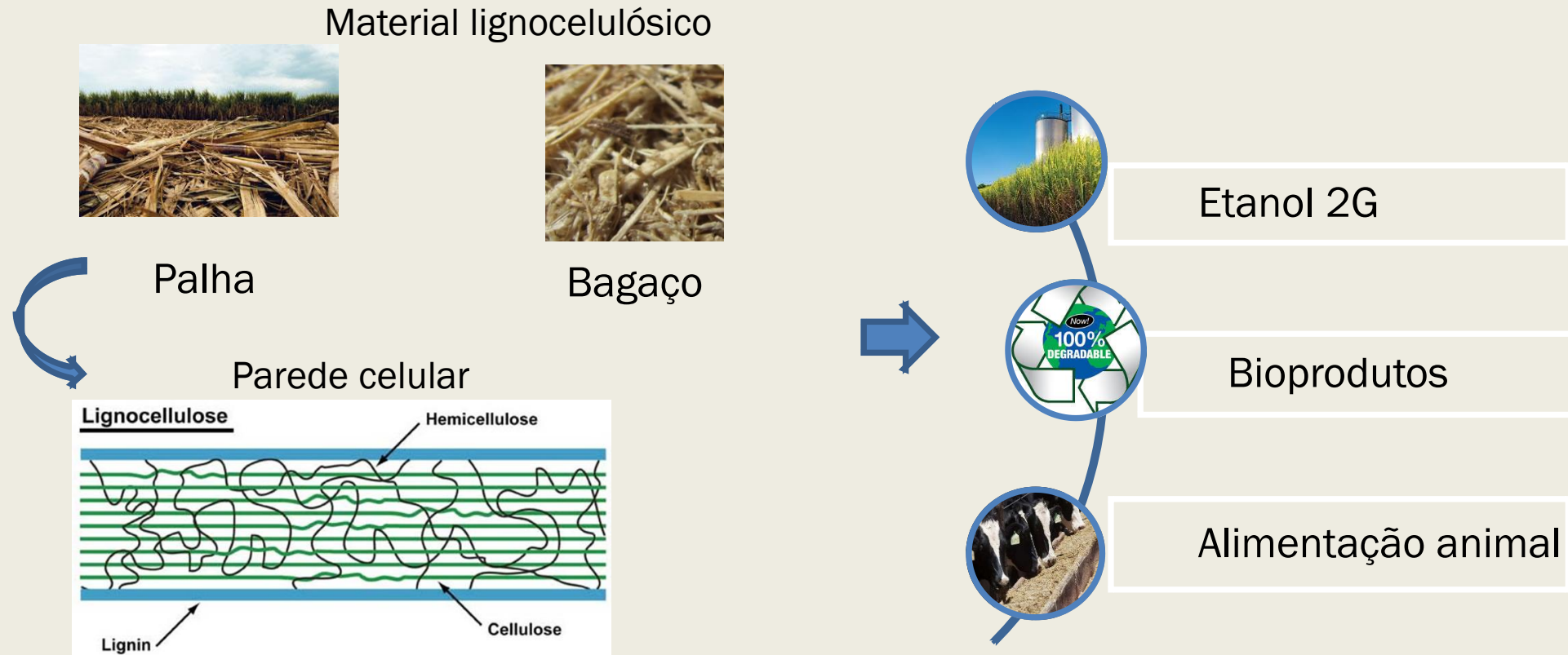


I Curso de Verão em Biotecnologia

# PROCURANDO PROMOTORES ESPECÍFICOS DE COLMO PARA APLICAÇÕES BIOTECNOLÓGICAS

Eng. Amanda Fanelli  
Eng. Maycow Berbert  
Prof.Dr. Michael dos Santos Brito

# O uso da biomassa de gramíneas (cana, milho, arroz, sorgo)



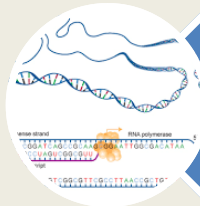
Problema: Estrutura complexa da parede!  
Como acessar os polímeros?

# Manipulação do DNA e engenharia metabólica

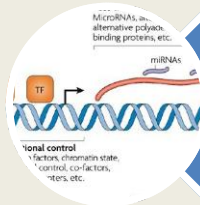
Obter uma planta com lignina/hemicelulose com estruturas diferentes ou quantidades diferentes



Conhecer a função dos genes  
(Parede celular)



Controlar a expressão gênica



Promotores e fatores de transcrição

# Análise de genes diferencialmente expressos em sorgo

- Sorghum bicolor  
(modelo para cana-de-açúcar)
- Plataforma GEO



Usar o R para criar uma ferramenta de mineração de dados

The screenshot shows the NCBI GEO Accession Display for series GSE49879. The page includes the NCBI and GEO logos, navigation links (HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, Email GEO), and a login status. The main content area displays the series title, status, platform, organisms, sample organism, experiment type, summary, overall design, contributor(s), citation(s), submission date, last update date, contact name, organization name, and street address.

Series GSE49879		Query DataSets for GSE49879
Status	Public on Dec 31, 2013	
Title	Expression data from vegetative tissues of grain, sweet and bioenergy sorghums	
Platform organisms	Sorghum bicolor; Saccharum hybrid cultivar; Zea mays subsp. mays	
Sample organism	Sorghum bicolor	
Experiment type	Expression profiling by array Non-coding RNA profiling by array	
Summary	We developed a commercially available whole-transcriptome sorghum microarray (Sorgh-WTa520972F) and generated this dataset to identify tissue and genotype-specific expression patterns for all identified Sorghum bicolor exons and UTRs. The genechip contains 1,026,373 probes covering 149,182 exons (27,577 genes) across the Sorghum bicolor nuclear, chloroplast and mitochondrial genomes. Specific probesets were also included for putative non-coding RNAs that may play a role in gene regulation (e.g., microRNAs), and confirmed functional small RNAs in related species (corn and sugarcane) were also included in our array design.	
Overall design	78 samples were analyzed from four different tissue types (shoot, seedling, leaves and stem), two dissected stem tissues (pith and rind) and six diverse genotypes (PI455230, Atlas, PI152611, AR2400, R159, and Fremont)	
Contributor(s)	Shakoor N, Feltus A, Kresovich S	
Citation(s)	Shakoor N, Nair R, Crasta O, Morris G et al. A Sorghum bicolor expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums. <i>BMC Plant Biol</i> 2014 Jan 23;14:35. PMID: 24456189	
Submission date	Aug 14, 2013	
Last update date	Jan 28, 2014	
Contact name	Nadia Shakoor	
Organization name	Chromatin, Inc.	
Street address	10 South LaSalle Street, Suite 2100	

# Quando visualizei a estrutura de dados do GEO pela primeira vez no R

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help


[CursoVeraoBiotec.R]
1 library("geoquery")
2 library("dplyr")
3 library("tidyselect")
4
5
6 #download gse database
7 gds <- getGEO("GSE49879", GSEMatrix = T)
8
9 #list all expressed genes
10 ExpressionSetData = as.data.frame(pData(pData(gds[[1]]))[,c(1,2,6,8)], stringsAsFactors=FALSE)
11 ExpressionSetData = data.frame("GSM" = ExpressionSetData$geo_accession, "title" = ExpressionSetData$title, stringsAsFactors=FALSE)
12 ExpressionSetData$title = as.data.frame(ExpressionSetData$title)
13
14
15 samples = as.matrix(ExpressionSetData)
16 samples <- select(ExpressionSetData, starts_with("pith"), starts_with("internode"), starts_with("shoot"), starts_with("rind"))
17
18 #looking for genes that occur in culm
19 listwordPatternMatching = c("pith", "internode", "shoot", "rind")
20
21 #I will search in the title of the data those that correspond with the key words of listwordPatternMatching
22 #and the retriever will be the position within the dataframe
23 listPositionGSM = NULL
24
25 for (i in 1:length(listwordPatternMatching)) {
26   catPosition = grep(listwordPatternMatching[i], ExpressionSetData$title)
27   listPositionGSM <- append(listPositionGSM, catPosition, after = length(listPositionGSM))
28 }
29
30
31 #with the list of positions I get the GSMs and create a list to be able to download the data of each one
32 listGSM = NULL
33
34 for (x in 1:length(listPositionGSM)) {
35   accessPoint = as.numeric(listPositionGSM[x])
36   listGSM <- append(listGSM, ExpressionSetData$geo_accession[accessPoint], after = length(listGSM))
37 }
38
39 [1061] [top level]
40
41 Console Terminal
42
43 GSM1208520 all
44 GSM1208521 all
45 GSM1208522 all
46 GSM1208523 all
47 GSM1208524 top
48 GSM1208525 top
49 GSM1208526 top
50 GSM1208527 middle
51 GSM1208528 bottom
52 GSM1208529 middle
53 GSM1208530 bottom
54 GSM1208531 middle
55 GSM1208532 bottom
56 GSM1208533 all
57 GSM1208534 all
58 [ reached getoption("max.print") -- omitted 55 rows ]
59 >
```



# Análises dos genes usando a plataforma do site

- Genótipos de biomassa PI455230, PI152611, AR2400

- Dois grupos



Colmo: Pith, rind, Shoot,  
internode

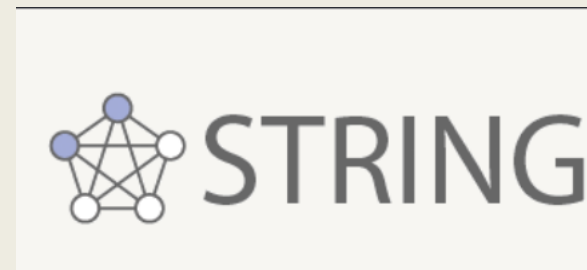
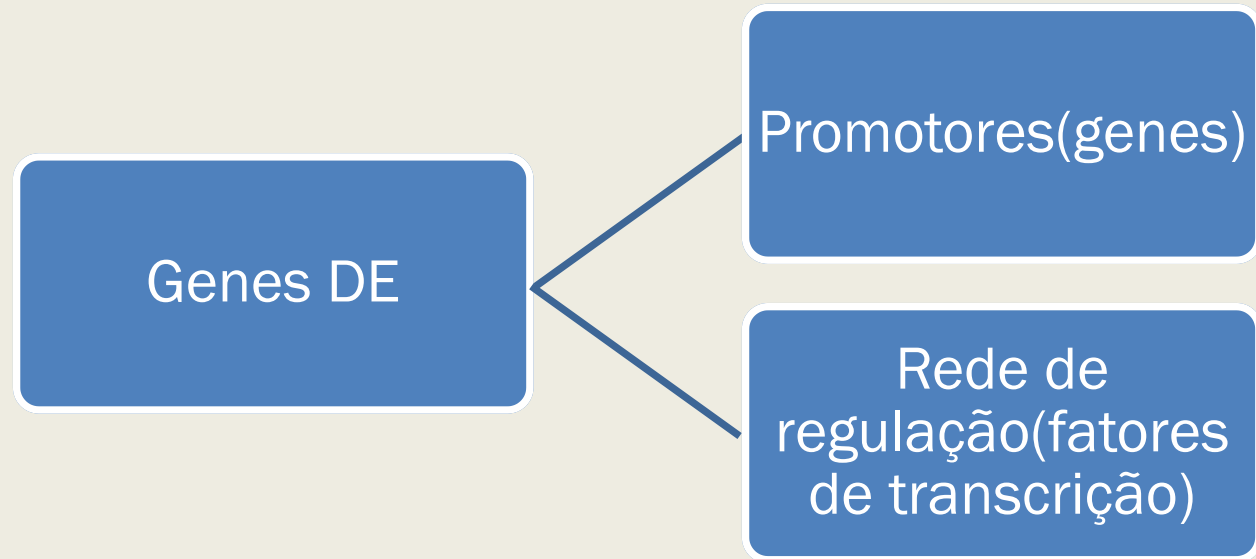
Outros: Leaf, root,  
seedling root

# Genes diferencialmente expressos com função de parede celular ou fator de transcrição

Proteins	Domains
Sb02g036010.1	PF03552 AT1G02730.1 ATCSLD5 ATCSLD5; 1,4-beta-D-xylan synthase/ cellulose synthase
Sb02g002200.1	PF03790,PF03791,PF03789 PTHR11850 KOG0773 AT4G08150.1 KNAT1 KNAT1 (KNOTTED-LIKE FROM ARABIDOPSIS THALIANA); transcription factor
Sb02g036000.1	PF03552 AT1G02730.1 ATCSLD5 ATCSLD5; 1,4-beta-D-xylan synthase/ cellulose synthase
Sb01g009460.1	PF03790,PF03791,PF03789 PTHR11850 KOG0773 AT4G08150.1 KNAT1 KNAT1 (KNOTTED-LIKE FROM ARABIDOPSIS THALIANA); transcription factor
Sb04g033930.1	PF00319,PF01486 PTHR11945 KOG0014 AT2G22540.1 SVP SVP (SHORT VEGETATIVE PHASE); transcription factor/ translation repressor, nucleic acid binding
Sb01g021720.1	AT1G08050.1 zinc finger (C3HC4-type RING finger) family protein
Sb01g006790.1	PF03790,PF03791,PF03789 PTHR11850 KOG0773 AT4G08150.1 KNAT1 KNAT1 (KNOTTED-LIKE FROM ARABIDOPSIS THALIANA); transcription factor
Sb04g008670.1	PF00249 AT2G38300.1 DNA binding / transcription factor
Sb02g012640.1	AT5G60710.1 zinc finger (C3HC4-type RING finger) family protein
Sb06g000920.1	PF00092 PTHR10166 AT5G60710.1 zinc finger (C3HC4-type RING finger) family protein
Sb10g021360.1	PF00249 AT2G38300.1 DNA binding / transcription factor
Sb07g005070.1	PF01370 PTHR10366 KOG1502 AT2G33600.1 cinnamoyl-CoA reductase family
Sb09g002080.1	PF00847 AT2G28550.1 RAP2.7 RAP2.7 (RELATED TO AP2.7); DNA binding / transcription factor
Sb03g003640.1	PF03106 AT4G39410.1 WRKY13 WRKY13; transcription factor
Sb05g002940.1	PF03479 AT4G12080.1 DNA-binding family protein
Sb03g003370.1	PF03106 AT5G15130.1 WRKY72 WRKY72; transcription factor
Sb03g009840.1	PF01370 PTHR10366 KOG1502 AT1G15950.1 CCR1 CCR1 (CINNAMOYL COA REDUCTASE 1); cinnamoyl-CoA reductase
Sb01g046040.1	PF05678 AT3G56710.1 SIB1 SIB1 (SIGMA FACTOR BINDING PROTEIN 1); binding / protein binding
Sb09g026100.1	PF02309 AT3G16500.1 PAP1 PAP1 (PHYTOCHROME-ASSOCIATED PROTEIN 1); transcription factor
Sb04g027540.1	PF00249 PTHR10641 KOG0048 K09422 AT1G34670.1 AtMYB93 AtMYB93 (myb domain protein 93); DNA binding / transcription factor
Sb06g001430.1	PF08240,PF00107 PTHR11695 KOG0023 K00095 AT4G39330.1 CAD9 CAD9 (CINNAMYL ALCOHOL DEHYDROGENASE 9); binding / catalytic/ oxidoreductase/ zinc ion binding
Sb01g009480.1	PF03790,PF03791,PF03789 PTHR11850 KOG0773 AT4G08150.1 KNAT1 KNAT1 (KNOTTED-LIKE FROM ARABIDOPSIS THALIANA); transcription factor
Sb08g001850.1	PF00931 AT3G07040.1 RPM1 RPM1 (RESISTANCE TO P. SYRINGAE PV MACULICOLA 1); nucleotide binding / protein binding
Sb01g014310.1	PF02365 AT3G18400.1 anac058 anac058 (Arabidopsis NAC domain containing protein 58); transcription factor
Sb07g024550.1	PF12171,PF00096 PTHR11389 KOG3576 AT2G02080.1 AtIDD4 AtIDD4 (Arabidopsis thaliana Indeterminate(ID)-Domain 4); transcription factor



# Análise dos promotores

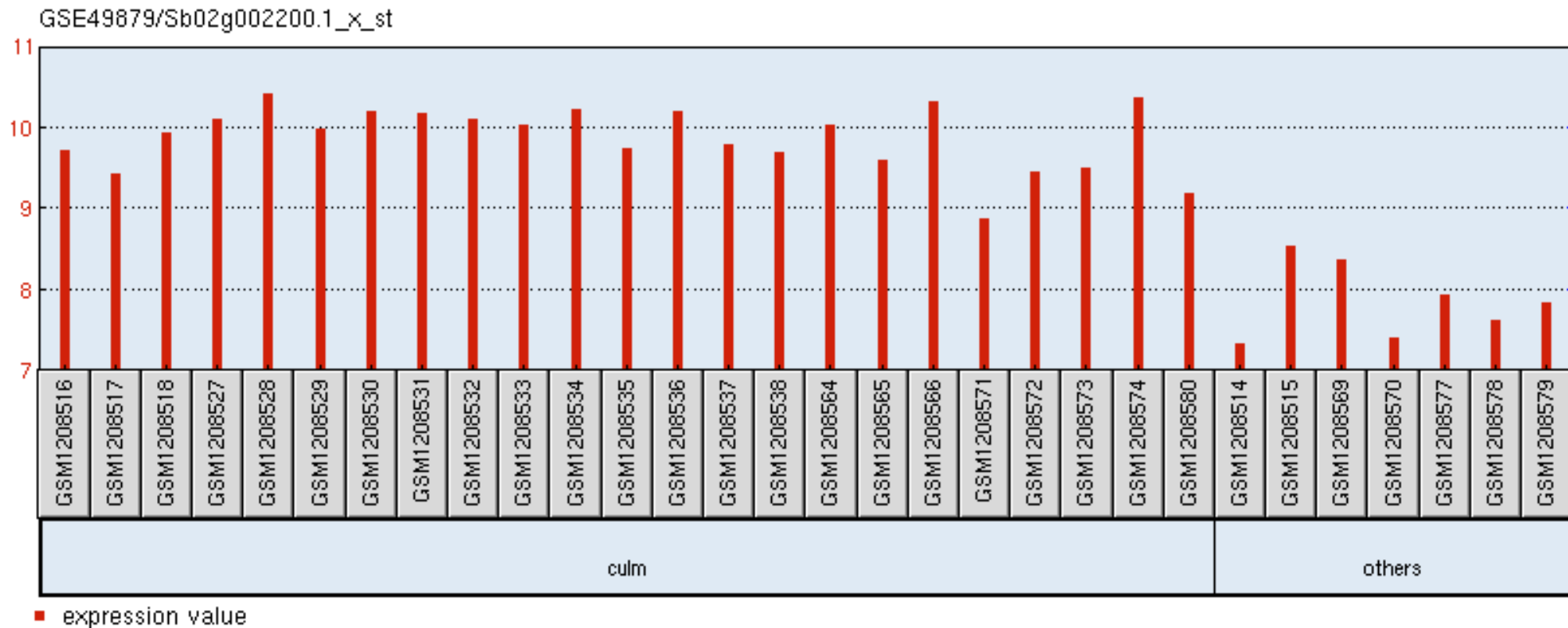




# KNAT1 Sb02g002201.1

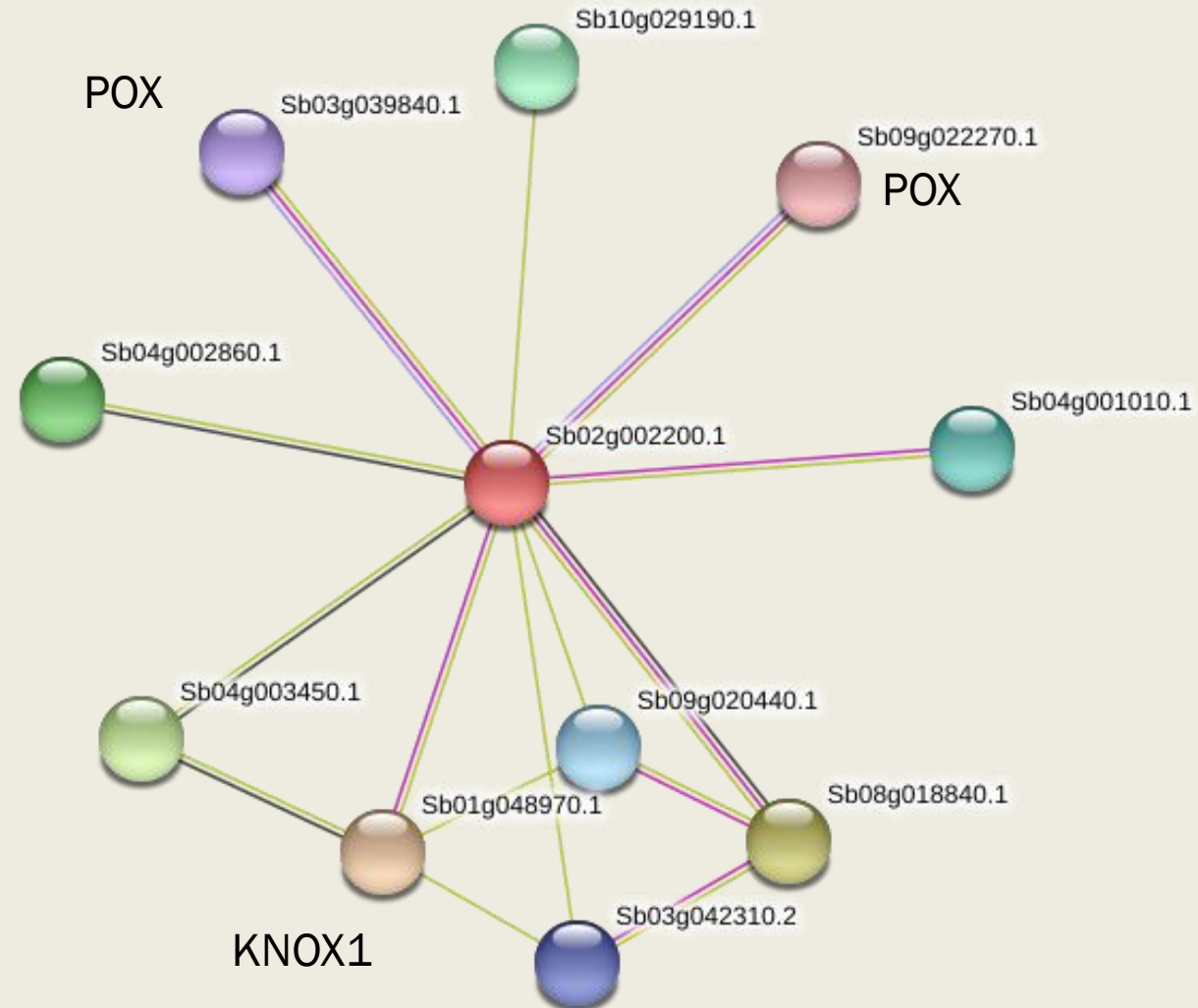
Fator de transcrição KNAT1:

classI Knox gene relacionado com o desenvolvimento de fibras (Em Arabidopsis)



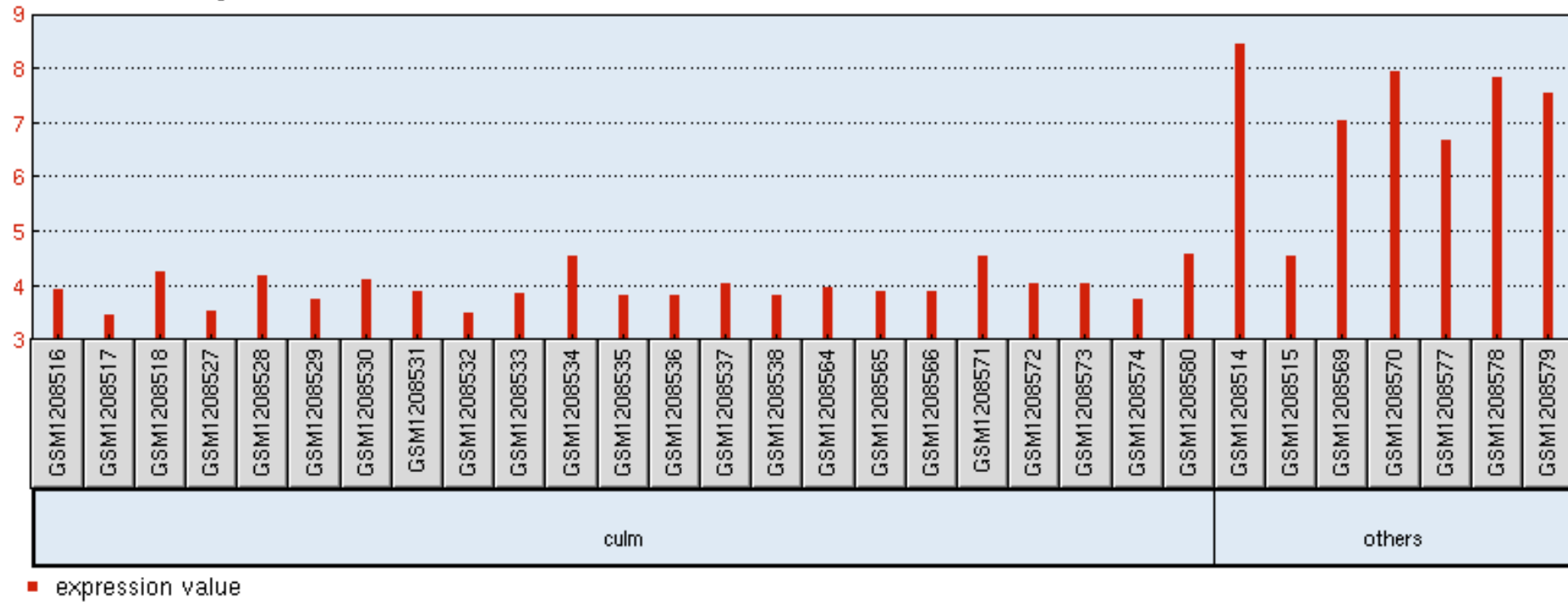
# KNAT1 Sb02g002200.1

## Análise da rede de regulação
































# Ces Sb02g036010.1

GSE49879/Sb02g036010.1\_s\_st



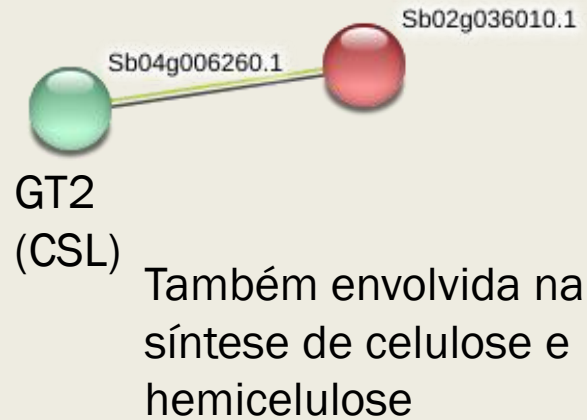
# Elementos cis

## Ces Sb02g036010.1

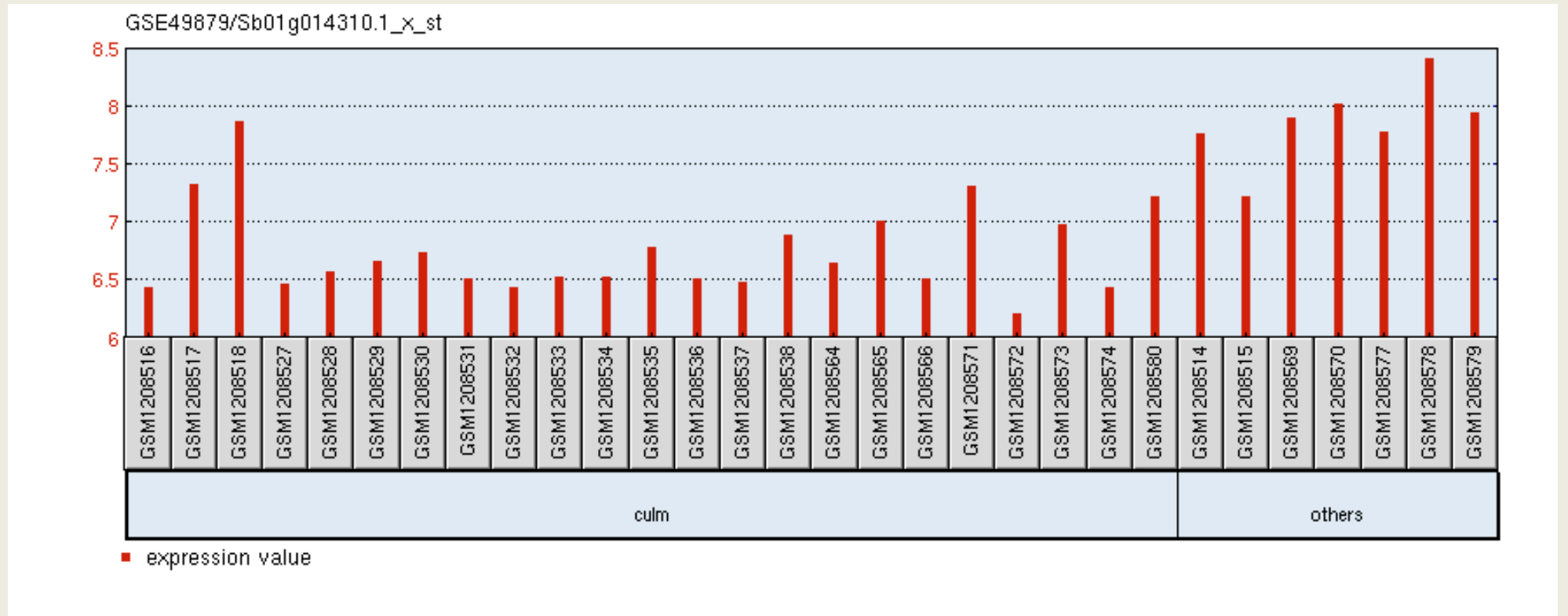
+		ABRE
+		ARE
+		AT~TATA-box
+		Box 4
+		CAAT-box
+		CAT-box
+		CCAAT-box
+		CCGTCC motif
+		CCGTCC-box
+		CGTCA-motif
+		DRE1
+		G-Box
+		G-box
+		GA-motif
+		GARE-motif
+		GATA-motif
+		HD-Zip 1
+		MBS
+		MYB
+		MYB recognition site
+		MYC
+		Myb
+		Myb-binding site
+		STRE
+		TATA-box
+		TC-rich repeats
+		TGACG-motif
+		Unnamed__1
+		Unnamed__4

# Ces Sb02g036010.1

## Análise da rede de regulação

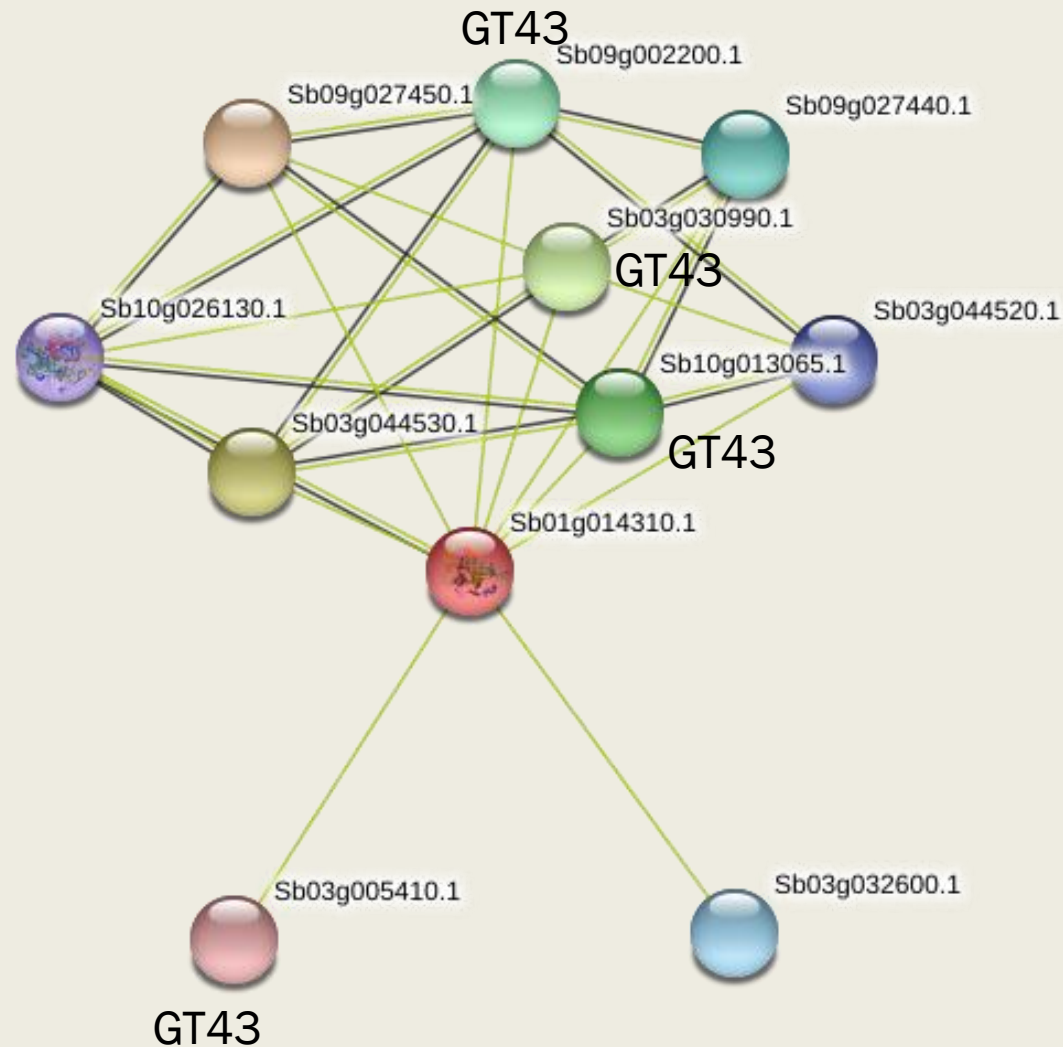


# NAC Sb01g014310.1



# NAC Sb01g014310.1

## Análise da rede de regulação



Genes de síntese  
de hemicelulose



# Conclusão

- Métodos de mineração de dados
- Estudar o promotor dos genes KNAT1
- Genes de celulose e lignina, fatores de transcrição possivelmente envolvidos na hemicelulose (NAC)

# Agradecimentos

- ICT-Unifesp
- Prof.Dr. Michael Brito
- Alunos do laboratório (Alice e Marinara)