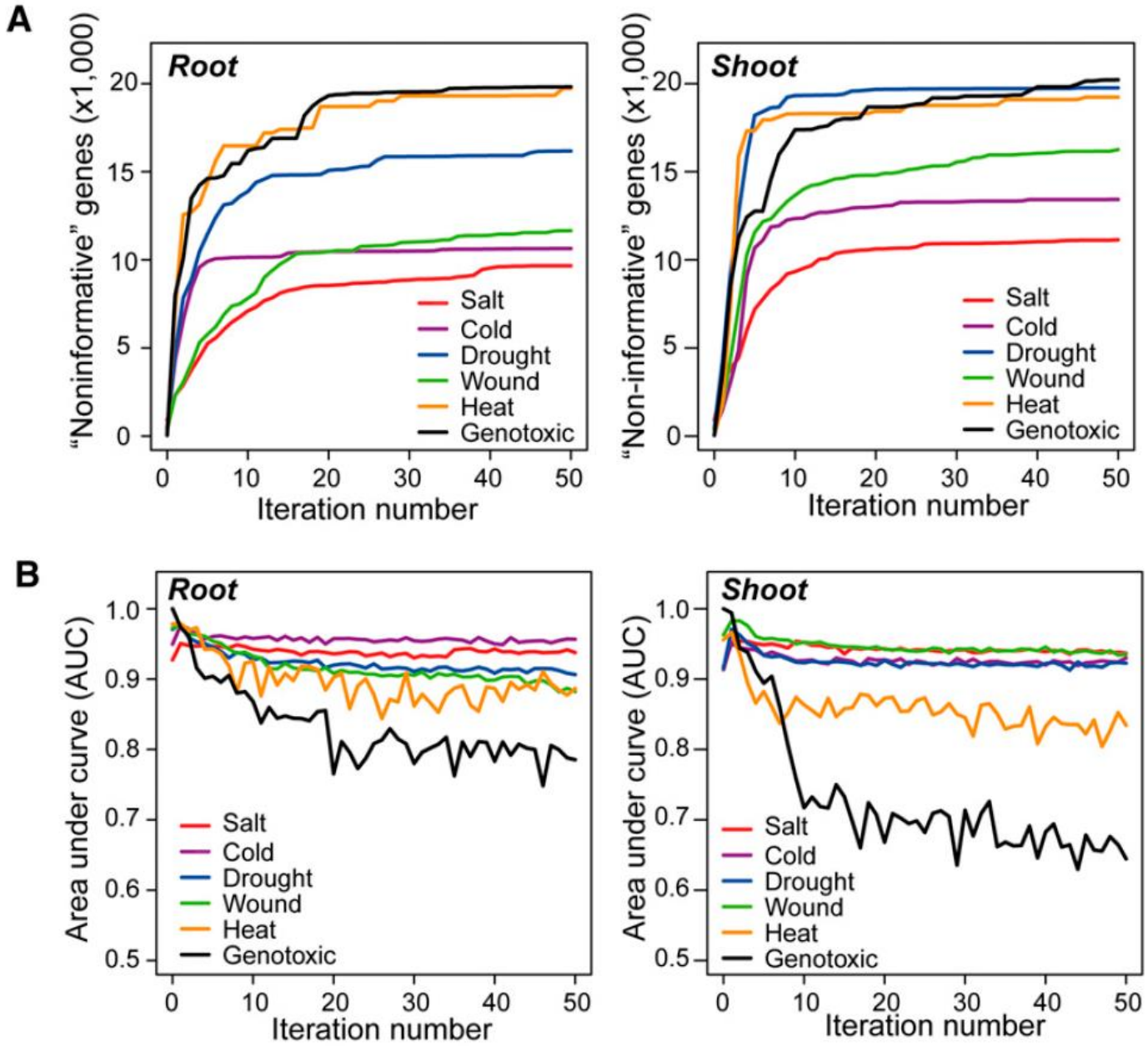
- Classificador foi rodado repetidas vezes para remover os genes “não informativos”.



Pré-seleção baseada em ML

- Classificador foi rodado repetidas vezes para remover os genes “não informativos”.

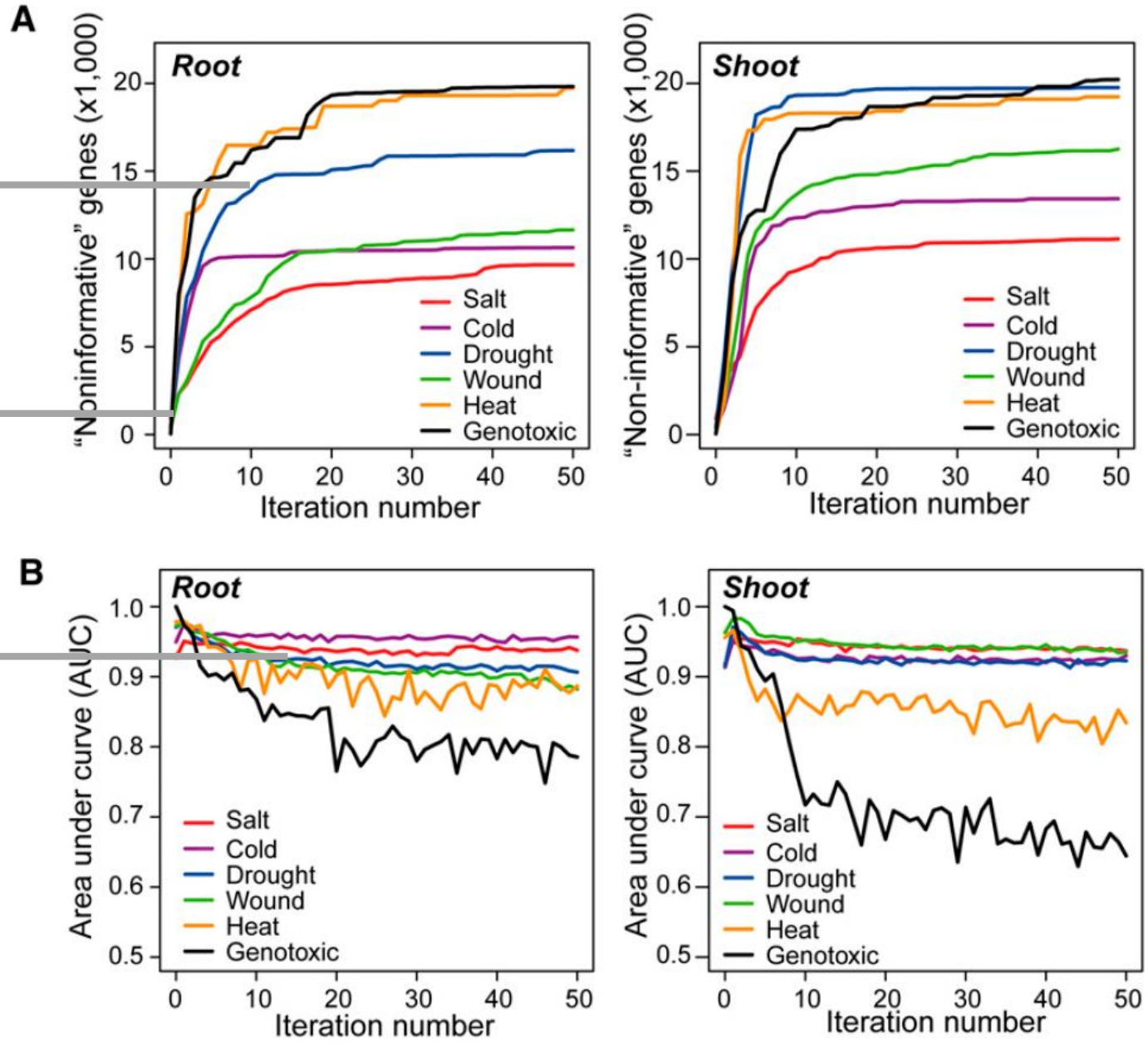



A
Root
Shoot
"Noninformative" genes (x1,000)
"Non-informative" genes (x1,000)
Iteration number
Salt
Cold
Drought
Wound
Heat
Genotoxic
B
Area under curve (AUC)

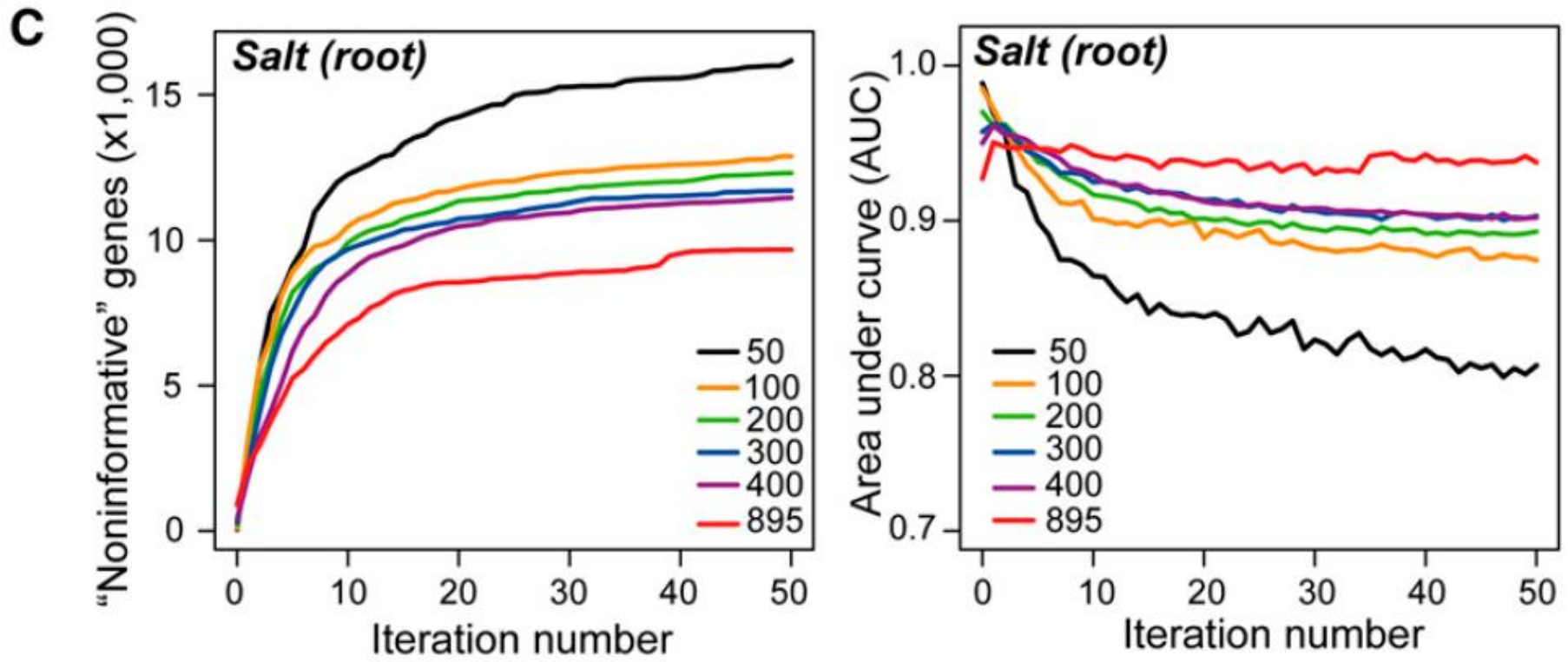
14.790 genes

397 genes
Data set dos
negativos

Ponto de
saturação

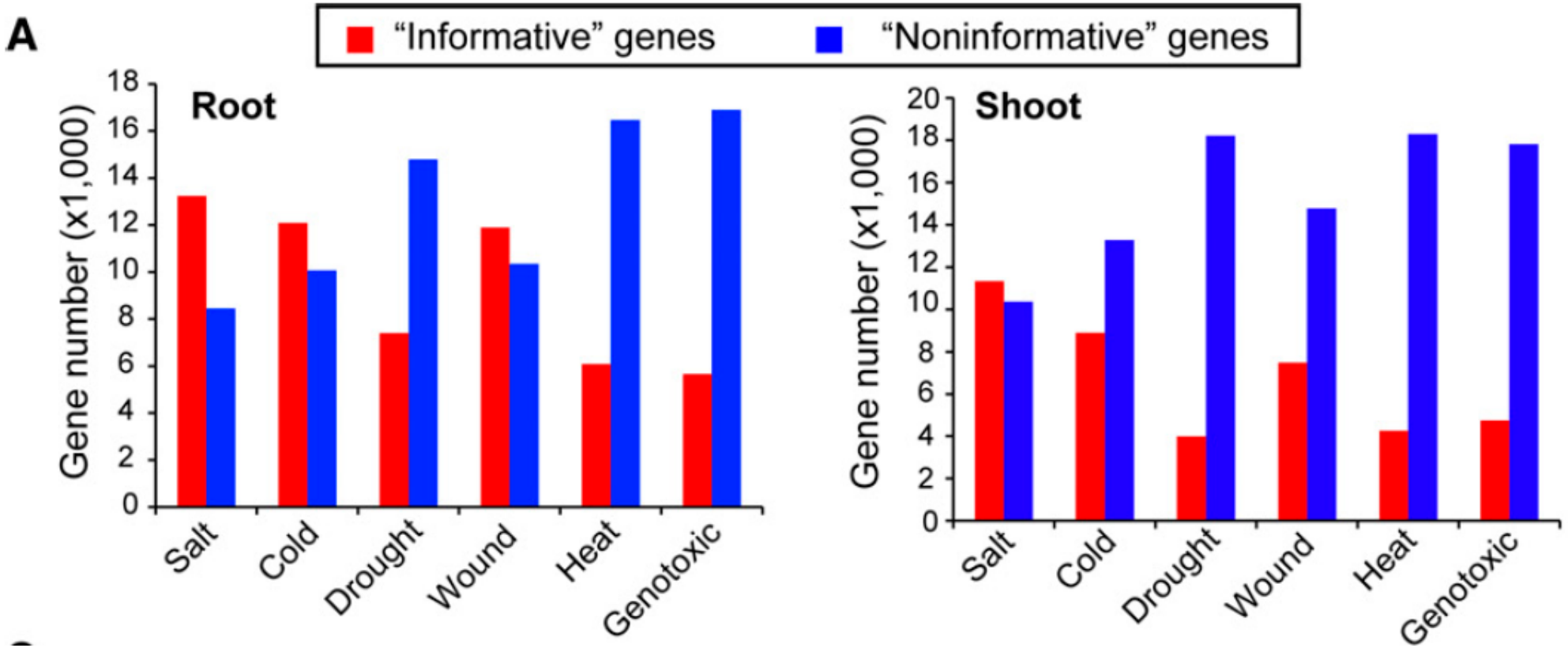


Data set de
Positivos
pequeno



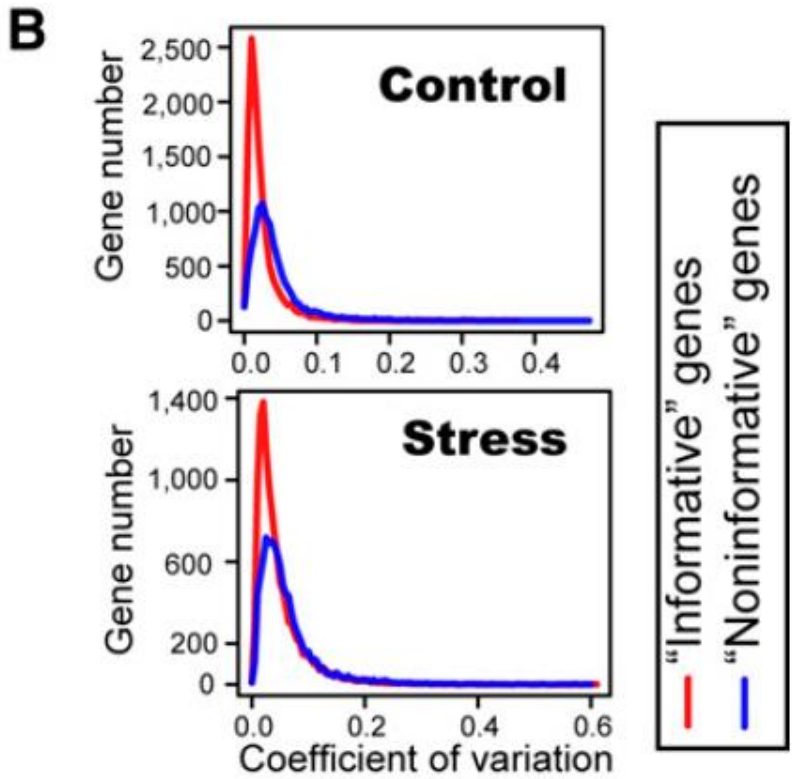
Ao menos 100 genes no data set de Positivos são necessários

Características de expressão



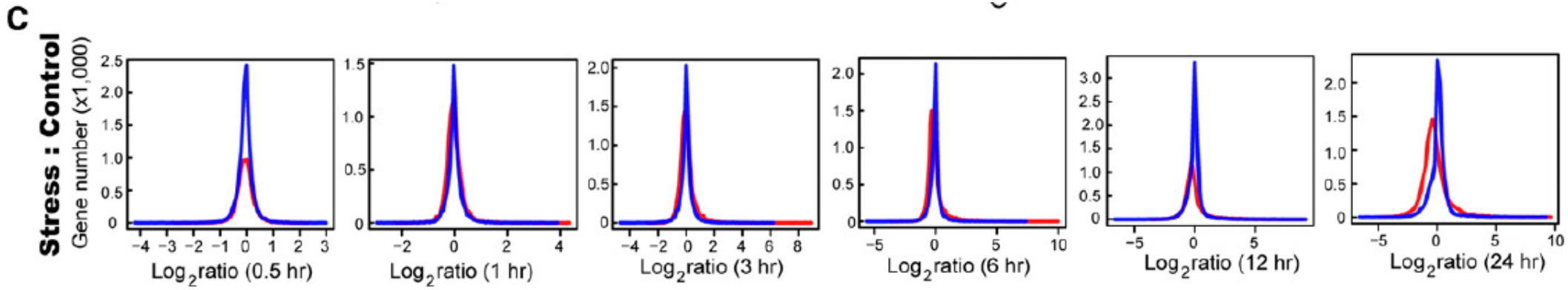
Características de expressão

- Validação do poder discriminatório das 32 características



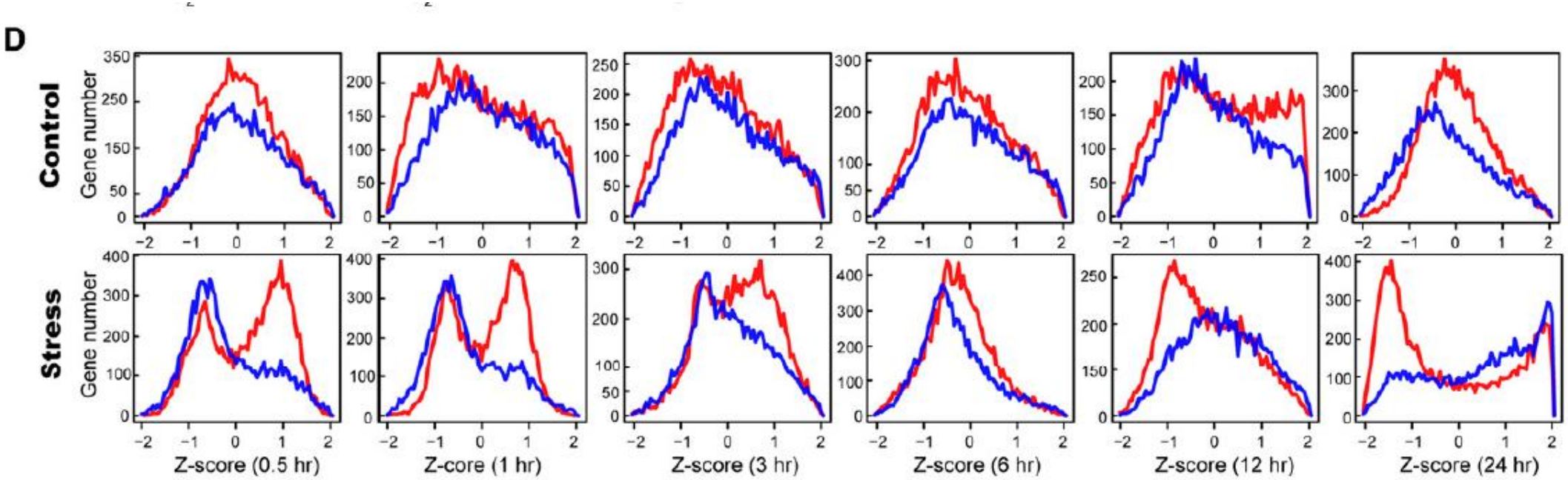
Características de coeficiente de variação

Características de expressão



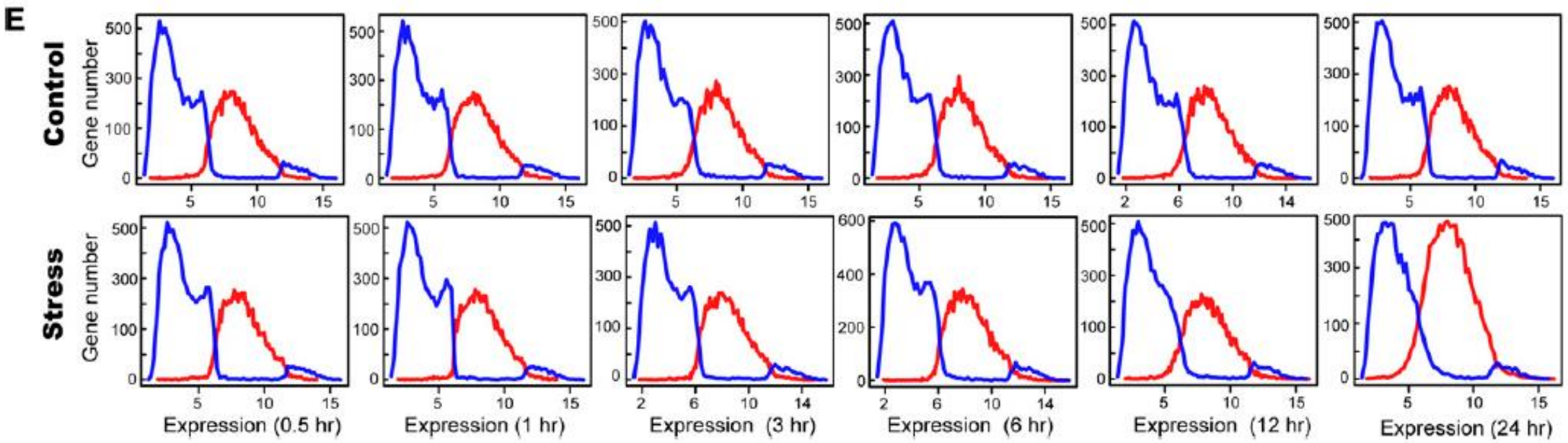
Características de *fold-changes*

Características de expressão



Características de z-score

Características de expressão



Características de expressão absoluta

Análise DN baseada em ML

- Filtragem resultou em torno de 4 a 14 mil genes informativos:

-Atividades fisiológicas
-Respostas induzidas por estresses

- Construção de *networks* para cada condição de estresse e condição de controle
- As correlações entre dois *networks* foi estabelecida preferencialmente pelo método GCC

Análise DN baseada em ML

- GCC: Gini correlation coefficient
- Utiliza tanto o Rank quanto os valores no calculo da correlação
- Vantagens:

- Independente da forma de distribuição dos dados
- Pouco influenciado por *outliers*
- Independente do tamanho da amostra
- Maior sensibilidade em detectar relações regulatórias momentâneas

Análise DN baseada em ML - mlDNA

- Matriz de características com 33 características de *network*, das quais:

- 10 características de genes nas condições de controle, estresse e na diferença entre as duas condições (3x10 = 30):

a) 7 características descrevendo as propriedades centrais do gene no network (Nível, conexão positiva, conexão negativa, etc.) (7x3 = 21)

b) 3 características descrevendo a relação entre o gene com genes conhecidos relacionados ao estresse (3x3 = 9)

- 2 características não centrais (ASC e corDistance)

- 1 característica de expressão (expDistance)

Análise DN baseada em ML - mlDNA

- Análises DN tradicionais geralmente utilizam apenas uma característica de *network*
- Sistema mlDNA permitiu a utilização simultânea de 33 características na predição de genes relacionados ao estresse:

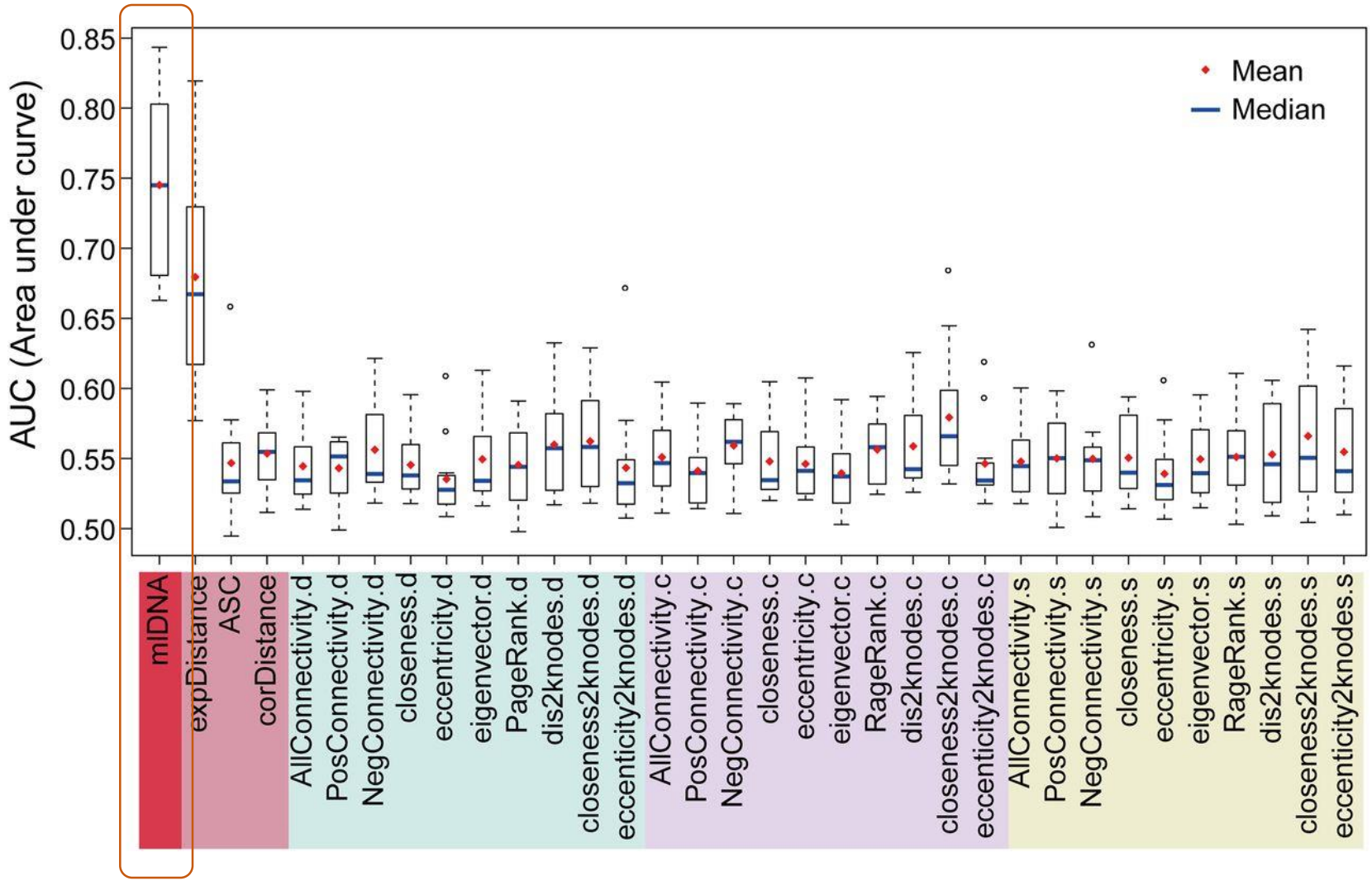
-Mudança no network
-Mudanças de relacionamento
-Mudanças na expressão

Análise DN baseada em ML - mlDNA

- Análise comparativa da performance nas predições de genes relacionados à estresses entre a utilização de apenas uma característica e as 33 simultaneamente
- Análise ROC dos resultados preditos por abordagem:

- Método 5-fold cross-validation

Análise DN baseada em ML - mlDNA

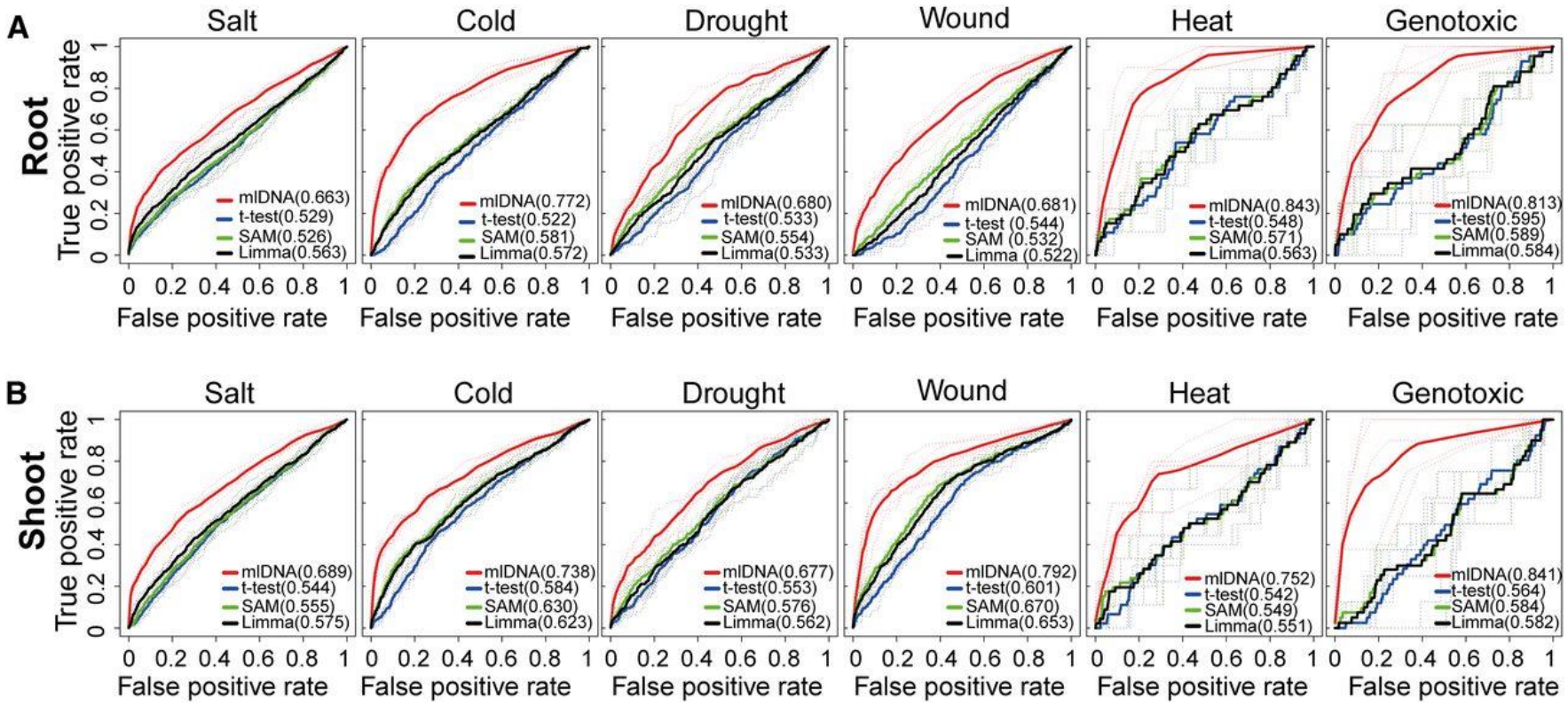


Análise DN baseada em ML - mlDNA

- Análise comparativa da performance do sistema mlDNA e três métodos convencionais de DE (*t* test, Limma e SAM)
- Análise ROC dos resultados preditos por cada abordagem:

- Método 5-fold cross-validation

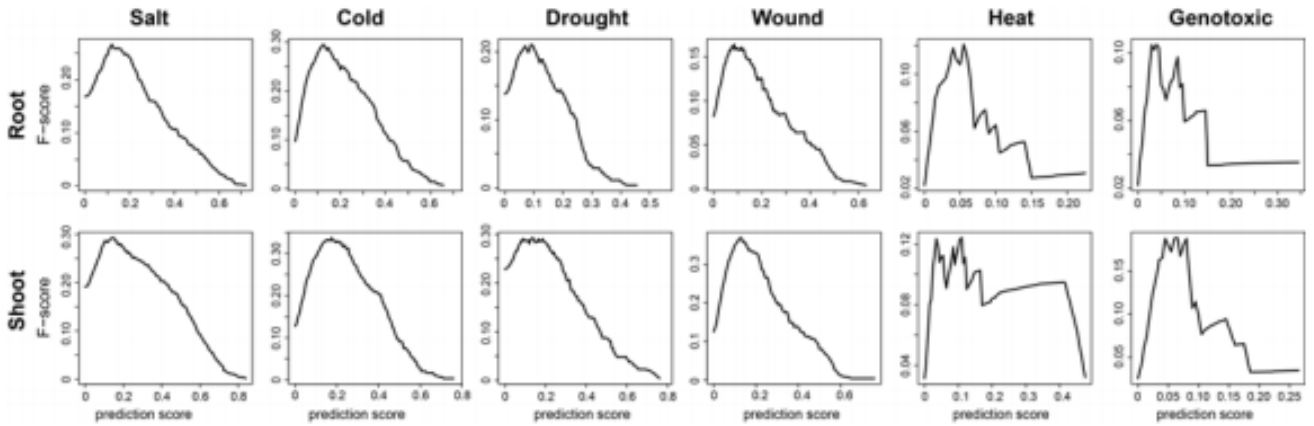
Análise DN baseada em ML - mlDNA



Análise DN baseada em ML – mlDNA: Predição de genes de interesse

- Método F-score para determinação do *score* ótimo para o classificador RF

Supplemental Figure 3. Determination of the Prediction Threshold of the RF Classification Model Using the F-score Algorithm.

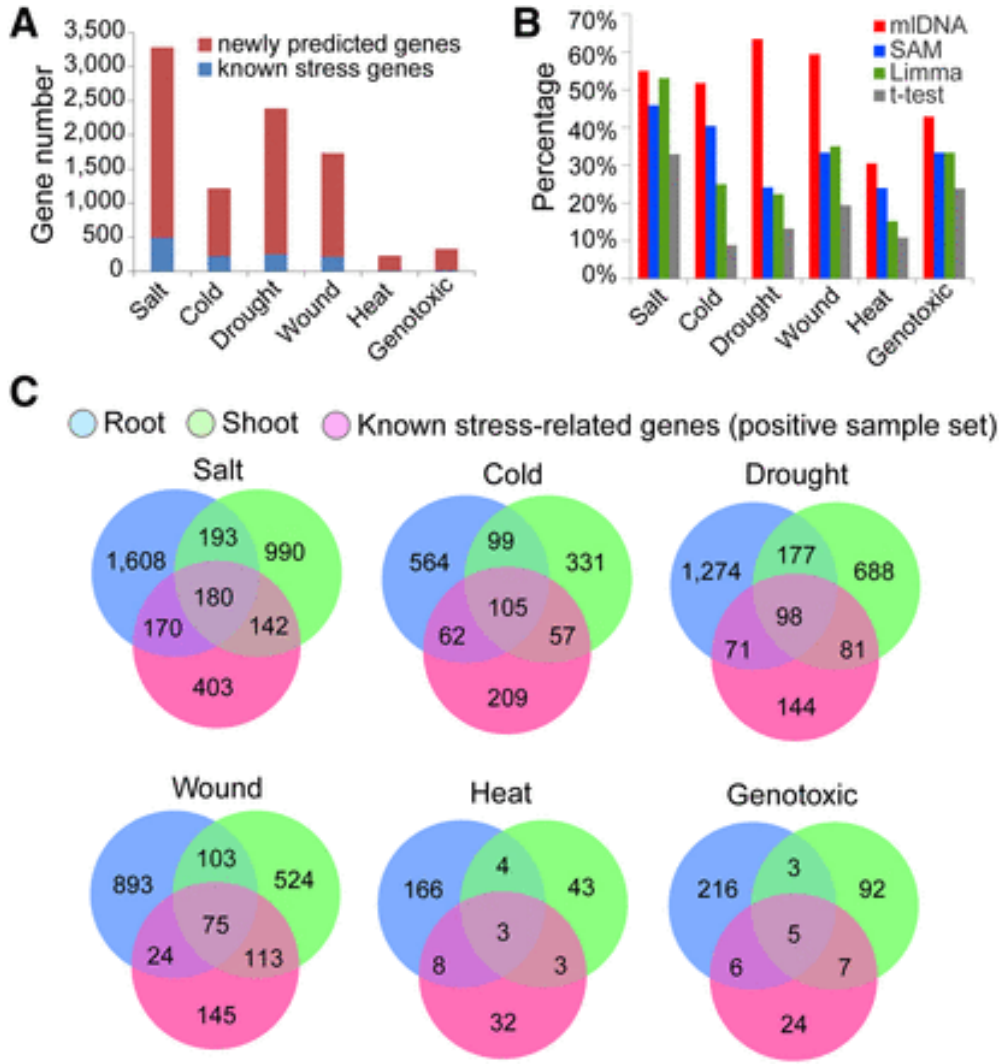


The optimal prediction score was selected when the F-score reaches the maximal value.

Análise DN baseada em ML – mlDNA: Predição de genes de interesse

- O sistema mlDNA identificou 3.283 (salino), 1.218 (frio), 2.389 (seca), 1.732 (ferimento), 227 (calor), 329 (genotóxico) genes candidatos relacionados a estresses nas raízes e brotos
- Novos genes foram maioria

Análise DN baseada em ML – mlDNA: Predição de genes de interesse



Análise DN baseada em ML – mlDNA: Predição de genes de interesse

- Comparação entre os genes de estresse salino e por frio preditos pelo sistema mlDNA e análises DE e genes encontradas no *screening* feito por Luhua et al. (2013):

- 16 genes relacionados a estresse salino e 13 relacionados a estresse por frio não foram identificados por análises DE
- 83 genes foram identificados por Luhua et al. (2013) e pelo sistema mlDNA, dos quais 19 eram conhecidos e 64 desconhecidos

- 75 a 85% dos genes identificados eram específicos do tipo de tecido

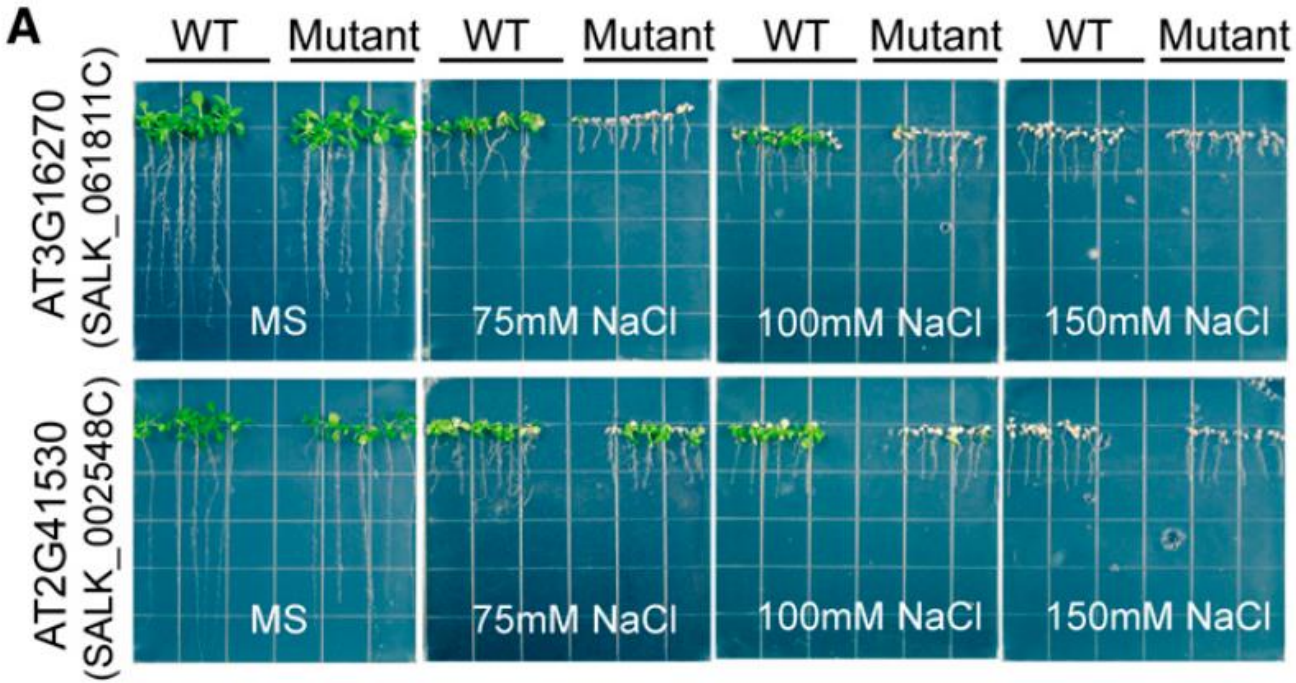
Screen fenotípico – validação experimental

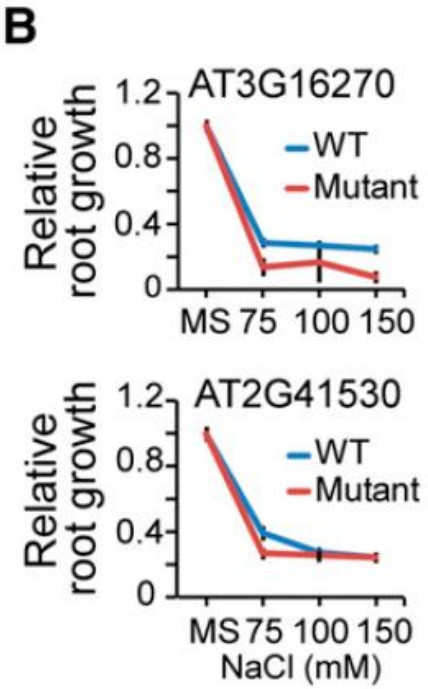
- Papel funcional dos genes relacionados ao estresse, preditos pelo mlDNA

- 89 candidatos de genes relacionados ao estresse salino, com uma linhagem T-DNA *knockout* correspondente e homozigota (Alonso et al., 2003).

- Plântulas transferidas para regimes de 0, 75, 100 e 150 mM NaCl

Screen fenotípico – validação experimental





Proteína com domínio VHS

S-formilglutationa hidrolase

Enriquecimento – validação computacional

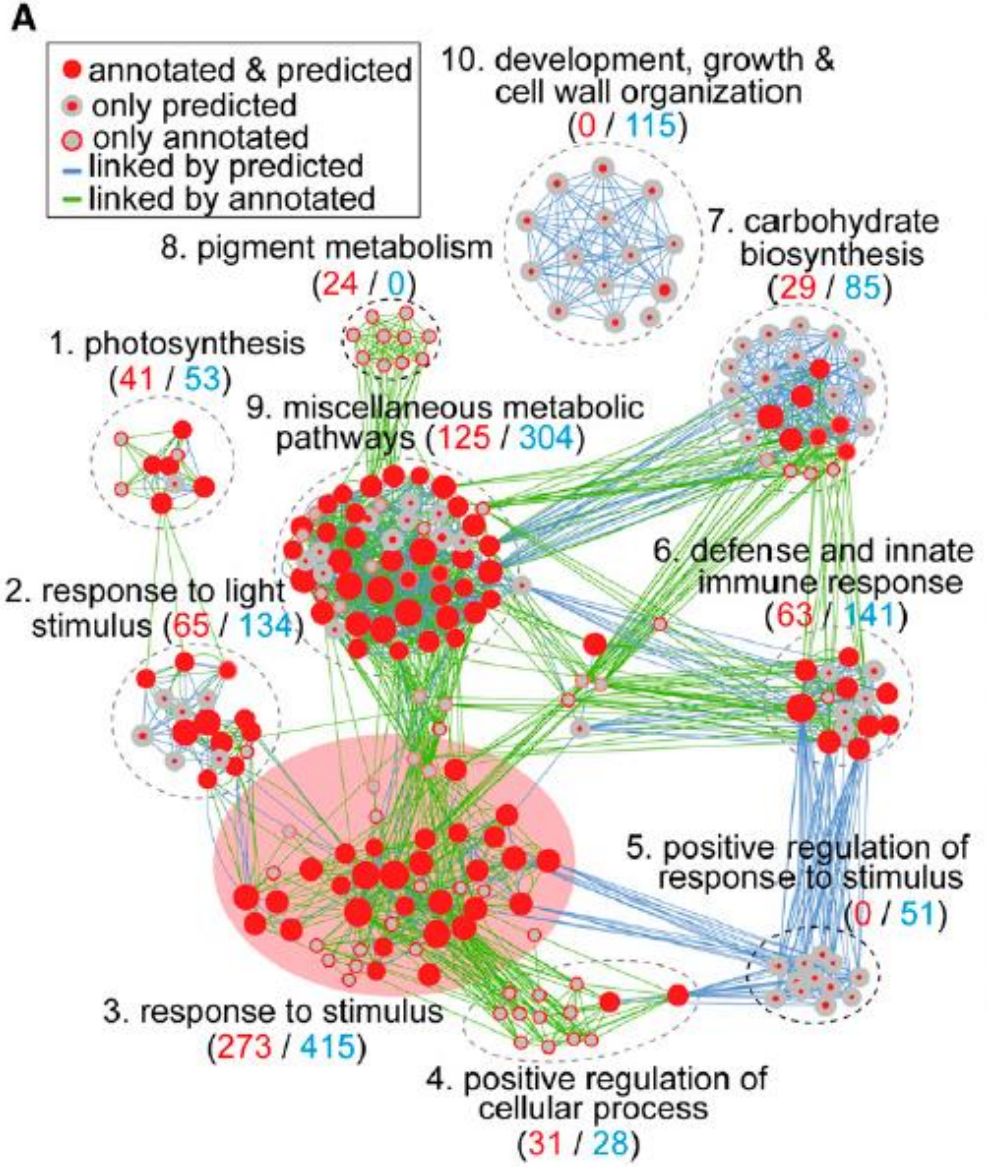
- Análise de enriquecimento GO → comparar as categorias GO dos genes relacionado ao estresse conhecidos e dos preditos

- Teste hipergeométrico → para detectar categorias GO com enriquecimento significativo

- Mapa de enriquecimento → para visualizar as categorias GO como um *network*

Enriquecimento – validação computacional

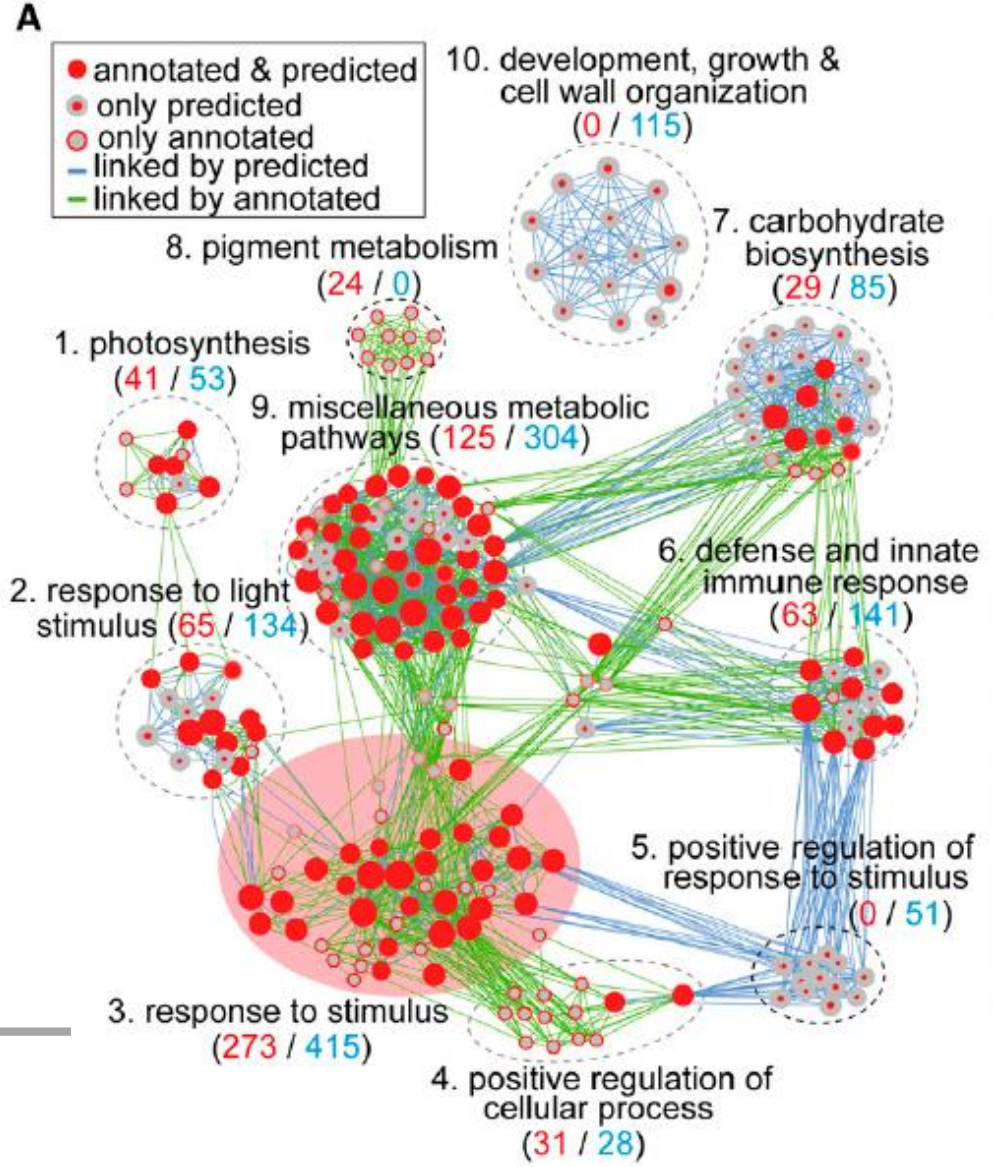
- 1061 genes relacionados ao estresse salino foram enriquecidos nas categorias GO → 10 módulos



Enriquecimento – validação computacional

- 1061 genes relacionados ao estresse salino foram enriquecidos nas categorias GO → 10 módulos

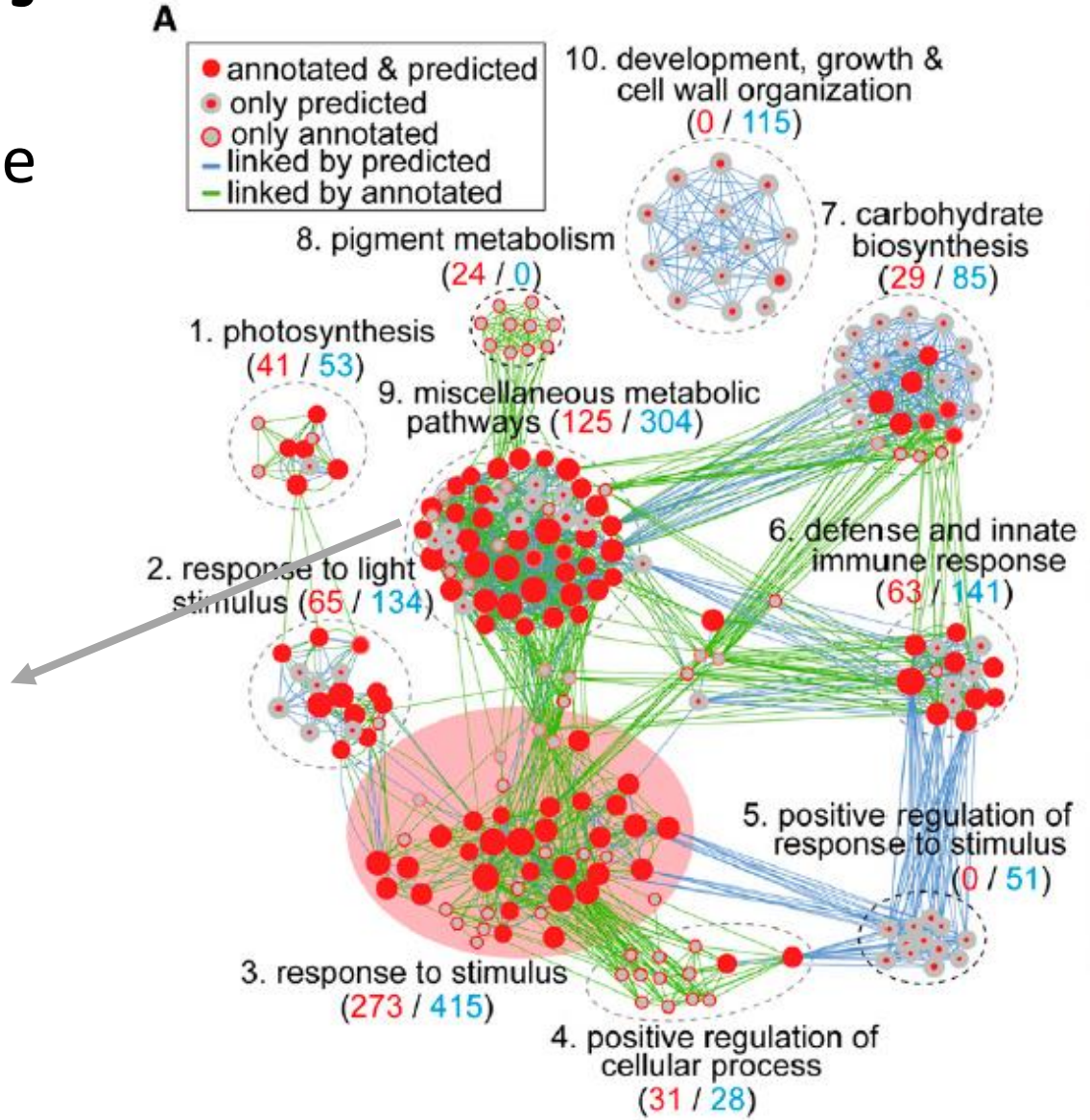
Maior categoria:
Inclui “resposta a estímulos hormonais”, “resposta a substancias inorganicas” e “respostas a estresse abiótico”



Enriquecimento – validação computacional

- 1061 genes relacionados ao estresse salino foram enriquecidos nas categorias GO → 10 módulos

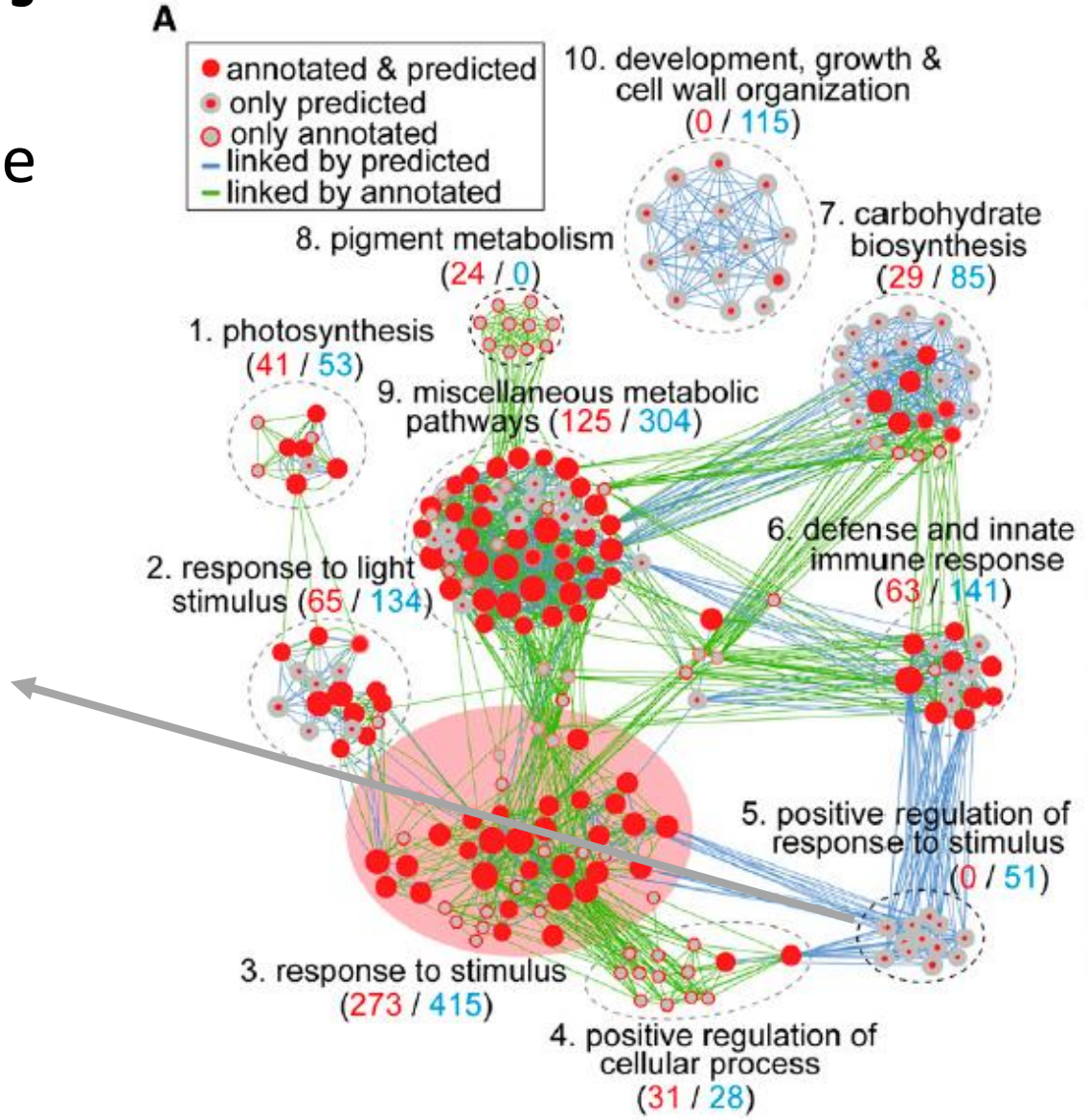
2ª maior categoria: Hormônios, vitaminas, ácidos graxos



Enriquecimento – validação computacional

- 1061 genes relacionados ao estresse salino foram enriquecidos nas categorias GO → 10 módulos

Apenas preditos por mlDNA: não envolvidos com as respostas primárias

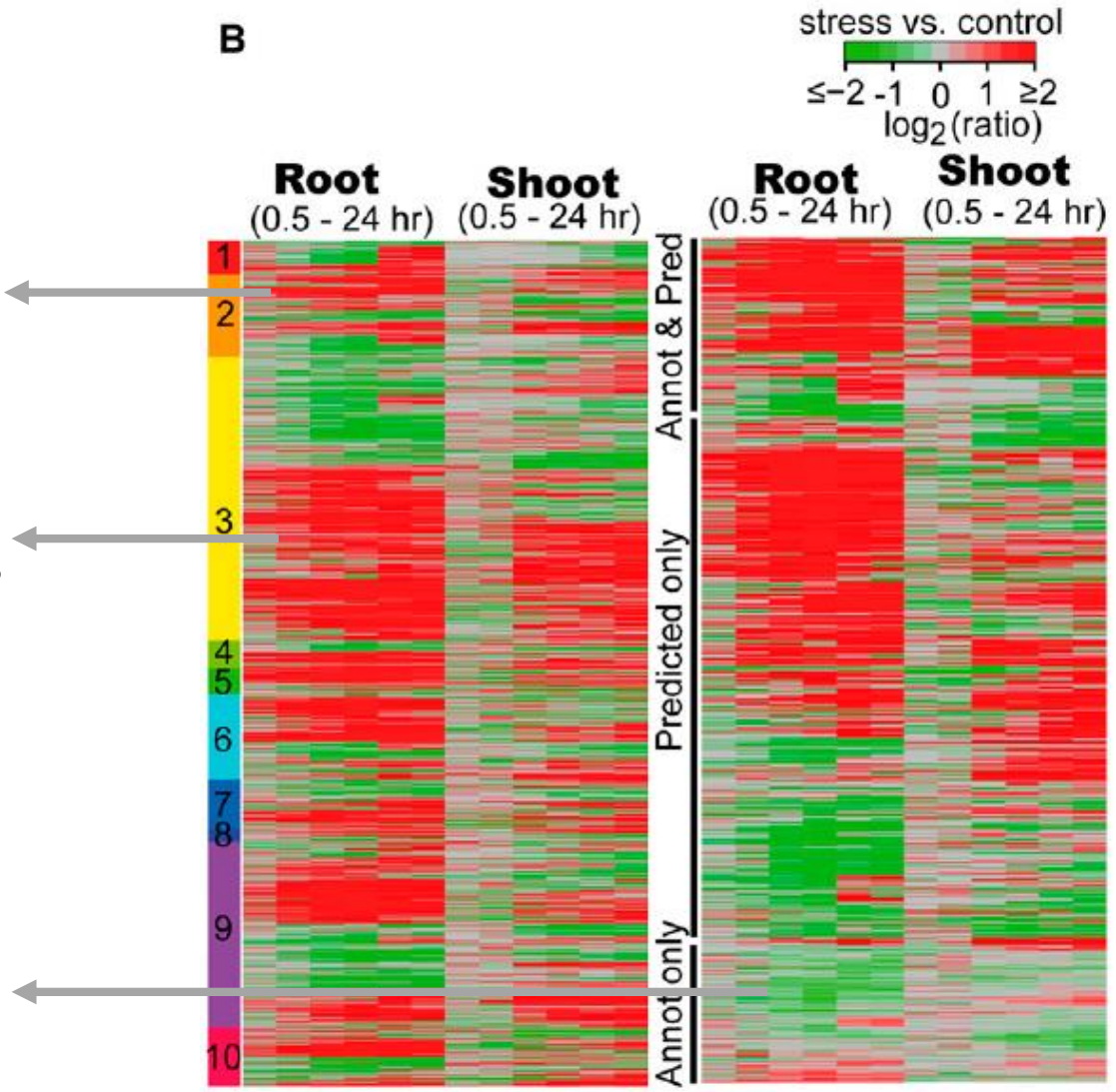


Enriquecimento – validação computacional

Respostas mais rápidas nas raízes

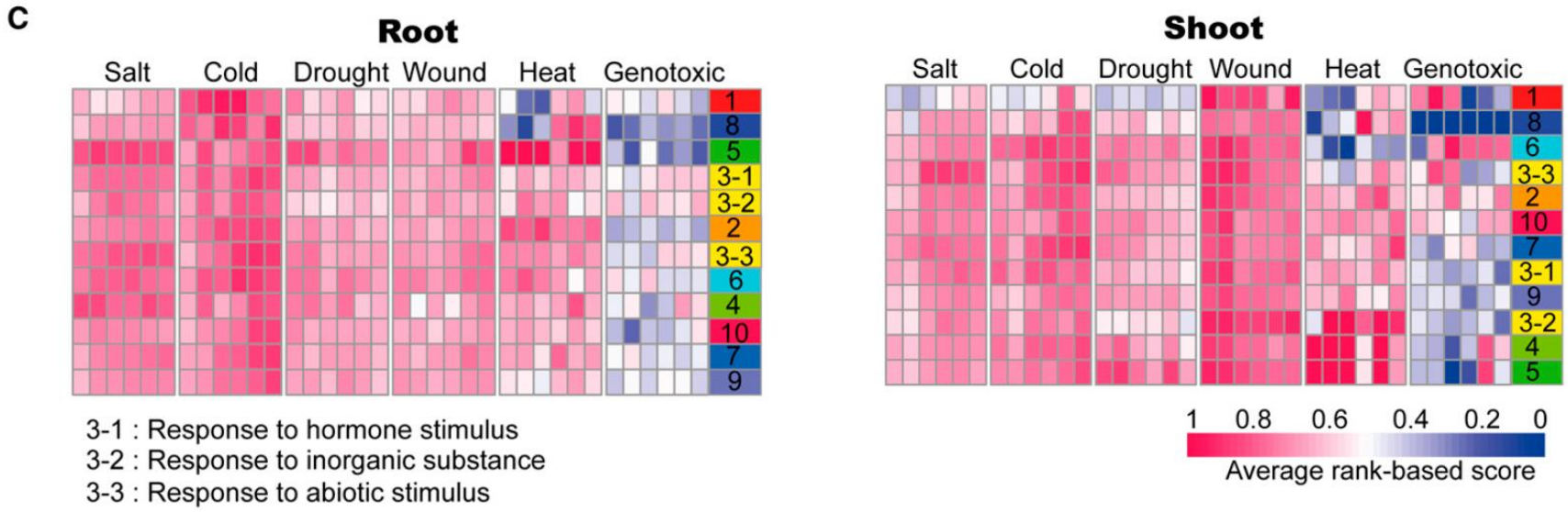
Respostas a estímulos também são mais rápidas

Genes não preditos tiveram baixa modificação



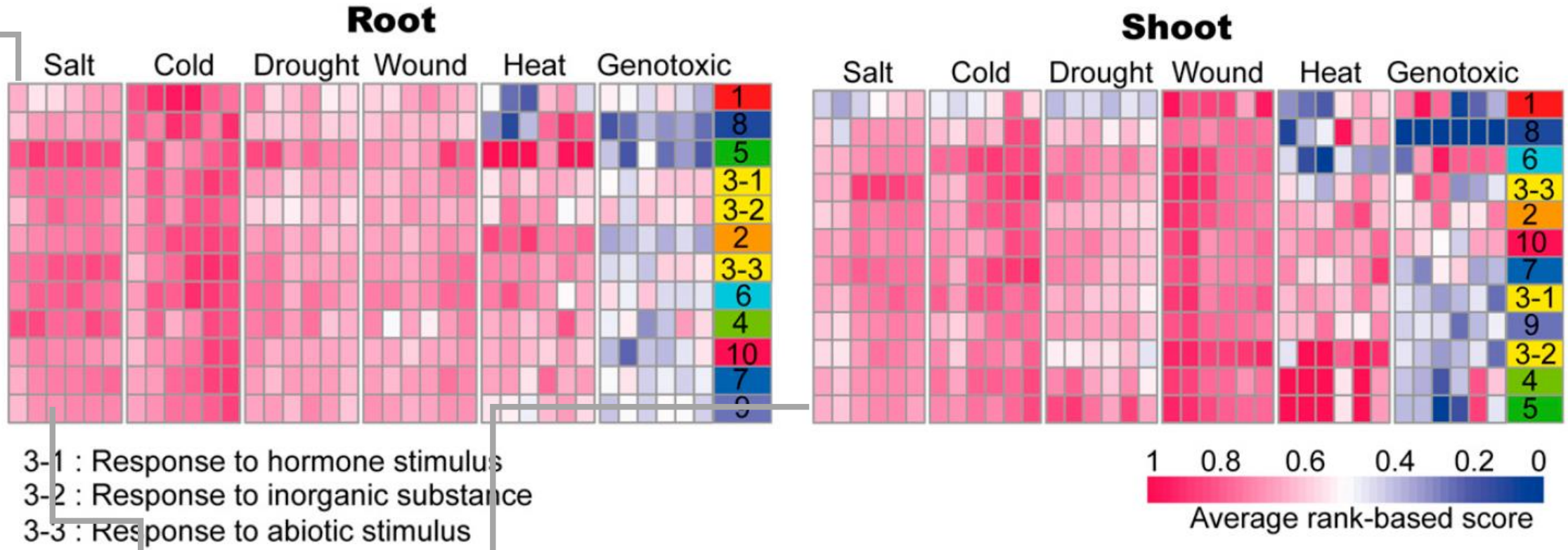
Diferentes respostas das vias biológicas

- Análise das mudanças de expressão a nível de via biológica, baseado nos módulos GO
- Assumiram que a regulação global de um módulo poderia ser refletida por mudanças na expressão de todos os genes de um módulo



0,5 h: Módulos 4 (regulação positiva de processos celulares), 5 (Regulação positiva de respostas a estímulos), 6 (Defesa e resposta imune inata) e 3-3 (Resposta a estímulo abiótico)

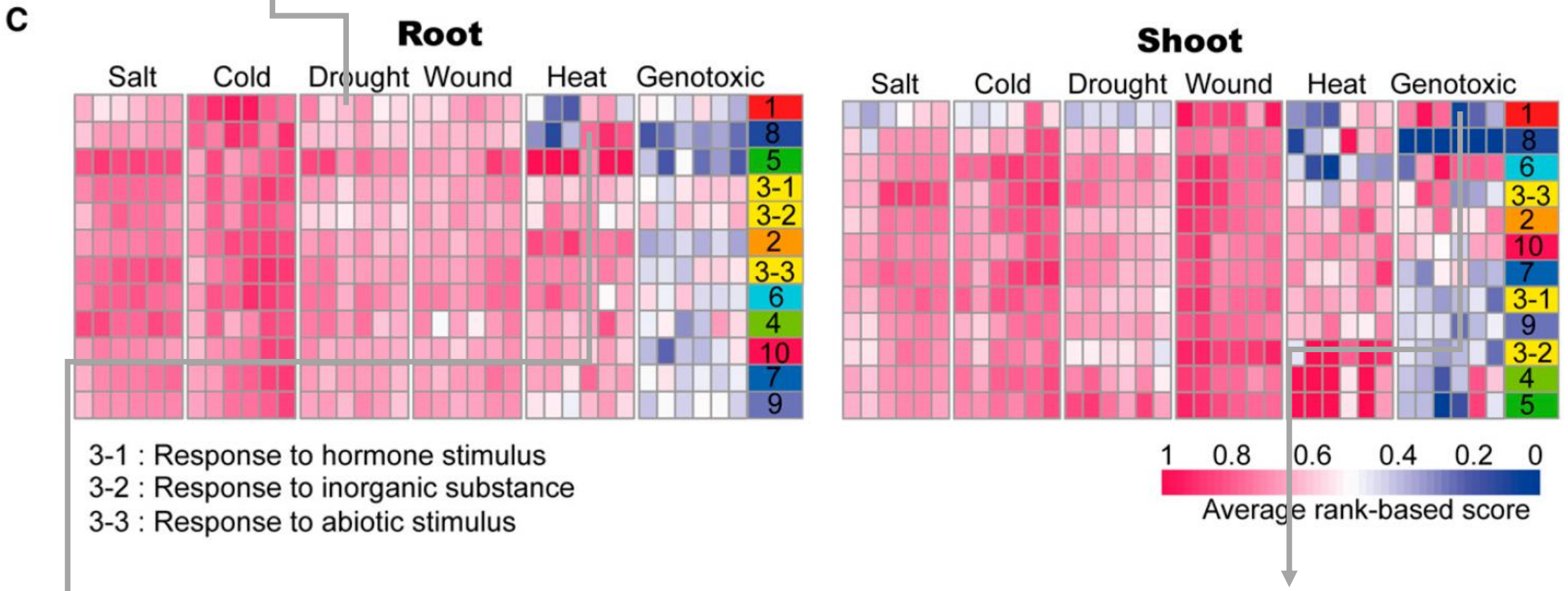
C



Aumento após 3h

10 (Desenvolvimento, crescimento e organização celular), 7 (síntese de carboidratos), 3 (resposta a estímulos hormonais) e 4 (regulação positiva de processos celulares)

Estresses por ferimentos e seca tendem a diminuir a expressão de genes após 3h

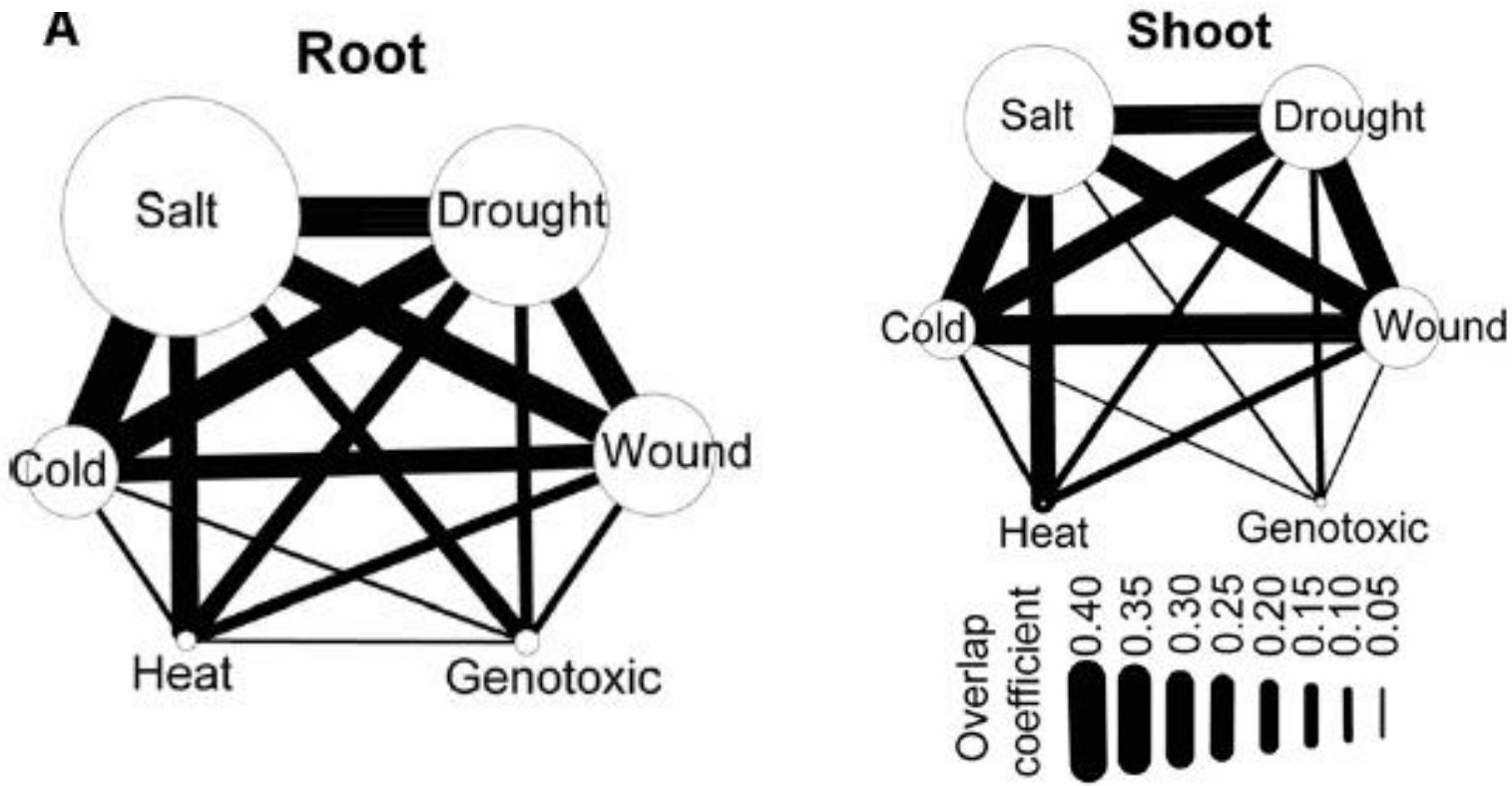


1 (fotossíntese) e 8 (metabolismo de pigmentos)

Diminuição: 1 (fotossíntese) e 3 (Resposta a estímulo abiótico)

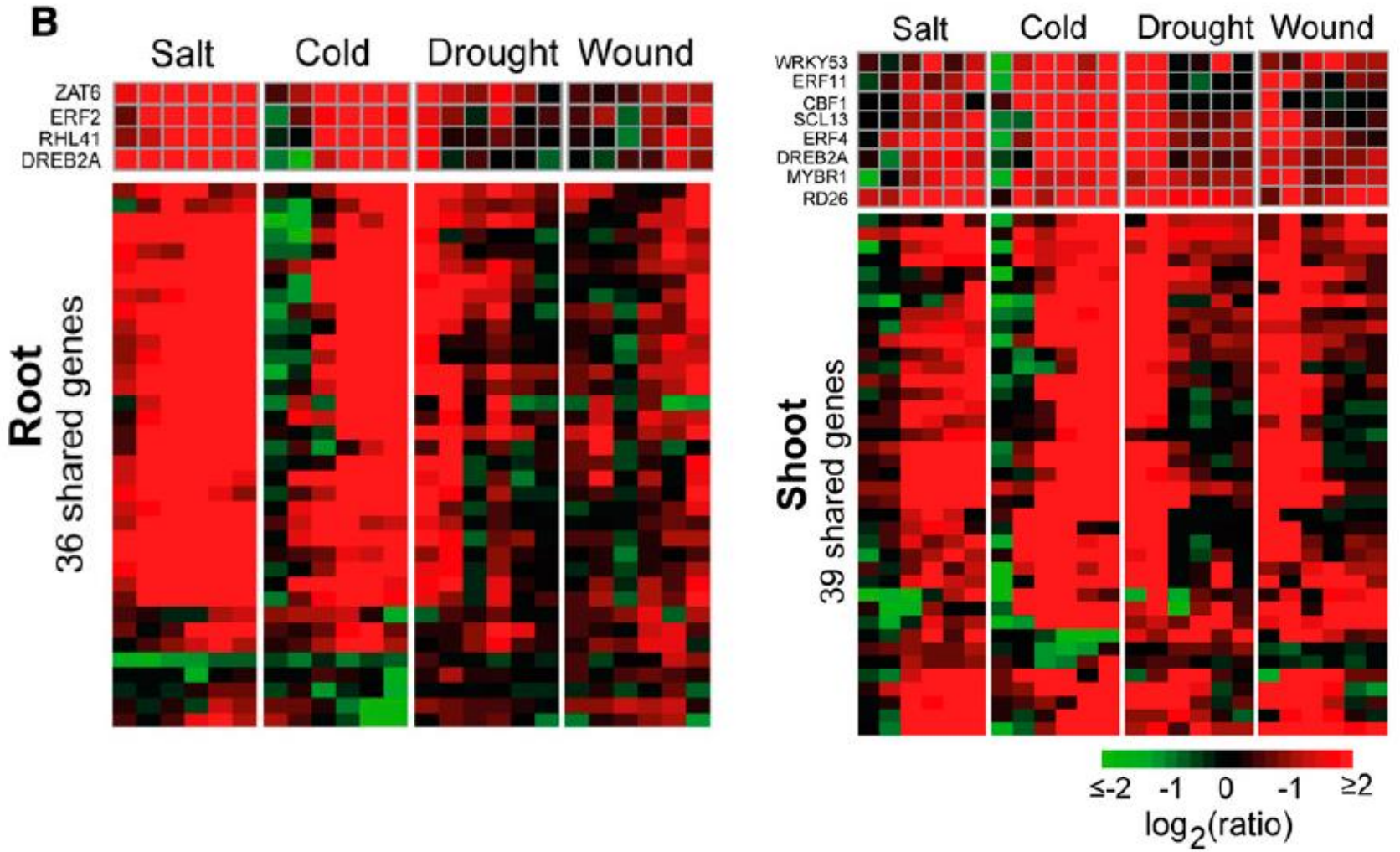
Genes compartilhados entre múltiplos estresses

- Baseado nos resultados preditos por mlDNA
- Grau de sobreposição (Merico, 2010)



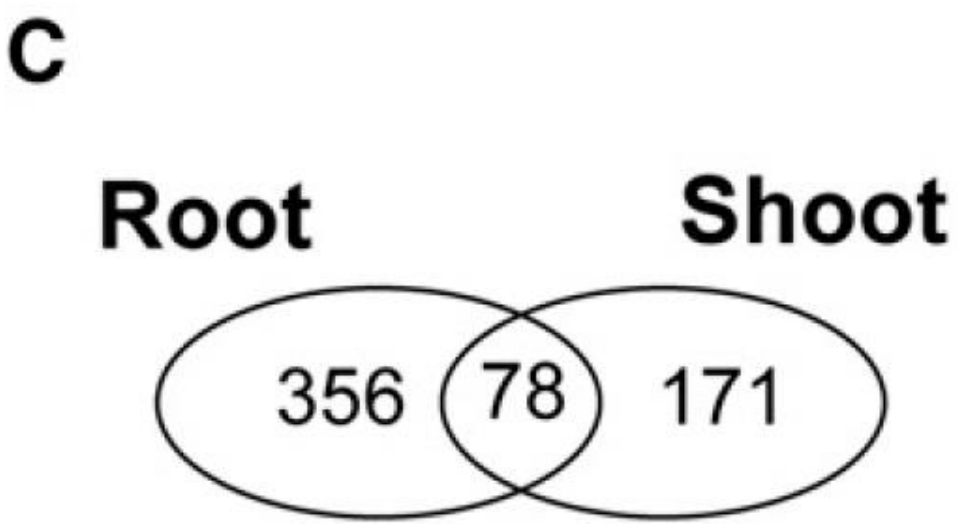
Genes compartilhados entre múltiplos estresses

- Condições dos experimentos interferiram nos resultados



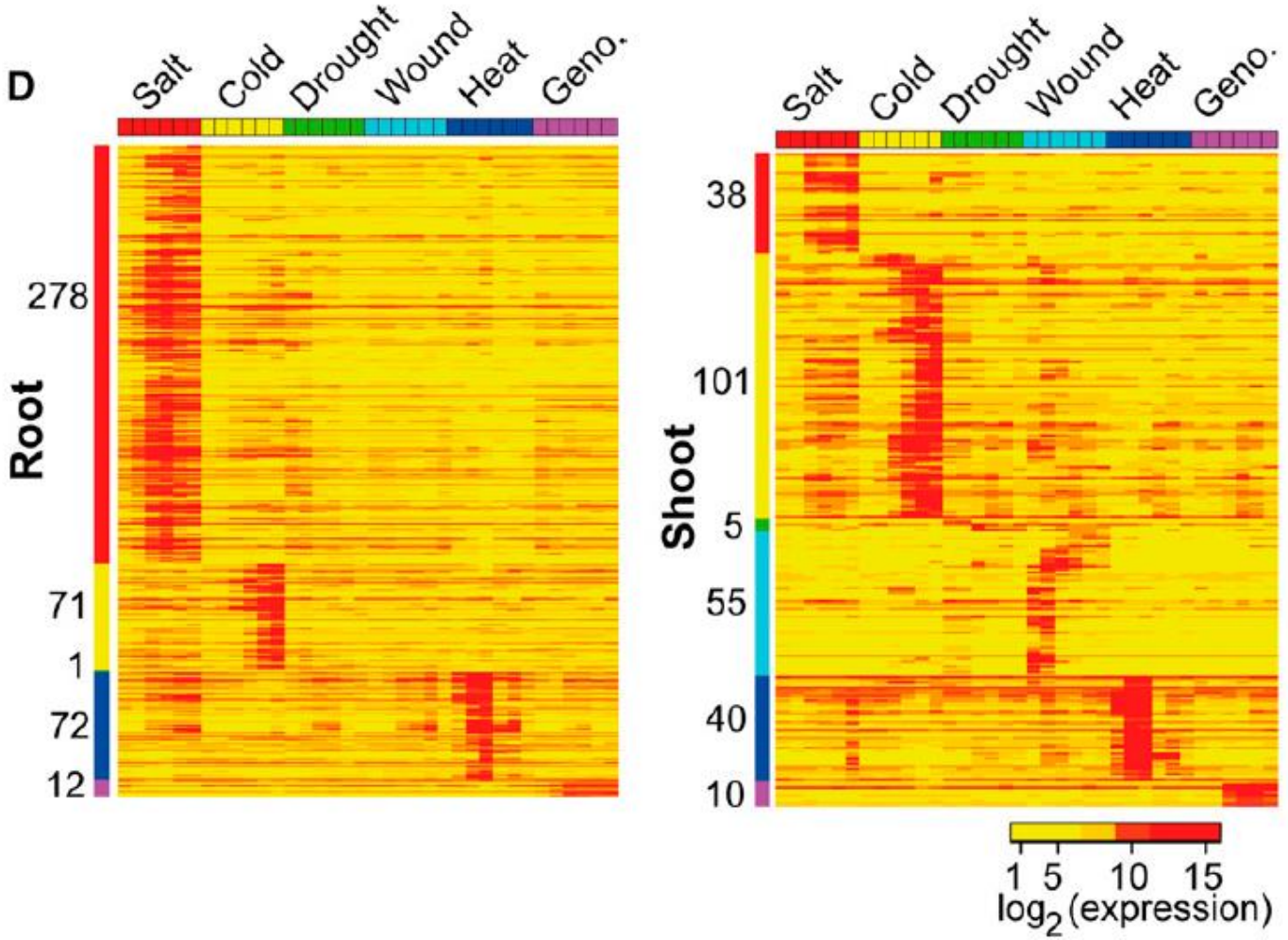
Genes expressos em estresses específicos

$$\frac{\text{Máximo valor de expressão de um gene sobre uma condição de estresse}}{\text{Máximo valor de expressão sobre outras cinco condições de estresse}}$$



Genes expressos em estresses específicos

Maior parte dos genes específicos a estresse salino



Maior parte dos genes específicos a estresse por frio

Implantação do ML

- Os procedimentos do mlDNA foram implementados como um pacote do R e estão disponíveis em:

http://cran.rproject.org/web/packages/mlDNA

- Tutorial:

http://www.cmbb.arizona.edu/mlDNA/

Obrigada!