GENE EXPRESSION PROFILING AND MICRO RNA DISCOVERY IN PLANTS USING ILLUMINA'S SHORT-READ SEQUENCING PLATFORM

Georgios Pappas Jr.

gpappas@cenargen.embrapa.br

EMBRAPA recursos genéticos e biotecnologia

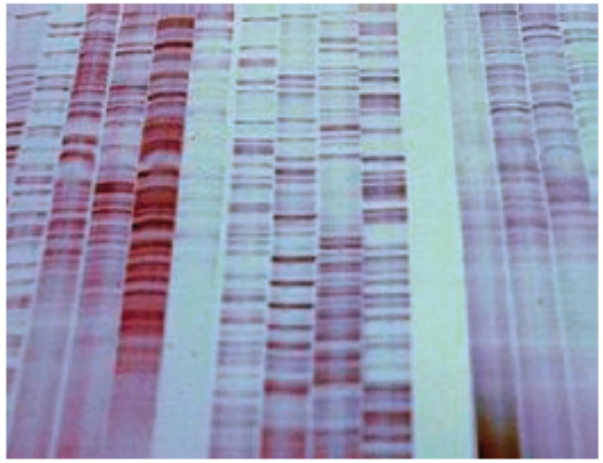
Universidade Católica de Brasília

31/08/2009

The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

In the toxicology building of North Carolina State University in Raleigh, Nigel Deighton, head of a small genome research facility, and a few others unpack the facility's first next-generation sequencing machine, a 454 GS FLX, on loan from Roche Diagnostics for three months. They train for a few days, nebulize a colleague's bacterial DNA and PCR-amplify "the living daylights out of it," Deighton recalls. They load the bead-bound PCR products onto a plate with holes that are not visible to the naked eye, pop the plate into the machine and close the drawbridge-like

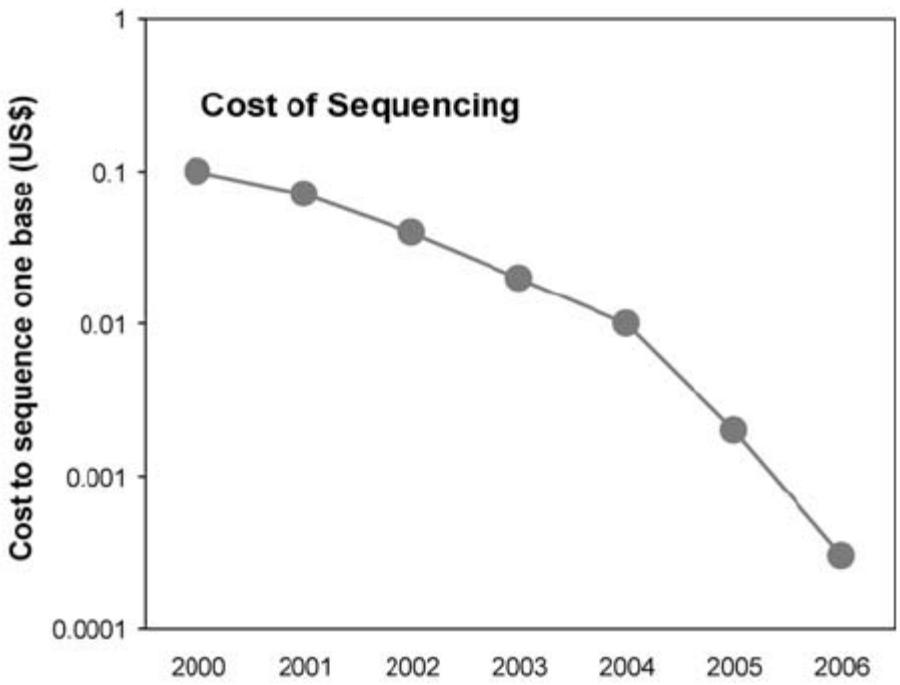


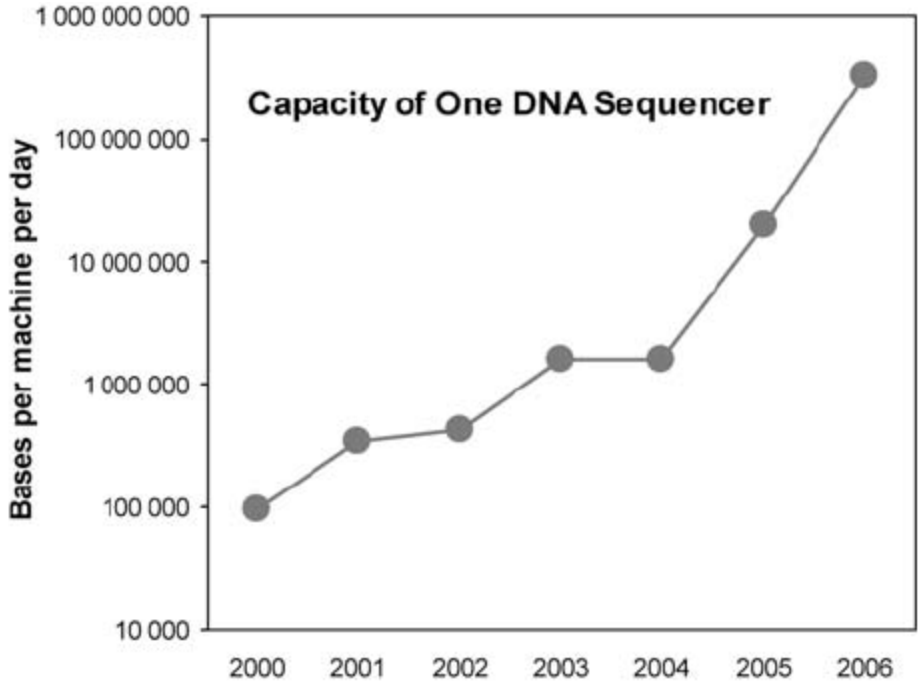
Sanger sequencing becomes the 'old' generation.

In 2007, researchers performed whole-genome human sequencing using old and new platforms. Researchers at Baylor College of Medicine and 454 Life Sciences sequenced James Watson's genome in two months, for about $1 million. Two other personal genomes were sequenced: Craig Venter's, at the Institute he founded, and that of a Chinese individual, at the Beijing Genomics Institute. The J. Craig Venter Institute used Sanger technology for sequencing Venter's DNA, which cost an estimated $70 million and took several years, but is now focusing its efforts

PUBLISHED ONLINE 19 DECEMBER 2007; DOI:10.1038/NMETH1154

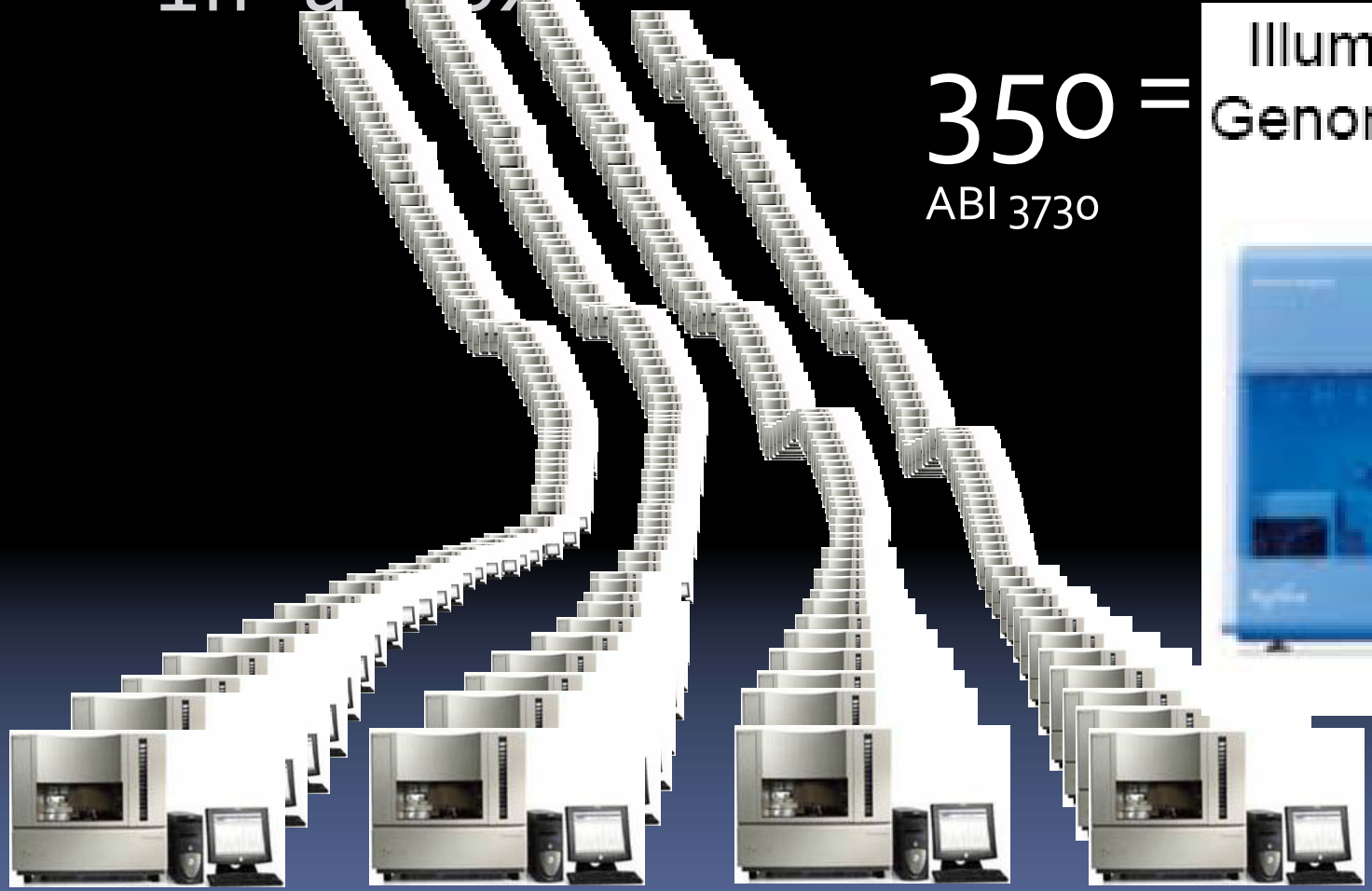
Sequencing evolution





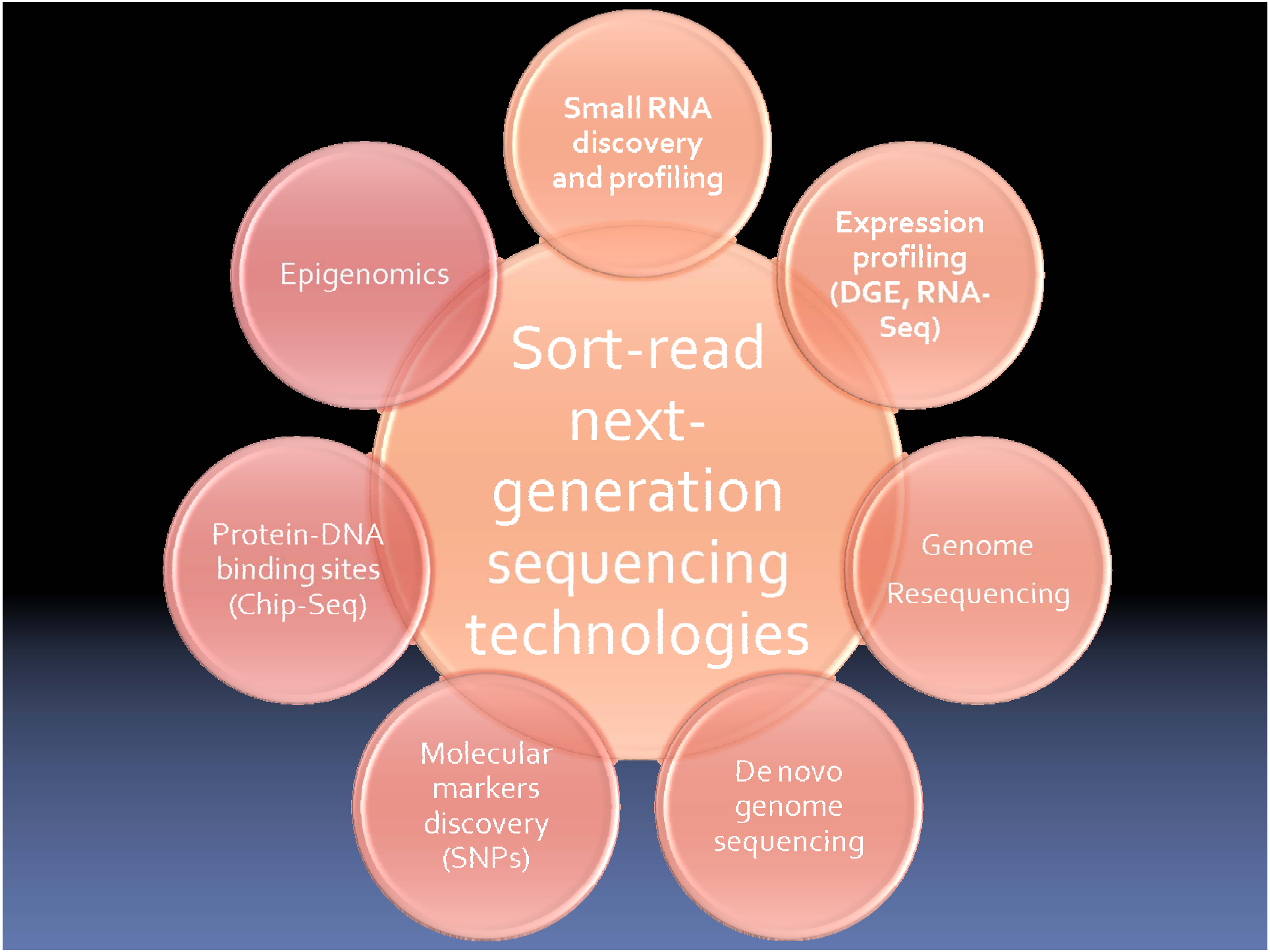
HUDSON, MATTHEW E. (2008)
Sequencing breakthroughs for genomic ecology and evolutionary biology.
Molecular Ecology Resources **8** (1), 3-17.

Illumina GA is a genome center in a box!



350 = ABI 3730

Illumina/Solexa Genome Analyzer


Small RNA discovery and profiling
Expression profiling (DGE, RNA-Seq)
Genome Resequencing
De novo genome sequencing
Molecular markers discovery (SNPs)
Protein-DNA binding sites (Chip-Seq)
Epigenomics
Sort-read next-generation sequencing technologies

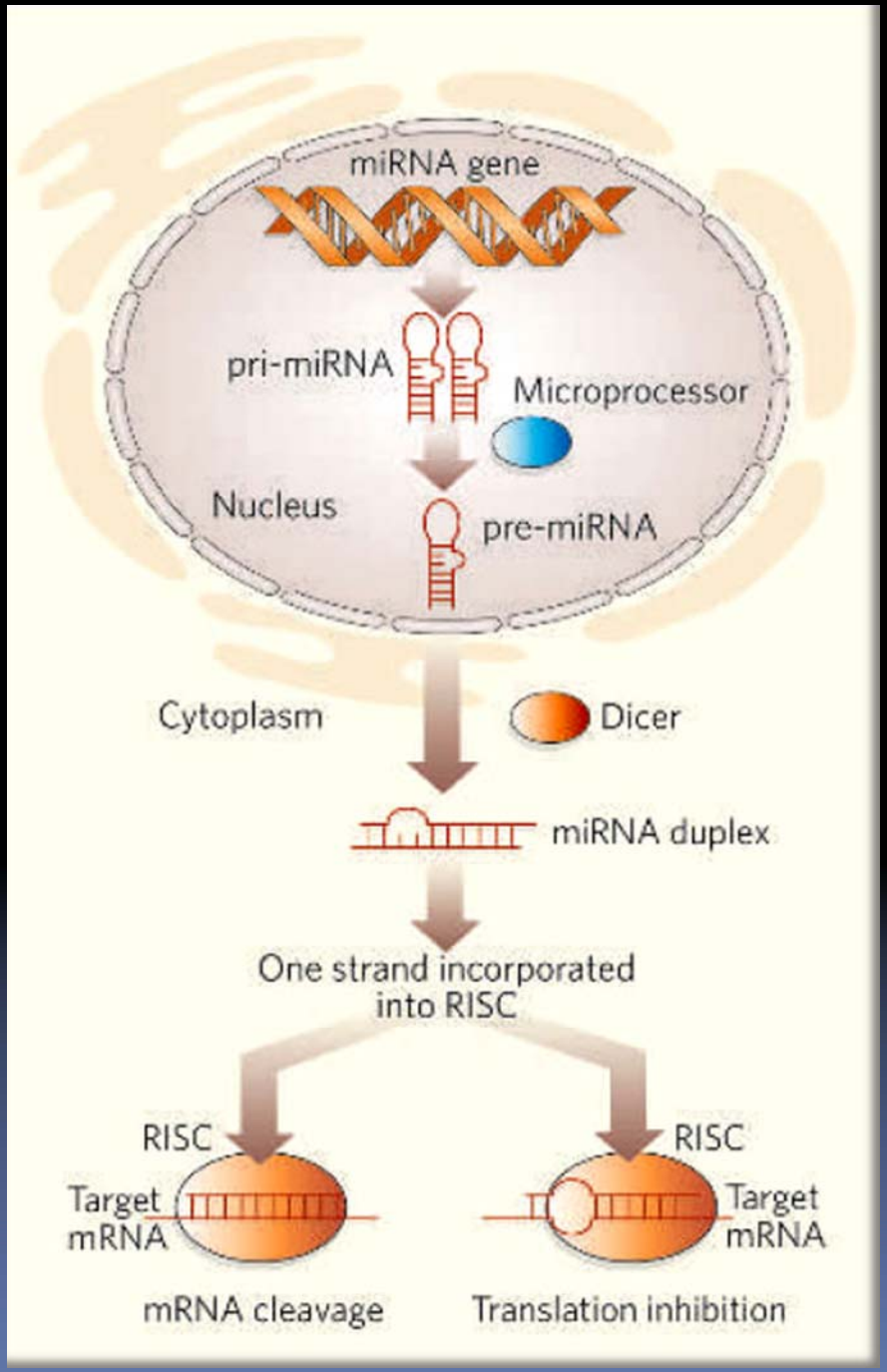
Experiments using Illumina GA

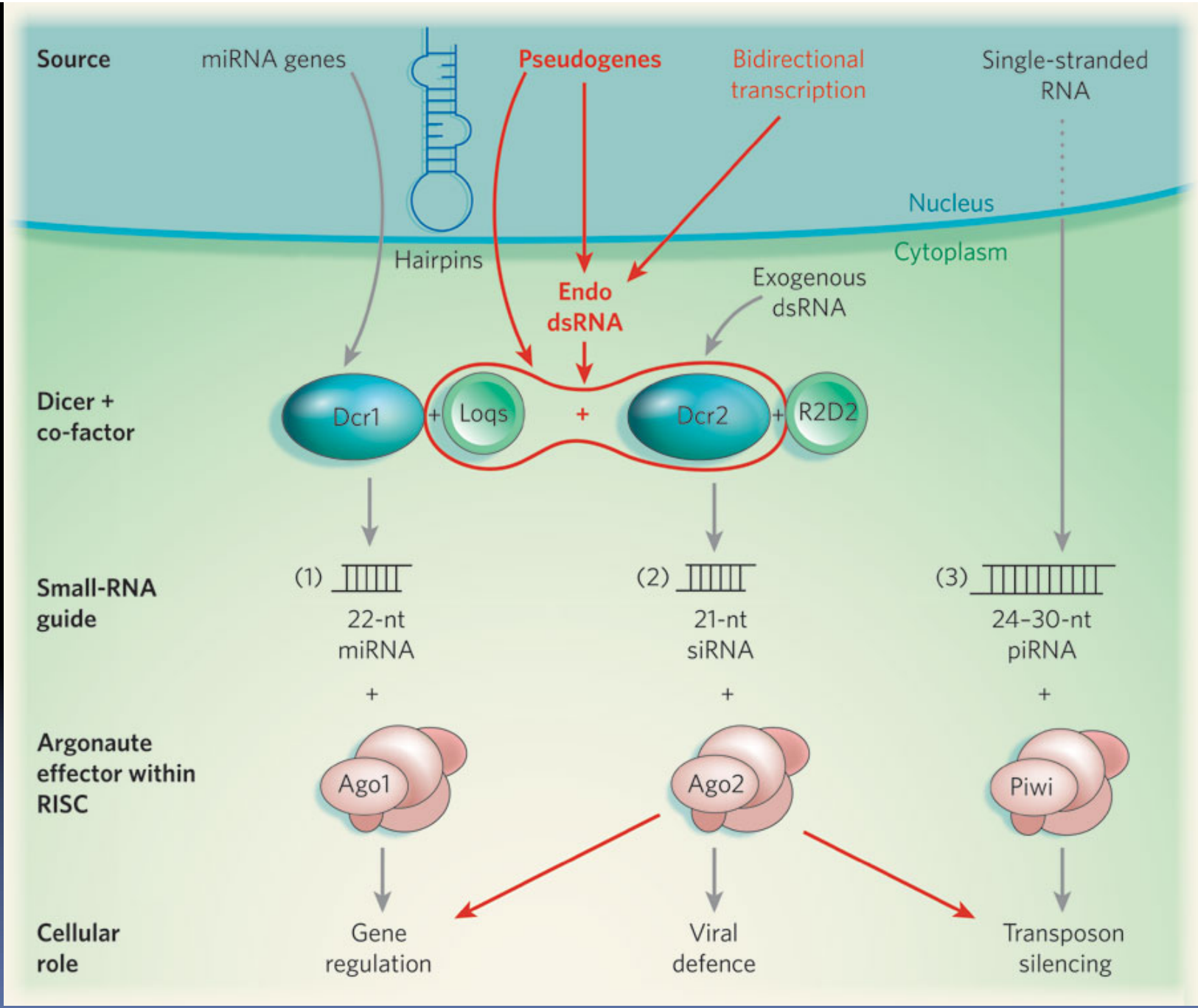
- The aim is to provide a brief overview of experiments designed at EMBRAPA using Illumina GA
- Sequencing was performed at Fasteris (http://www.fasteris.com)

Micro RNA discovery in *Eucalyptus* sp.

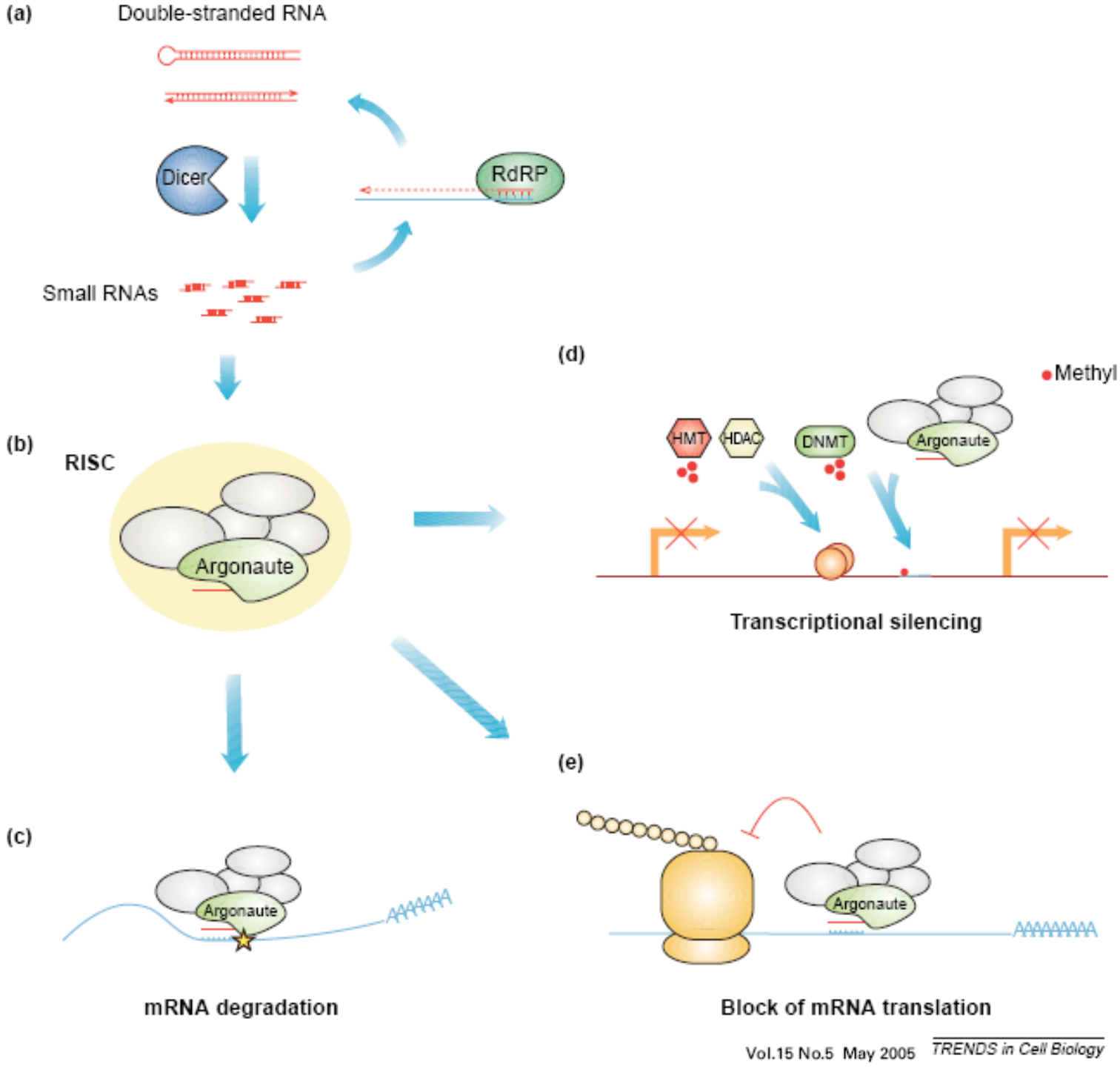
- Marília Pappas (EMBRAPA)
- Alessandra Reis (Universidade Católica de Brasília-UCB)
- Laurent Farinelli (Fasteris)
- Dario Grattapaglia (EMBRAPA/UCB)

Small RNA are increasingly being recognized as major players in gene regulation and genome homeostasis





Genomics: Protein fossils live on as RNA R Sasidharan & Gerstein. Nature 453, 729-731(5 June 2008)


(a)
Double-stranded RNA
Dicer
RdRP
Small RNAs
(b)
RISC
Argonaute
(c)
Argonaute
mRNA degradation
(d)
Methyl
HMT
HDAC
DNMT
Argonaute
Transcriptional silencing
(e)
Argonaute
Block of mRNA translation

Experiment 1: Objective

- Large scale discovery and expression profiling of small RNAs in two species of *Eucalyptus*, including the genotype that is being sequenced at JGI (*Eucalyptus grandis*; BRASUZ1)

BRASUZ1
E. grandis

Experimental design

Eucalyptus small RNA discovery and profiling

Tissue specific variability

vascular cambium

X

leaves

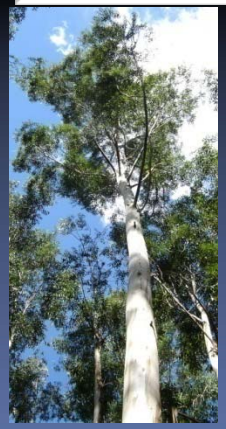
Intra-specific variability

E. globulus A2

X

E. globulus C3

Inter-specific variability

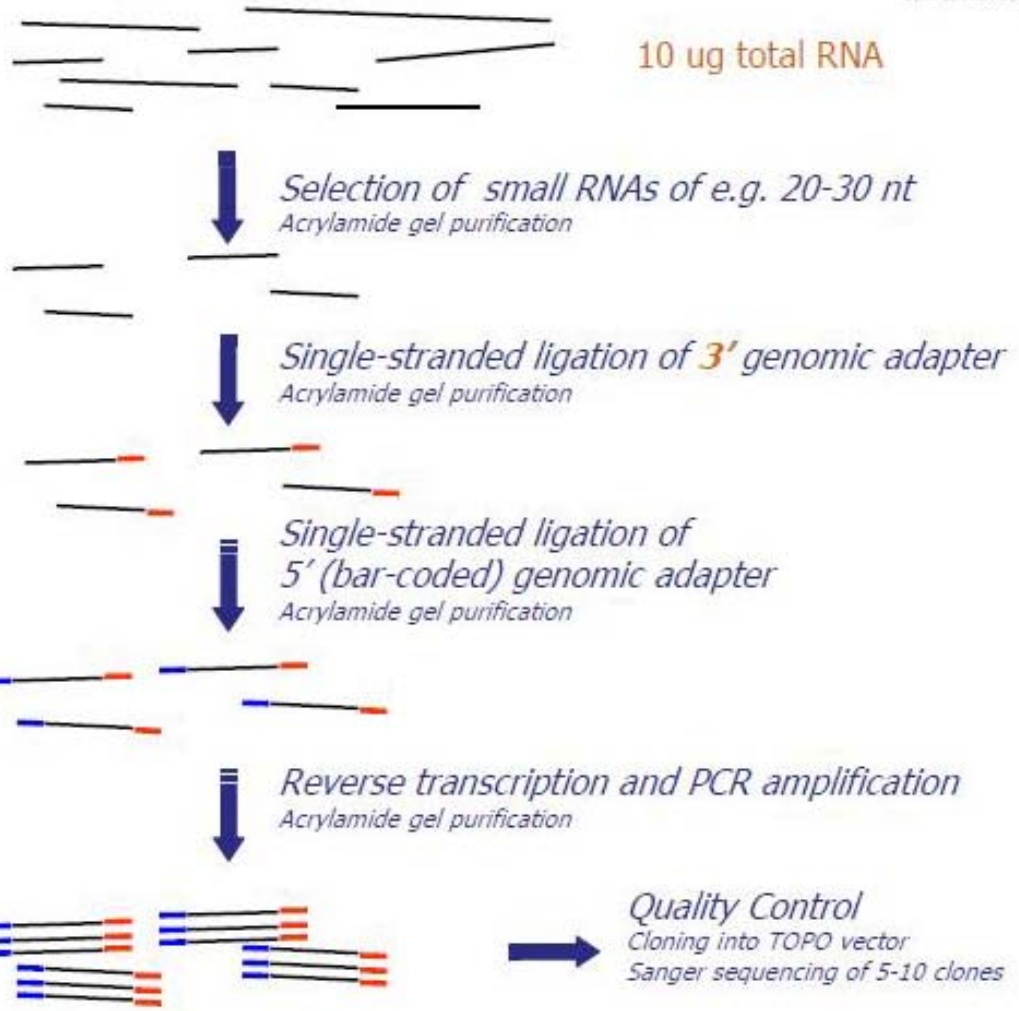


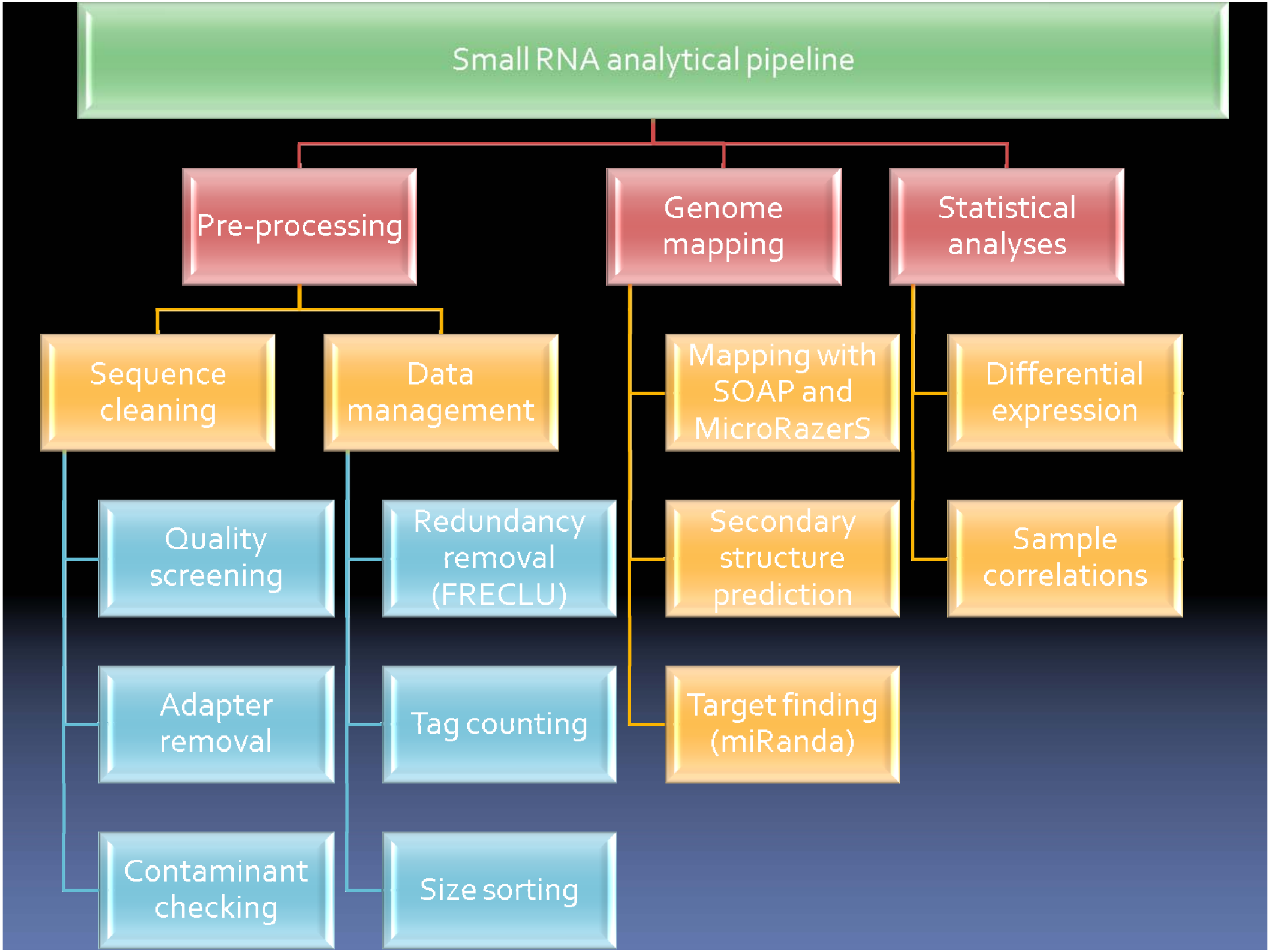
E. grandis - BRASUZ

X

E. globulus

Fasteris modification of small RNA Sample Preparation




Small RNA analytical pipeline
Pre-processing
Genome mapping
Statistical analyses
Sequence cleaning
Data management
Mapping with SOAP and MicroRazerS
Differential expression
Quality screening
Redundancy removal (FRECLU)
Secondary structure prediction
Sample correlations
Adapter removal
Tag counting
Target finding (miRanda)
Contaminant checking
Size sorting

Data analysis

- Hardware
  - Need of a complex computational infra-structure, both in terms of storage and parallel processing
- Software
  - New software should be developed to integrate several steps of the pipelines
  - In our case for a single pipeline programs were developed in PERL, JAVA, R and Groovy
- Peopleware
  - ?

EDITORIAL

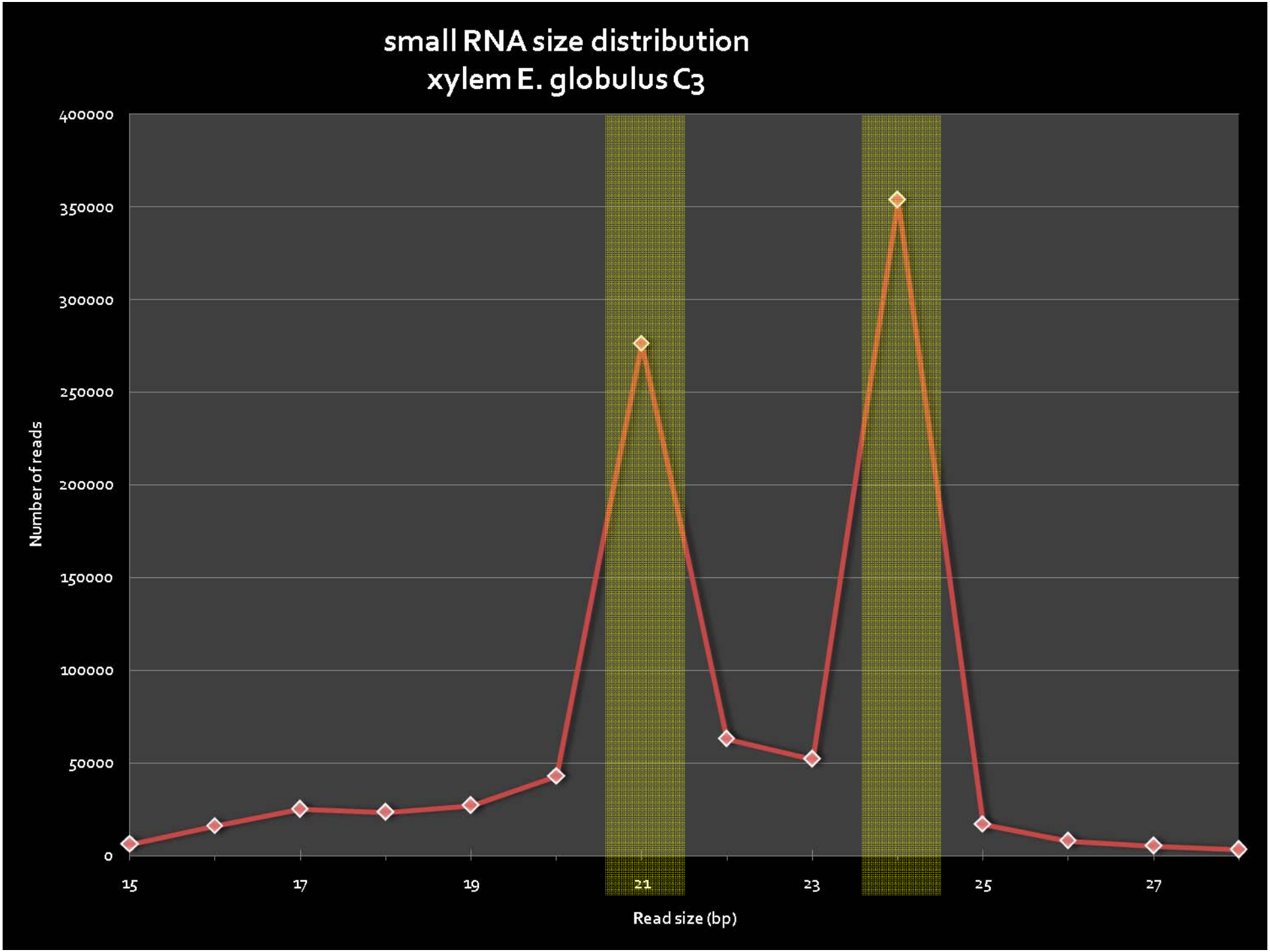
nature biotechnology

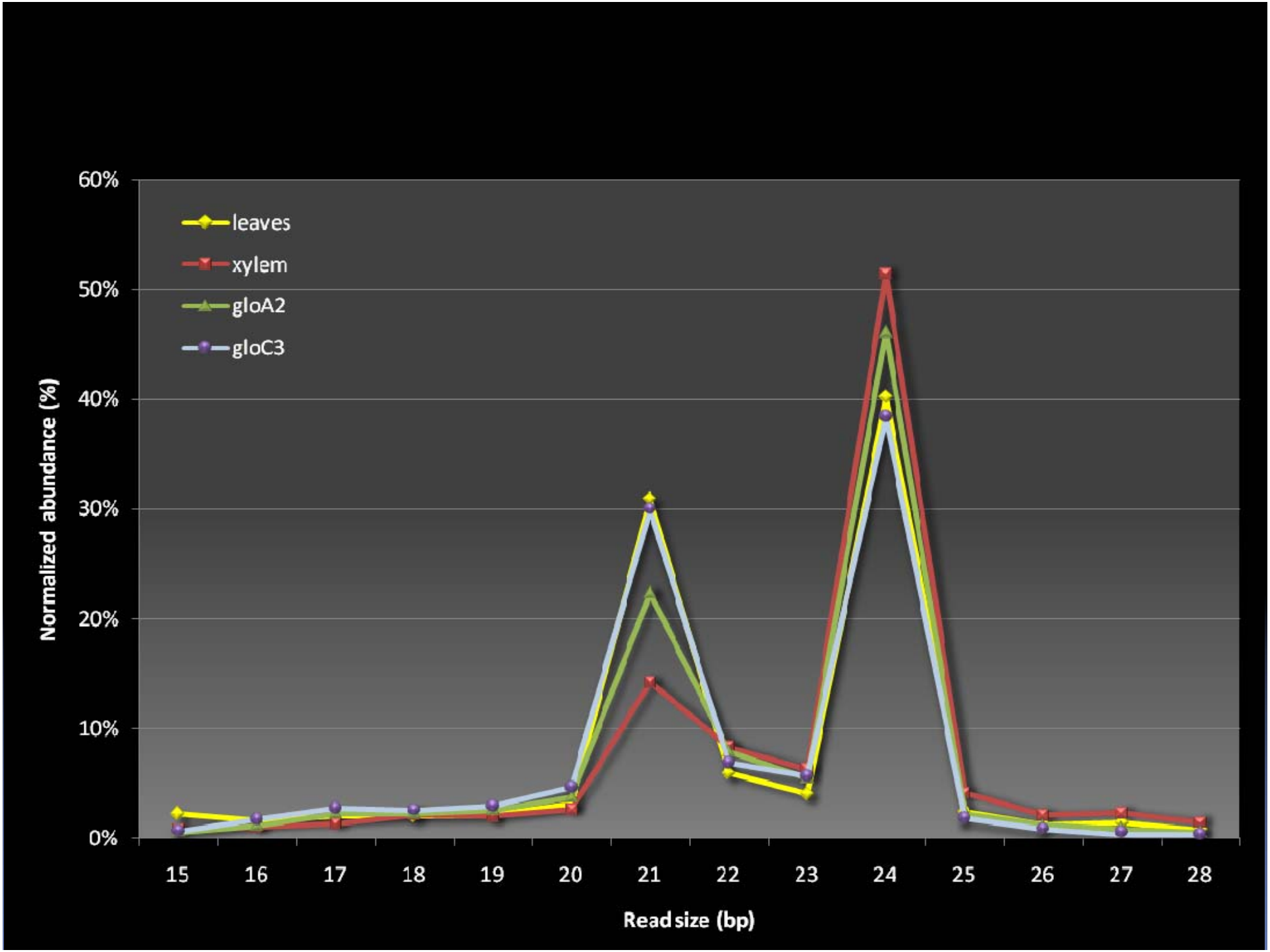
Prepare for the deluge

The gobs of data produced by next-generation sequencing are a key problem limiting wider adoption.

Overall sequencing results

Sample	Total reads	Unique sequences	Unique sequences (filtered)
BRASUZ - leaves	1,484,867	338,256	283,375
BRASUZ - xylem	1,766,355	485,321	402,158
globulus A2 - xylem	1,737,872	834,917	789,762
globulus C3 - xylem	1,115,404	540,931	505,663


small RNA size distribution
xylem E. globulus C3
400000
350000
300000
250000
200000
150000
100000
50000
0
Number of reads
15
17
19
21
23
25
27
Read size (bp)


60%
50%
40%
30%
20%
10%
0%
Normalized abundance (%)
leaves
xylem
gloA2
gloC3
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Read size (bp)

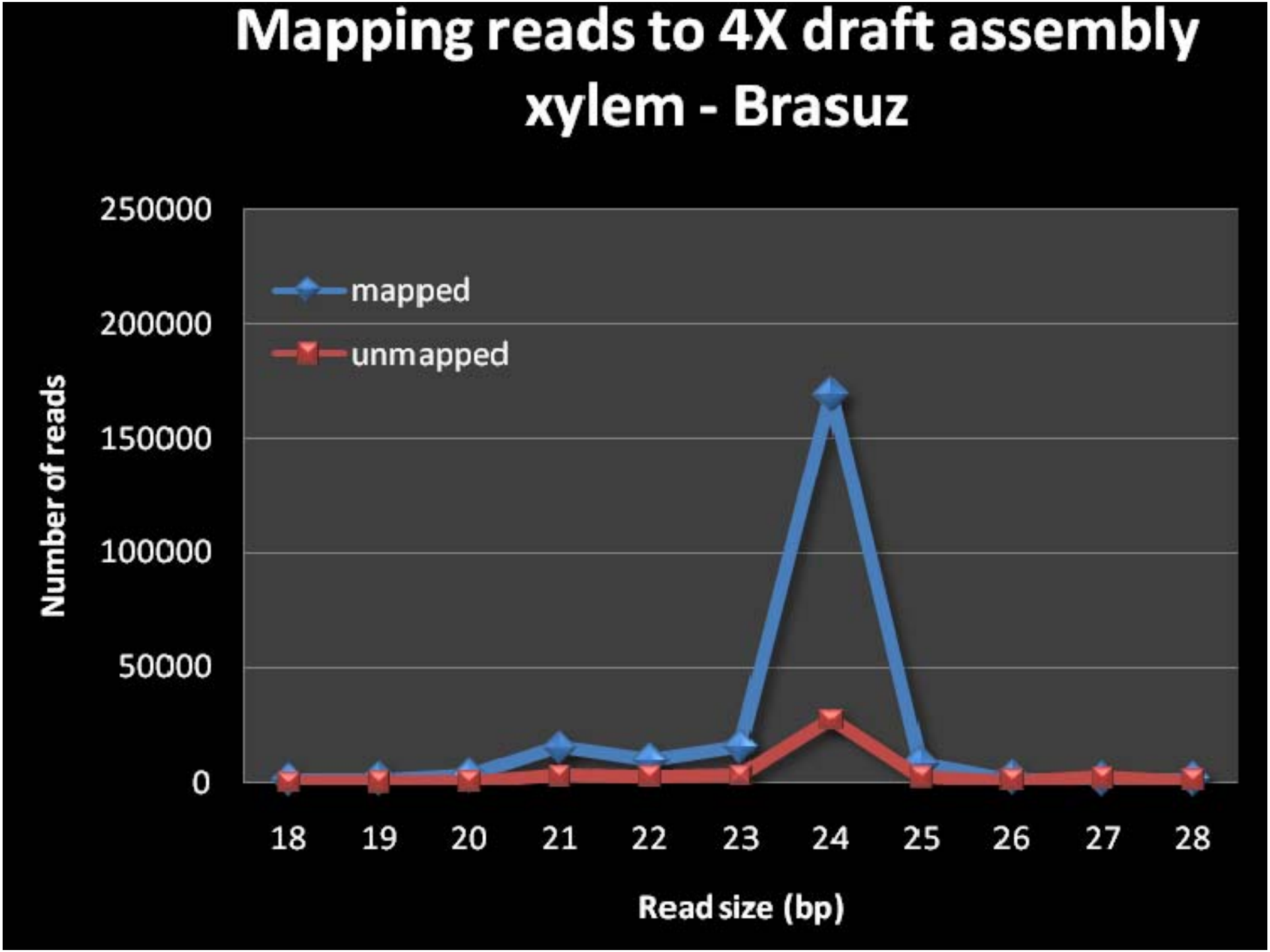
Not all small RNAs are miRNAs!

The Plant Cell, Vol. 20: 3186–3190, December 2008, www.plantcell.org

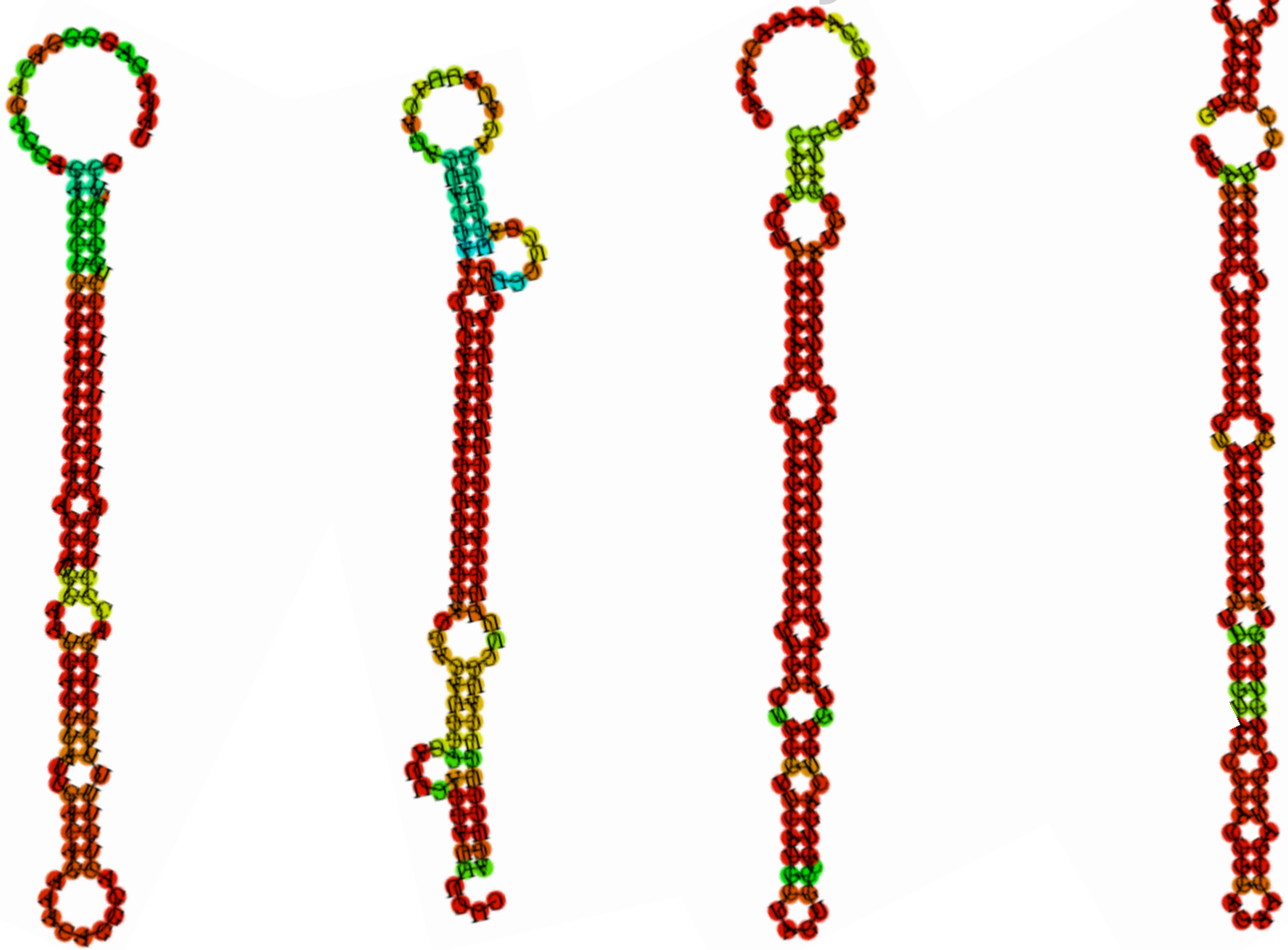
COMMENTARY

Criteria for Annotation of Plant MicroRNAs

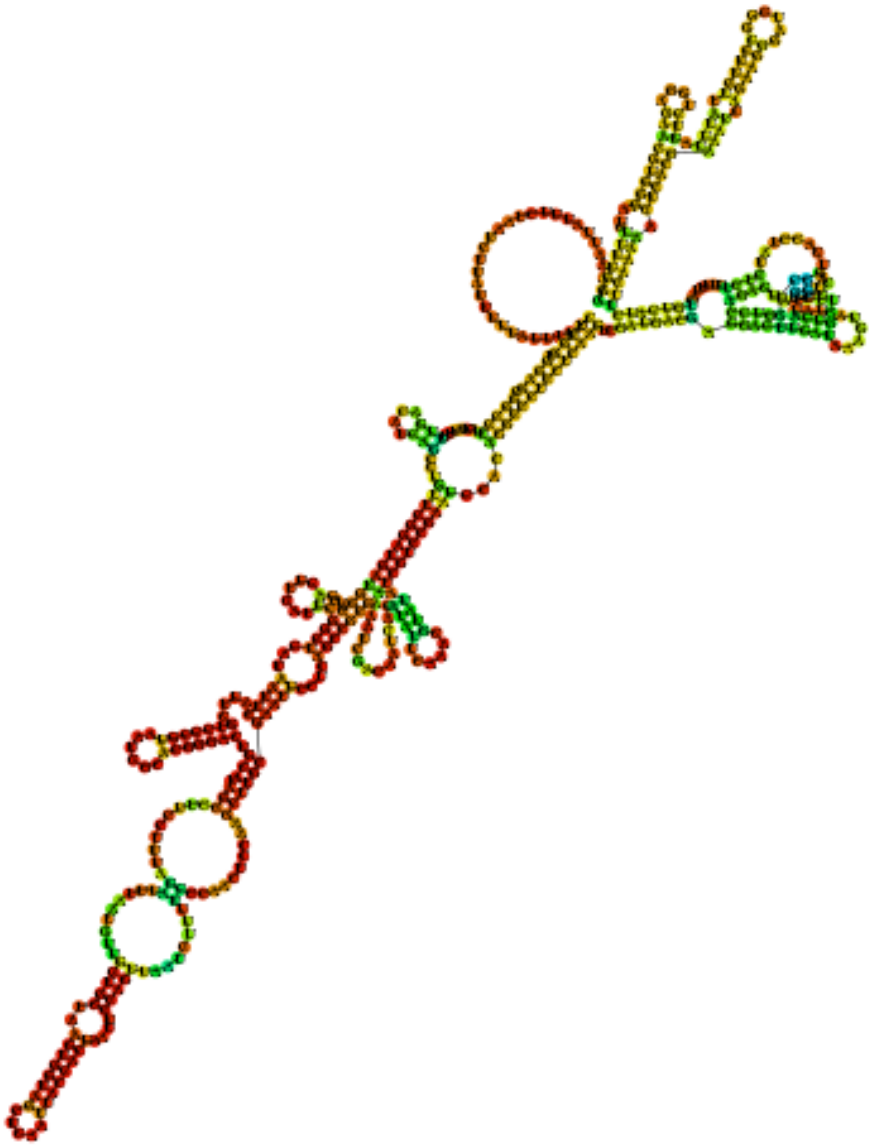
Blake C. Meyers,[a,1] Michael J. Axtell,[b,1] Bonnie Bartel,[c] David P. Bartel,[d] David Baulcombe,[e] John L. Bowman,[f] Xiaofeng Cao,[g] James C. Carrington,[h] Xuemei Chen,[i] Pamela J. Green,[a] Sam Griffiths-Jones,[j] Steven E. Jacobsen,[k] Allison C. Mallory,[l] Robert A. Martienssen,[m] R. Scott Poethig,[n] Yijun Qi,[o] Herve Vaucheret,[l] Olivier Voinnet,[p] Yuichiro Watanabe,[q] Detlef Weigel,[r] and Jian-Kang Zhu[i]


Mapping reads to 4X draft assembly xylem - Brasuz
250000
200000
150000
100000
50000
0
Number of reads
mapped
unmapped
18
19
20
21
22
23
24
25
26
27
28
Read size (bp)

Precursor secondary structure from draft assembly

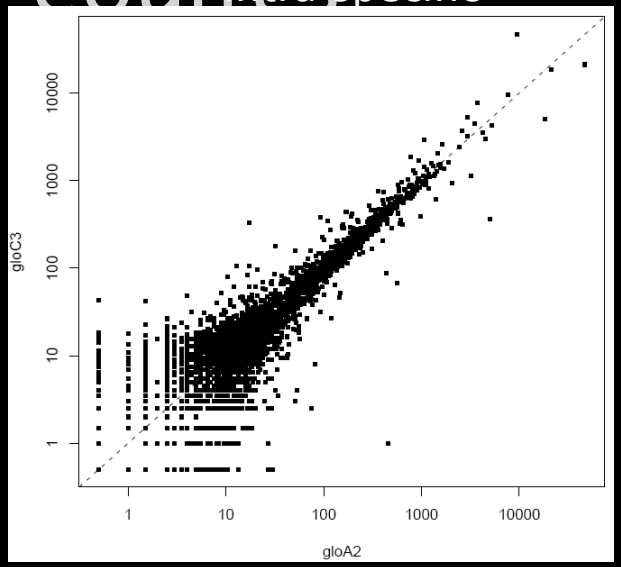


False positive

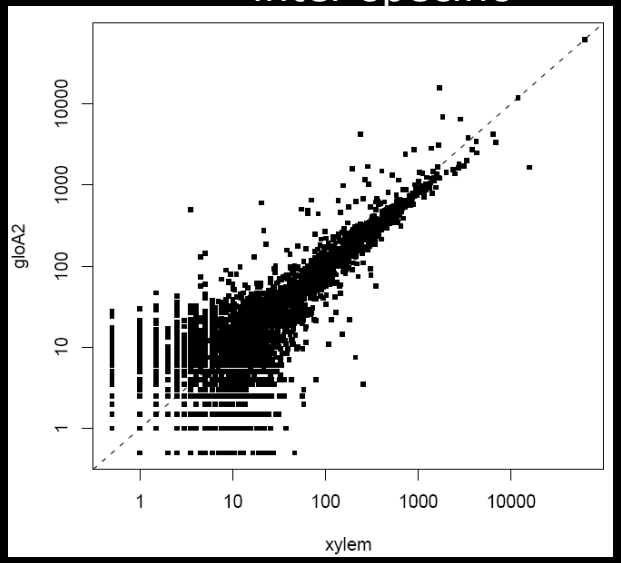


Pairwise comparison of tag counts

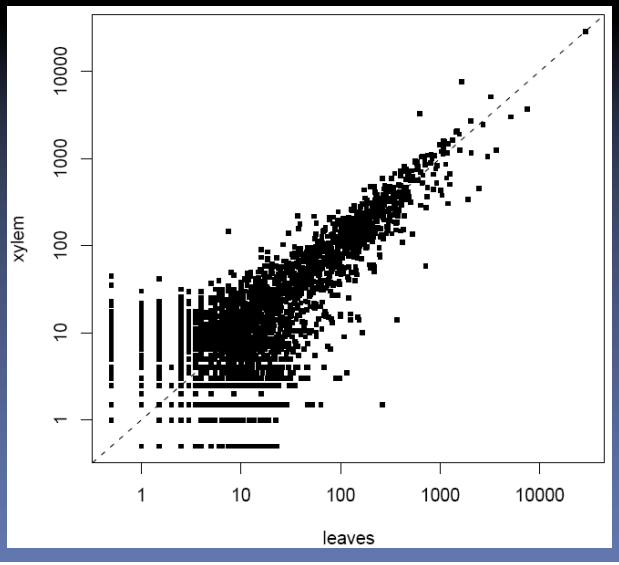
Intra-specific



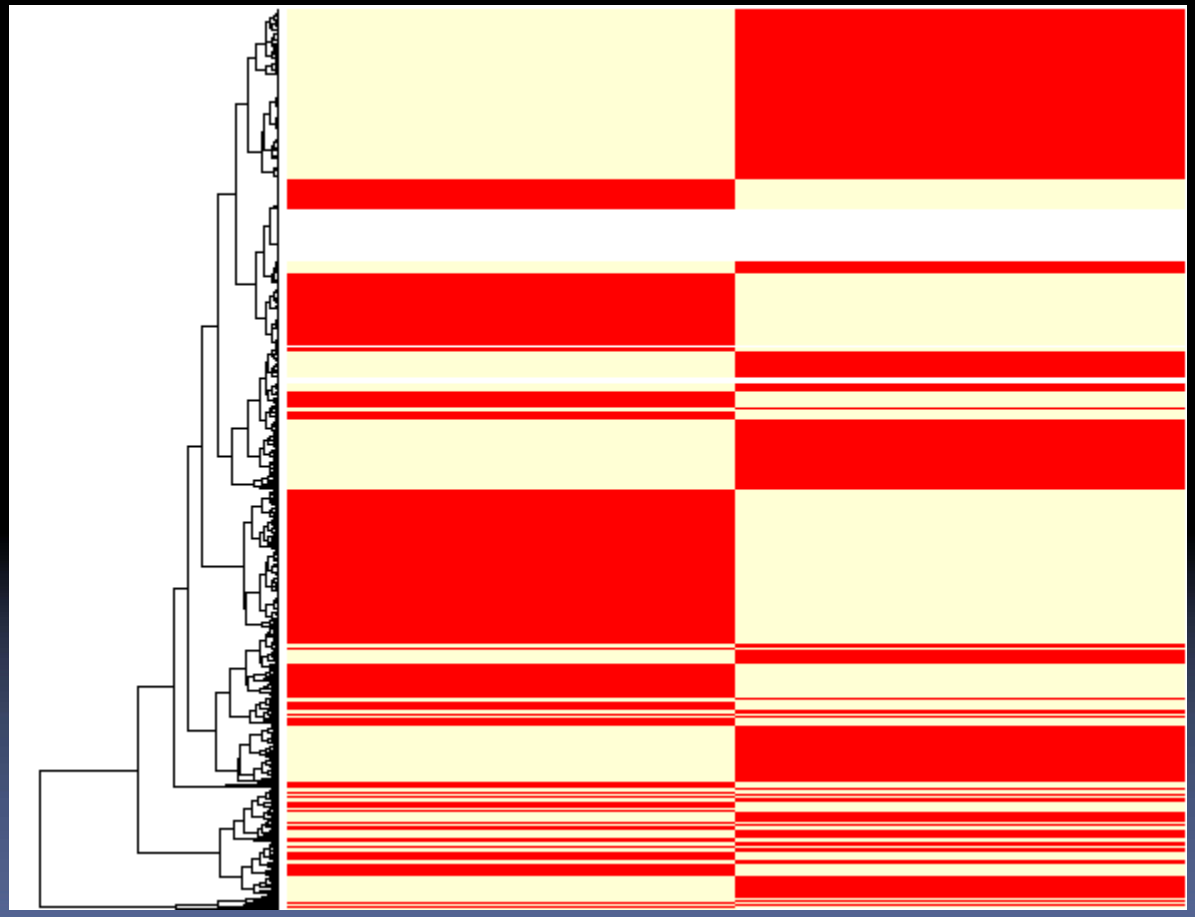
Inter-specific

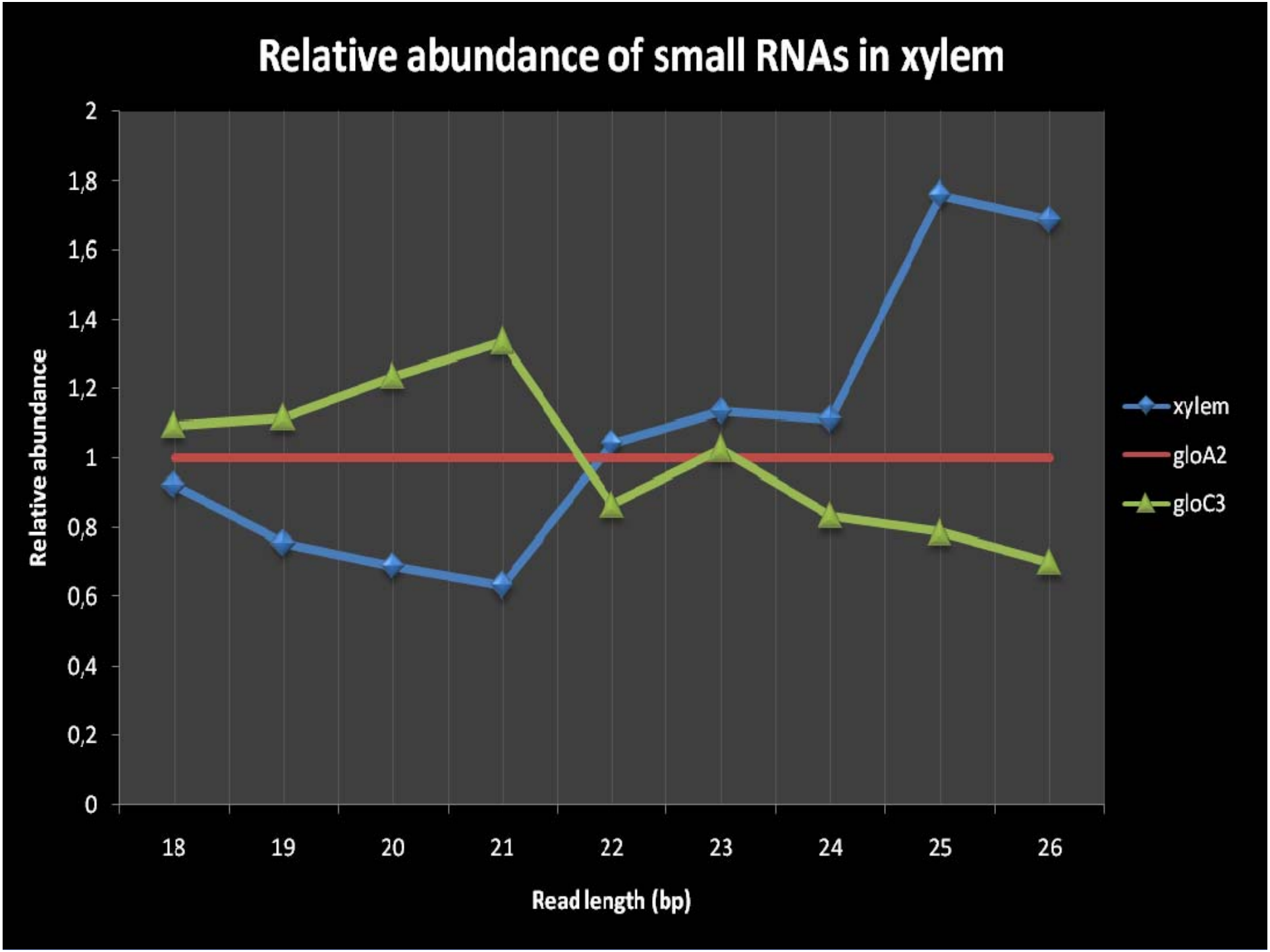


Tissue-specific

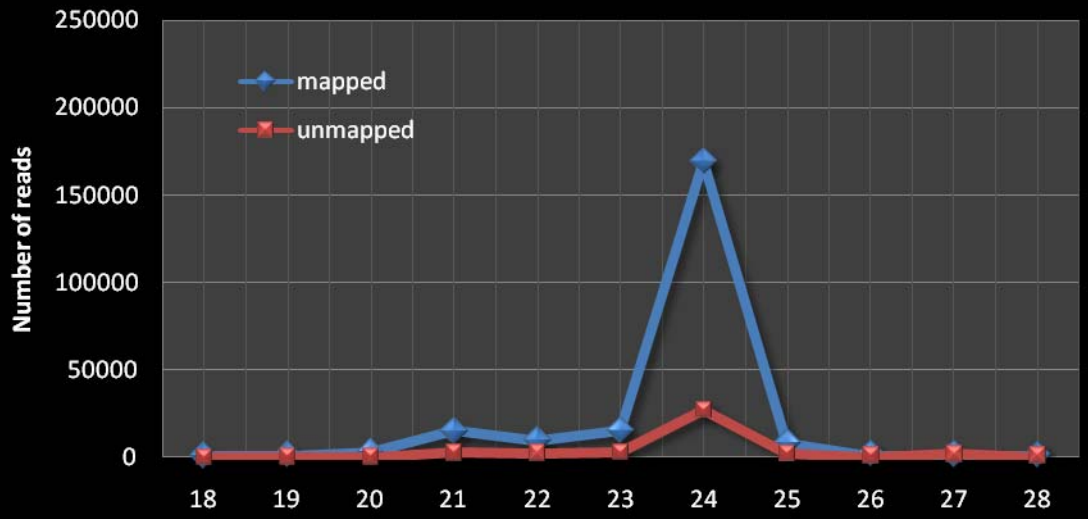


Differential expression in miRNAs

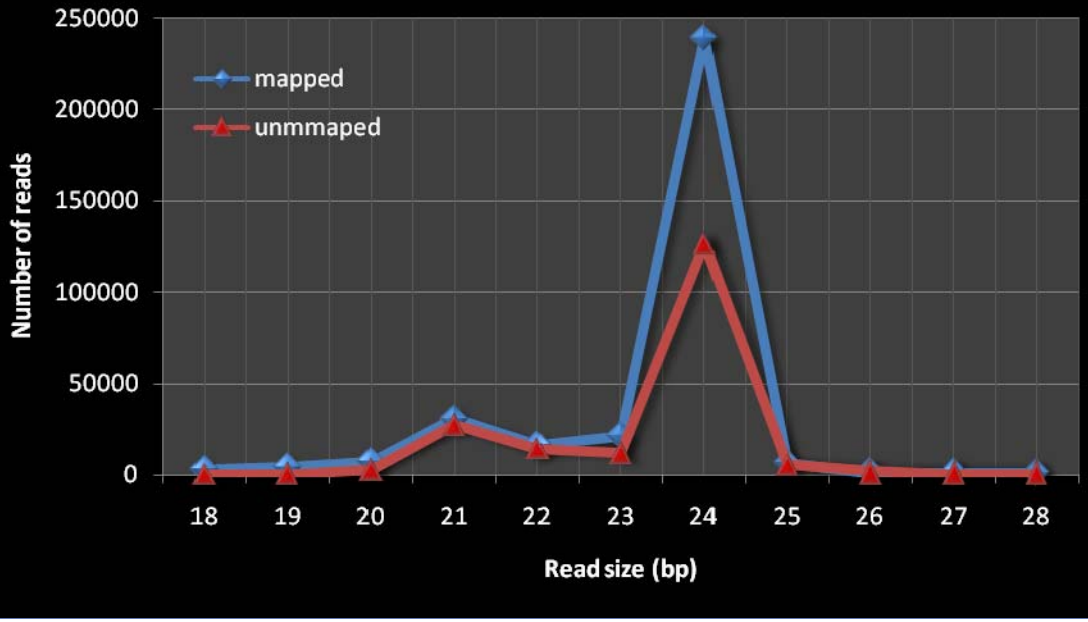



Relative abundance of small RNAs in xylem
Relative abundance
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
18
19
20
21
22
23
24
25
26
Read length (bp)
xylem
gloA2
gloC3

Mapping reads to 4X draft assembly xylem - Brasuz



Mapping reads to 4X draft assembly xylem - gloA2



Expression profiling in rice-*Magnaporthe grisea* pathosystem

- Rosângela Bevitori (EMBRAPA)
- Maria de Fátima Grossi de Sá
- Rodrigo da Rocha Fragoso
- Isabela Tristan
- Blake Meyers (University of Delaware)

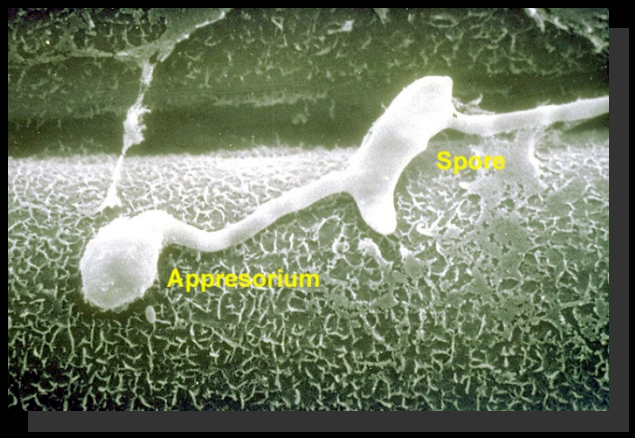
Experiment 2: Objective

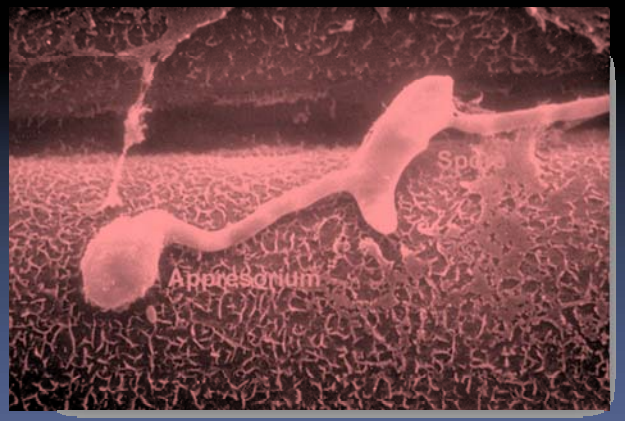
- Study the global patterns of gene expression of the plant-pathogen interaction at the onset of infection

Biological material

Cultivar METICA 1

Virulent strain





Avirulent strain

New horizons for gene expression profiling

The beginning of the end for microarrays?

NATURE METHODS | VOL.5 NO.7 | JULY 2008 | 585

Jay Shendure

Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. **Heidi Ledford** looks at a genome technology facing intense competition.

NATURE|Vol 455|16 October 2008

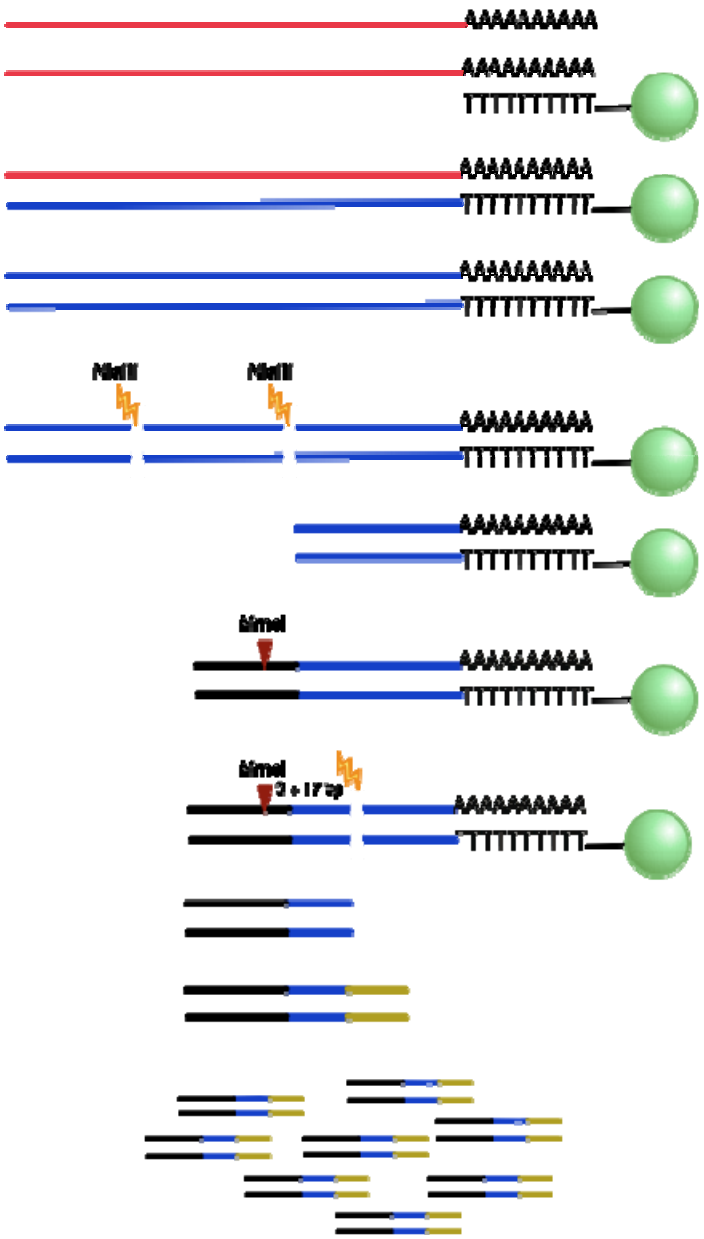
Microarray drawbacks in transcript quantification

- Indirect quantification by light intensity
  - Analogic!
- Hybridization-based
  - Relative expression measurement
  - Background and cross-hybridization problems
  - Problems with paralogs and sequence polymorphisms
  - Closed platform (only predefined sequences are detected)
- Reproducibility issues
- Dynamic range

Digital Gene Expression (DGE)

- Measures global expression profiles by direct sequencing
- Tag based technology generating 21 bp from 3' end of an mRNA population
- Similar to SAGE (*Serial Analysis of Gene Expression*), but by using Illumina GA it avoids time consuming technical manipulations

Steps In Preparing a Tag Profile Library



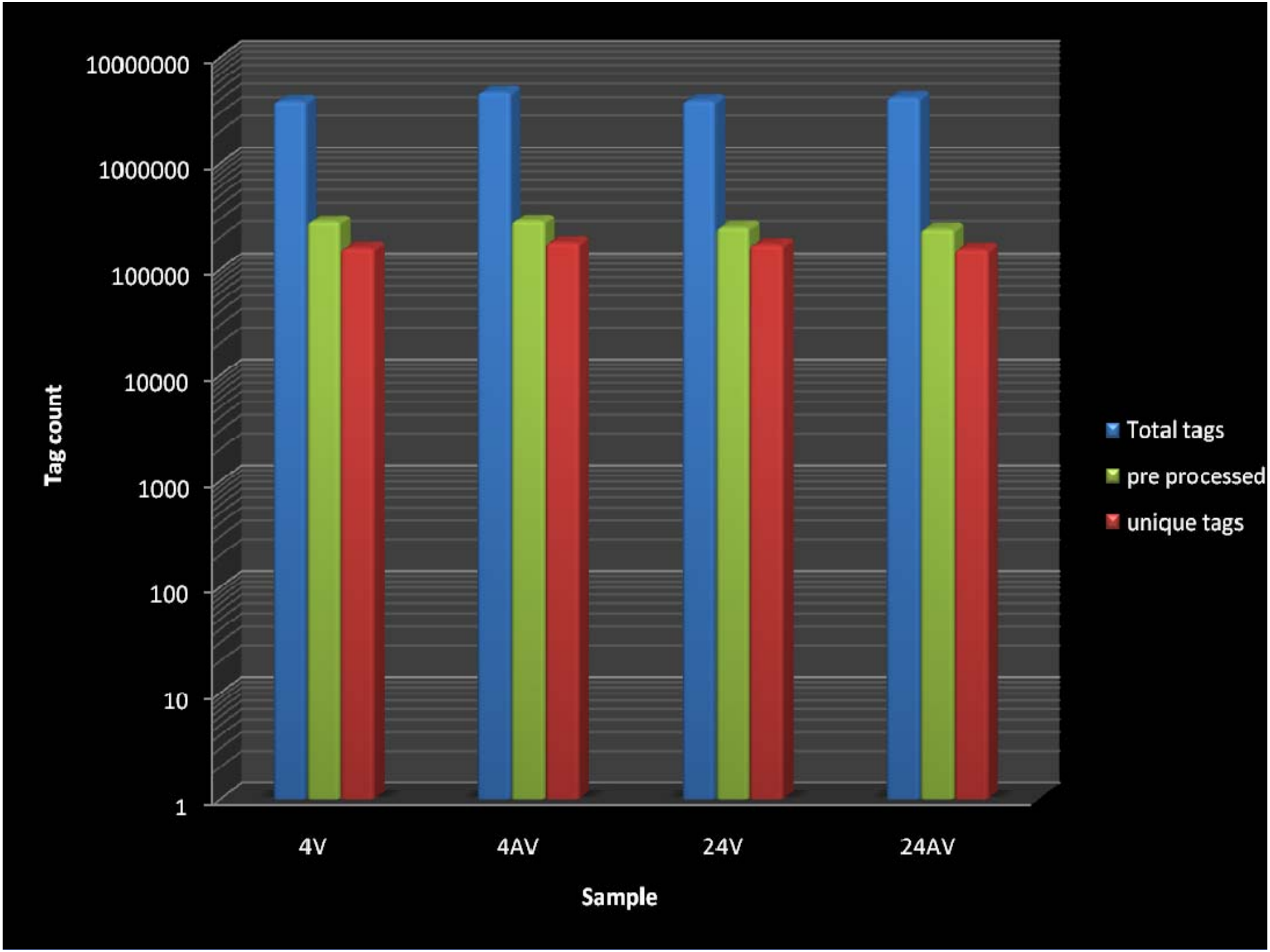
1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Synthesize first strand cDNA
4. Synthesize second strand cDNA
5. Digest with NlaIII (CATG)
6. Isolate 3' fragment
7. Ligate Vmel adaptor
8. Digest with Mmel (TCCRAC[N]$_{20}$)
9. Isolate fragments with 17 bp tags
10. Ligate adaptors
11. PCR amplify fragments, sequence
12. Analyze millions of 17 bp tags

Results

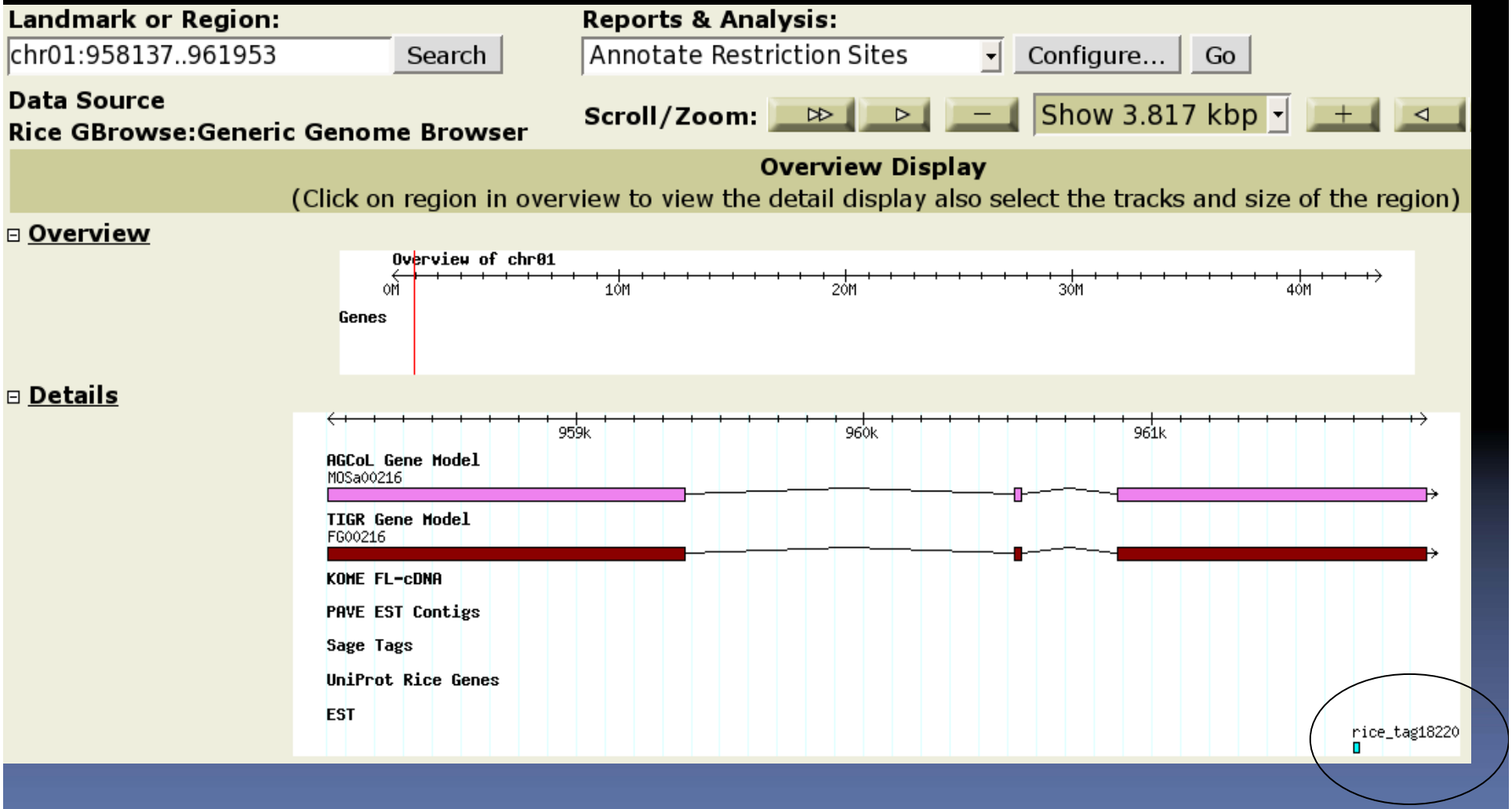
Sample	Total tags
4V	3,924,878
4AV	4,715,500
24V	3,920,177
24AV	4,220,170

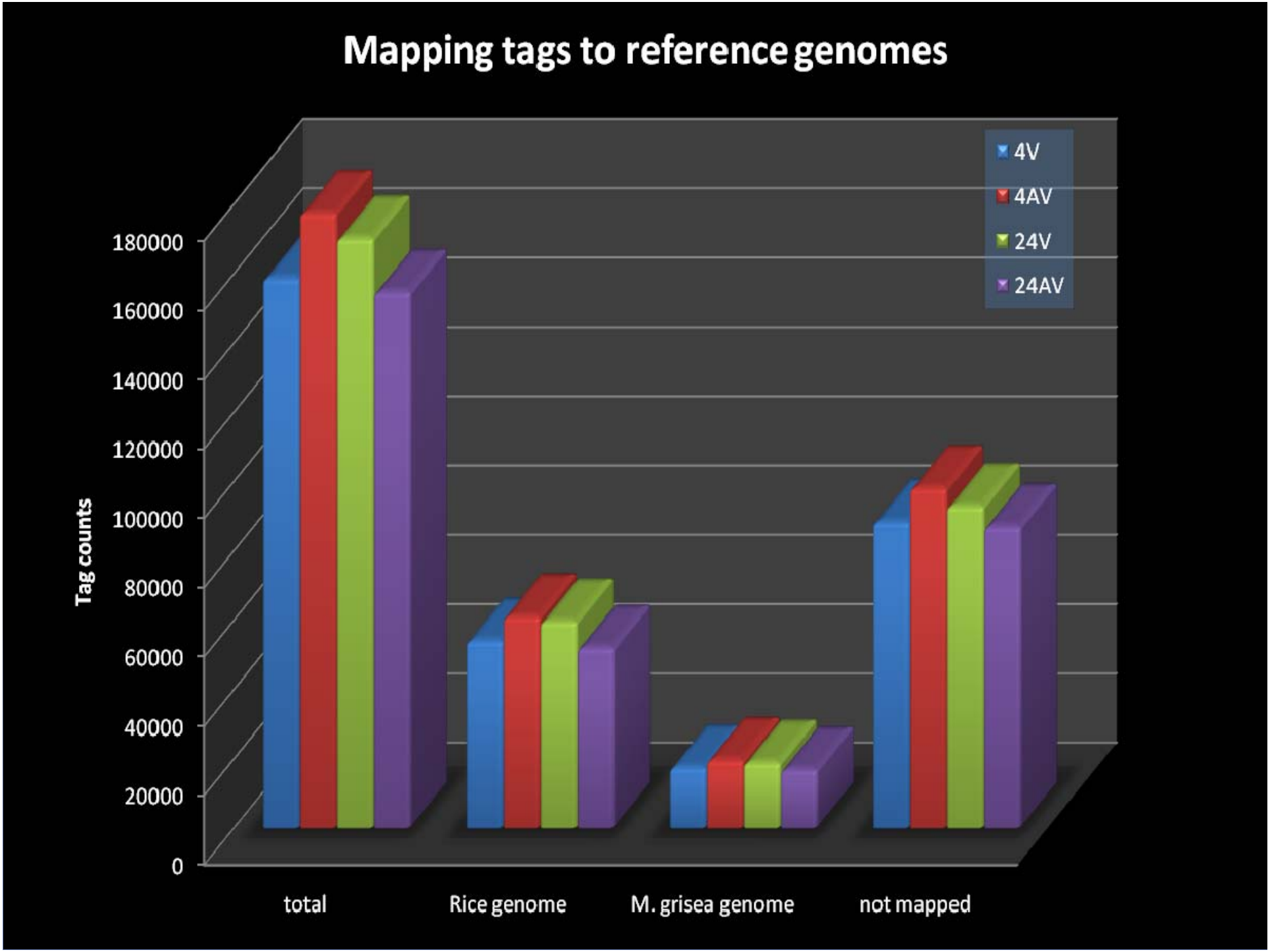
"… Before the development of deep sequencing technology, construction of a large scale SAGE library containing up to 100 000 canonical tags would Typically take 1 year and a considerable finacial investment."

Hoen et al. (2008) – doi:10193/nar/gkn705

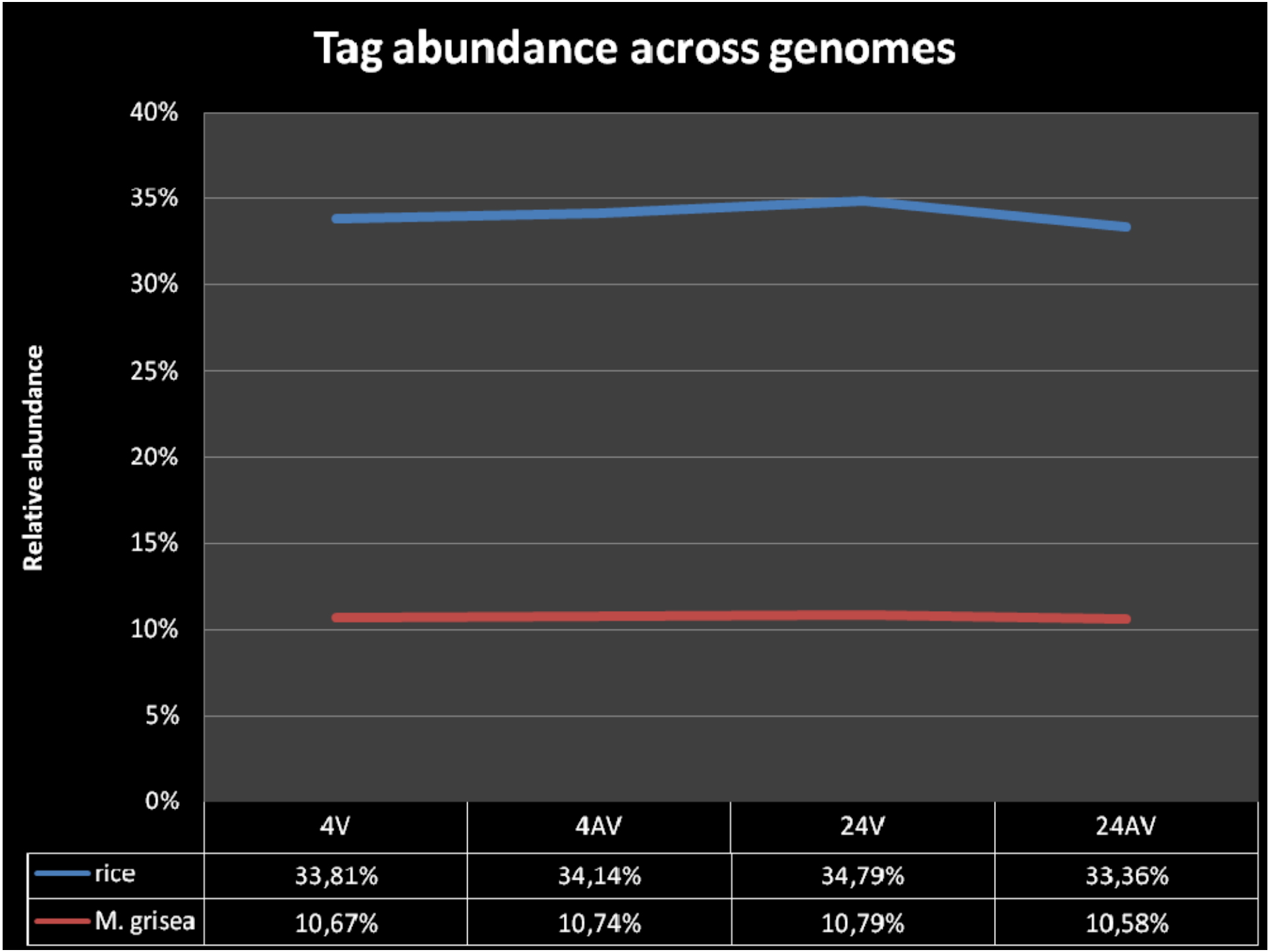

10000000
1000000
100000
10000
1000
100
10
1
Tag count
4V
4AV
24V
24AV
Sample
Total tags
pre processed
unique tags

Mapping a tag to a canonical region in the rice genome



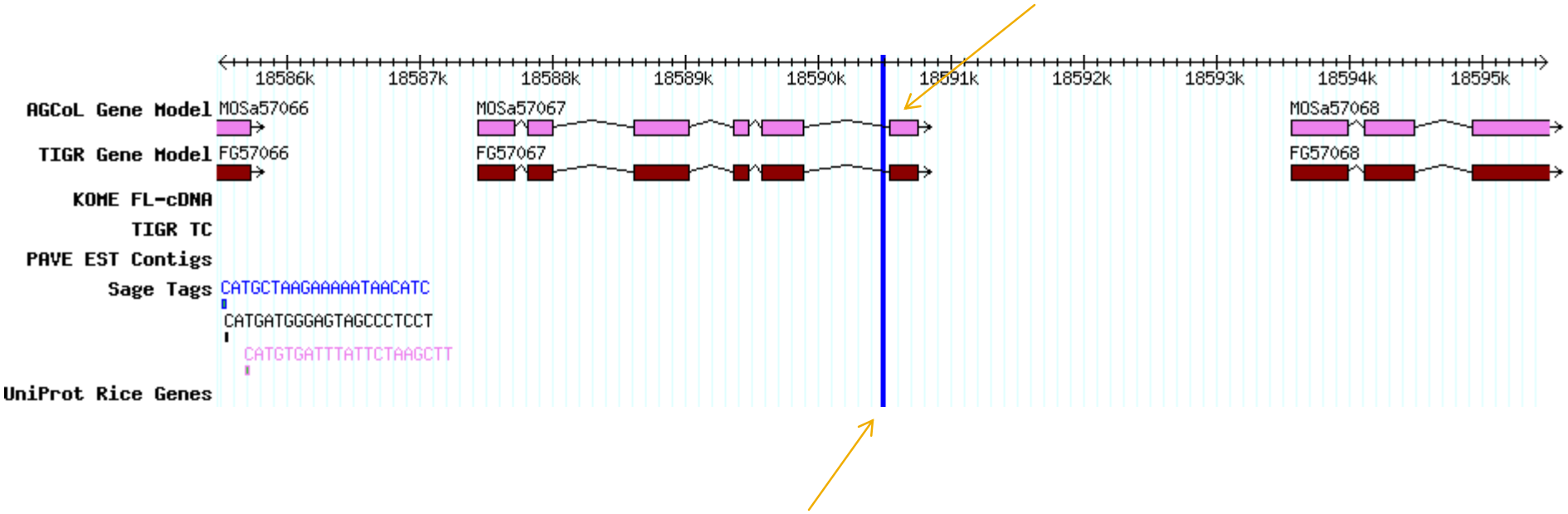

Mapping tags to reference genomes
Tag counts
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
4V
4AV
24V
24AV
total
Rice genome
M. grisea genome
not mapped

Tag abundance across genomes



	4V	4AV	24V	24AV
rice	33,81%	34,14%	34,79%	33,36%
M. grisea	10,67%	10,74%	10,79%	10,58%

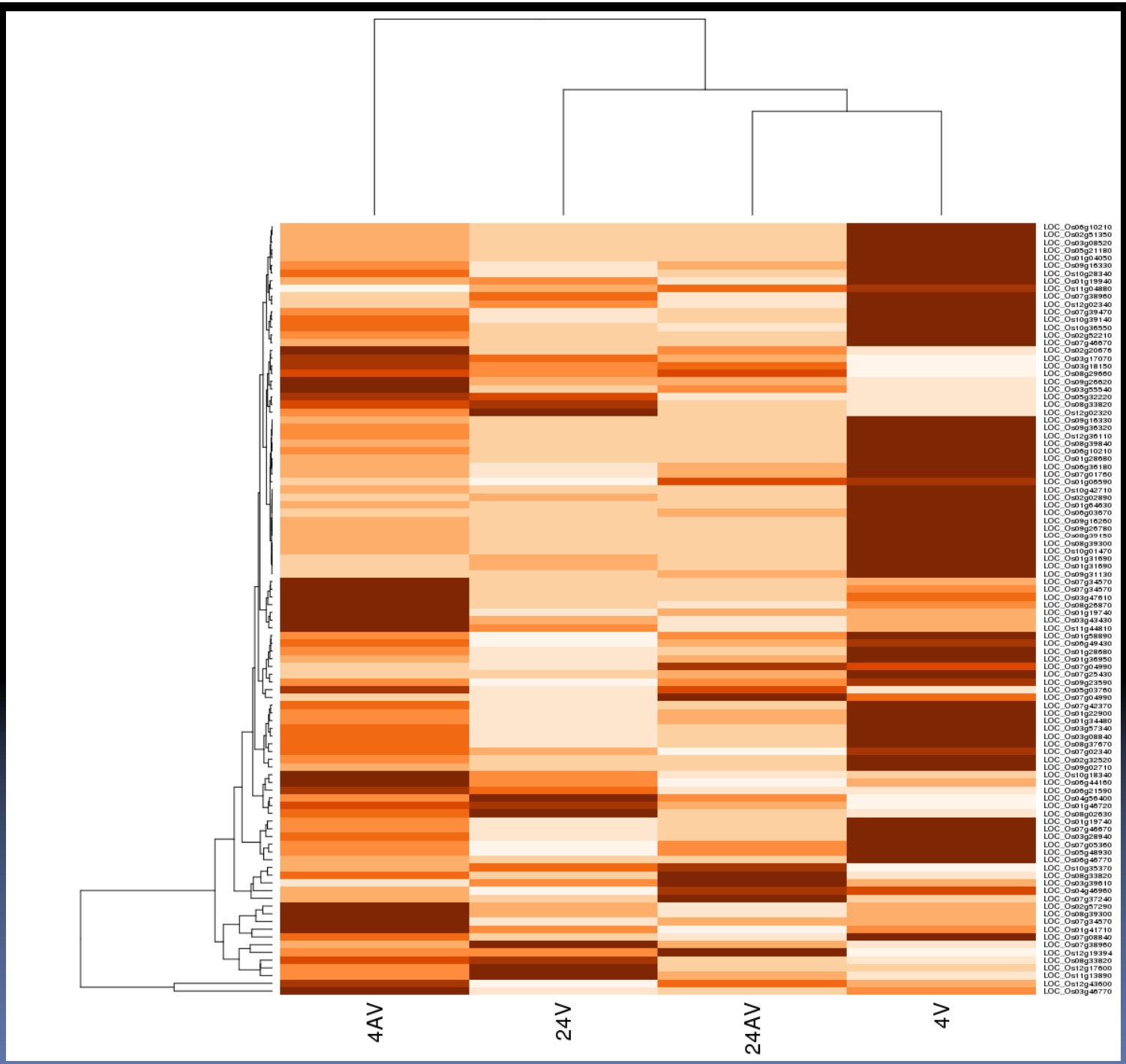
Intron mapping tags



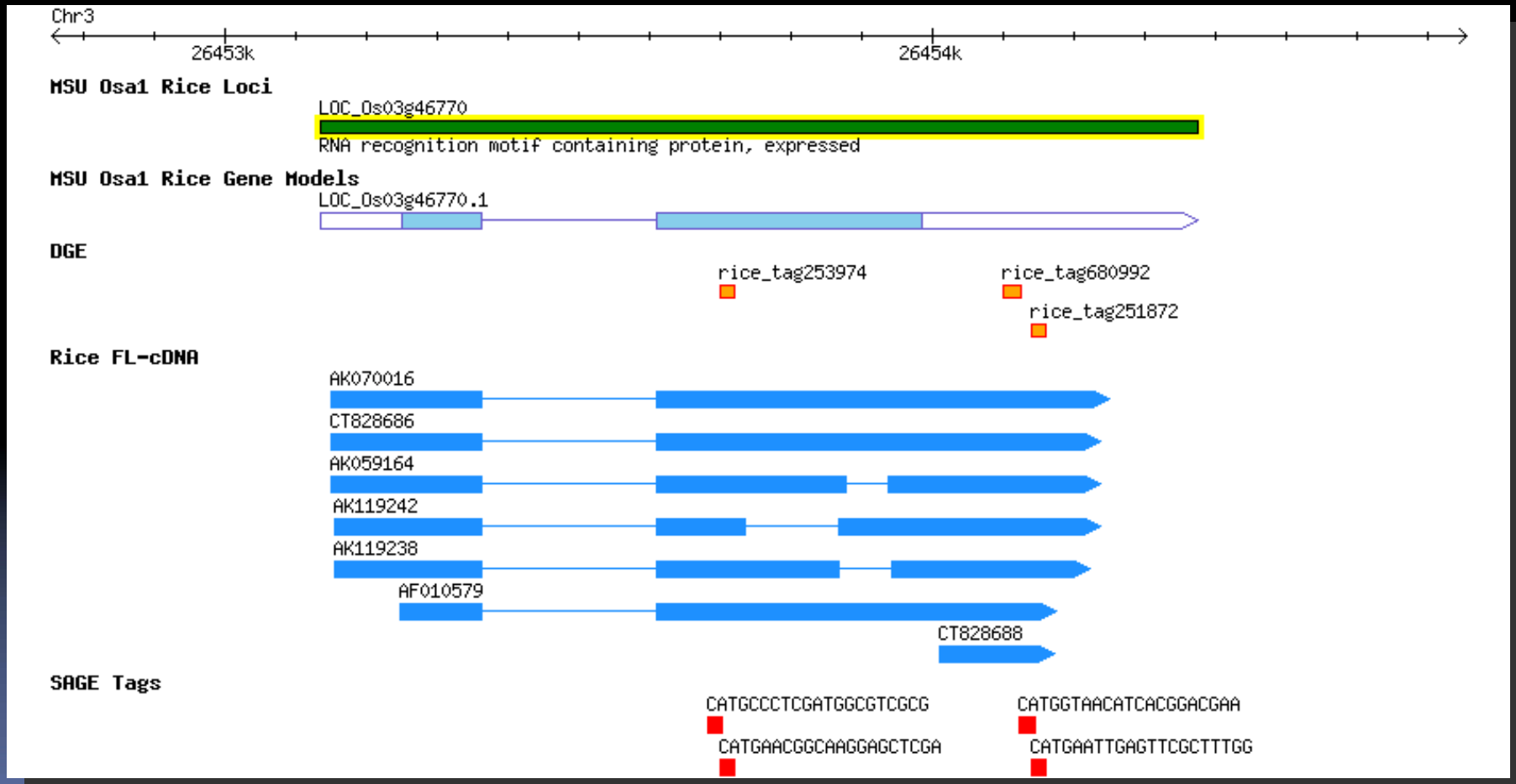
- This example clearly shows the advantage of new types of transcripts that can be detected by DGE, and not by microarrays
- Captures relevant biological phenomena like alternative splicing, alternative poly-adenylation, anti-sense transcription

Differentially expressed genes after pairwise sample comparison using Audic-Claverie test and FDR

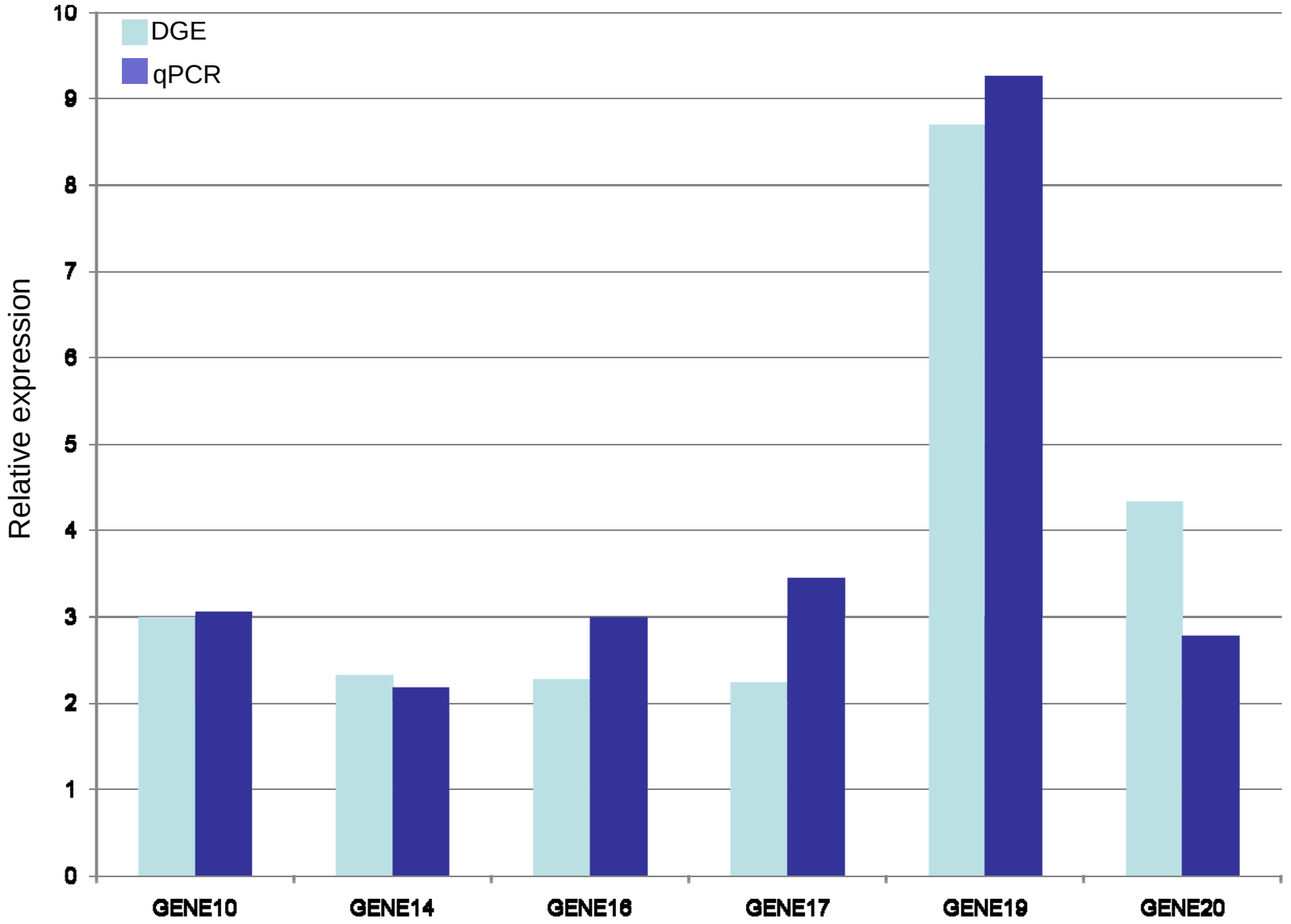
id	tag	4V	4AV	24V	24AV
rice_tag251872	CATGAATTGAGTTCGCTTTGG	24343	50953	3325	4669
rice_tag323395	CATGAGCCACCTGAATGAGAA	7319	3075	709	370
rice_tag344276	CATGTAGAAGCAAAGCAAATG	11064	7130	1089	1506
rice_tag273881	CATGCTAACTCTGGTTGTTAG	15551	9358	5582	4316
rice_tag411704	CATGTGTACTAATGATTCTAC	10936	4967	542	1737
rice_tag53092	CATGTACTGTGGTTGTTGATC	7049	15047	21939	18160
rice_tag449667	CATGCTTTTTTTCTTTTGCC	4762	2142	111	833
rice_tag501077	CATGTGTTAGTATTTGTATAT	12172	5223	1420	3159
rice_tag176600	CATGGAGAATGTTATTATTTT	6703	1553	580	701
rice_tag197132	CATGAAAGAAAGAAATGGTCT	3223	1079	202	454
rice_tag124497	CATGATAGTAACAATTGATAA	3092	964	230	422
rice_tag613208	CATGTATCAATGTATTTTTAT	7168	3107	1743	1944
rice_tag244520	CATGTACATTAGTATAGGTTT	5232	2317	301	1362
rice_tag344329	CATGTGTACTAATGATTCTTC	3511	639	59	233
rice_tag596387	CATGAATAAAGGTCTTTAATC	3924	878	355	382
rice_tag326469	CATGAATACATTTTTTGTTG	2485	286	81	115
rice_tag253336	CATGTATAAATTGTAATTGTA	2324	258	52	131
rice_tag188431	CATGTGTACTAATGATTCTCG	2173	313	22	170
rice_tag696977	CATGTGATGAATATTACAAAT	4765	1462	1520	1459
rice_tag558113	CATGCATAATTAAGCTTGAAC	8031	4534	3519	3869
rice_tag610478	CATGGCGCAGGAGGTGCTTCT	1806	6763	6982	5882
rice_tag634235	CATGGAAGAACTATAATGAAT	9437	2936	2123	2626
rice_tag171902	CATGTATATAATCTTCACTAG	9320	5570	2127	5290
rice_tag602908	CATGTAAATACGAATGGGAGA	8441	4700	1142	5292
rice_tag590883	CATGTTCGGCTGCACCGATGC	22700	47158	62966	18734
rice_tag203051	CATGGCCAAGTTATCTATCTA	34094	61737	12449	43973
rice_tag36135	CATGGATCCGTCTCTCTGGGA	4596	18088	395	3164
rice_tag585302	CATGGTAGCAGCAGTGTTCTA	3194	9868	10443	4473
rice_tag138765	CATGACGTTTCCGCGGGACTG	2744	7851	11167	3803
rice_tag42830	CATGCTCCGTTTCCTCGATAA	3841	1376	1192	677
rice_tag687978	CATGGGGTTGTAATTAATTGC	2071	6433	744	8070
rice_tag410557	CATGGTATATATGAATAAAAA	3901	1046	2155	2598
rice_tag412884	CATGGACCGCTTCGCCCCTGC	214	2509	941	239
rice_tag497205	CATGGTGTTTTTGAATAGCAG	6460	16919	7210	2883
rice_tag531188	CATGGATGAGTGGGGAGTGGA	5001	14122	4573	2416
rice_tag405722	CATGCGAAATCGATTCCGAGT	14084	10337	7280	5692
rice_tag330081	CATGTCAATAAATTTCCTTGC	8338	5155	2193	9257
rice_tag474974	CATGTATAATACTTTGAGACC	3413	1271	752	1268


4AV
24V
24AV
4V

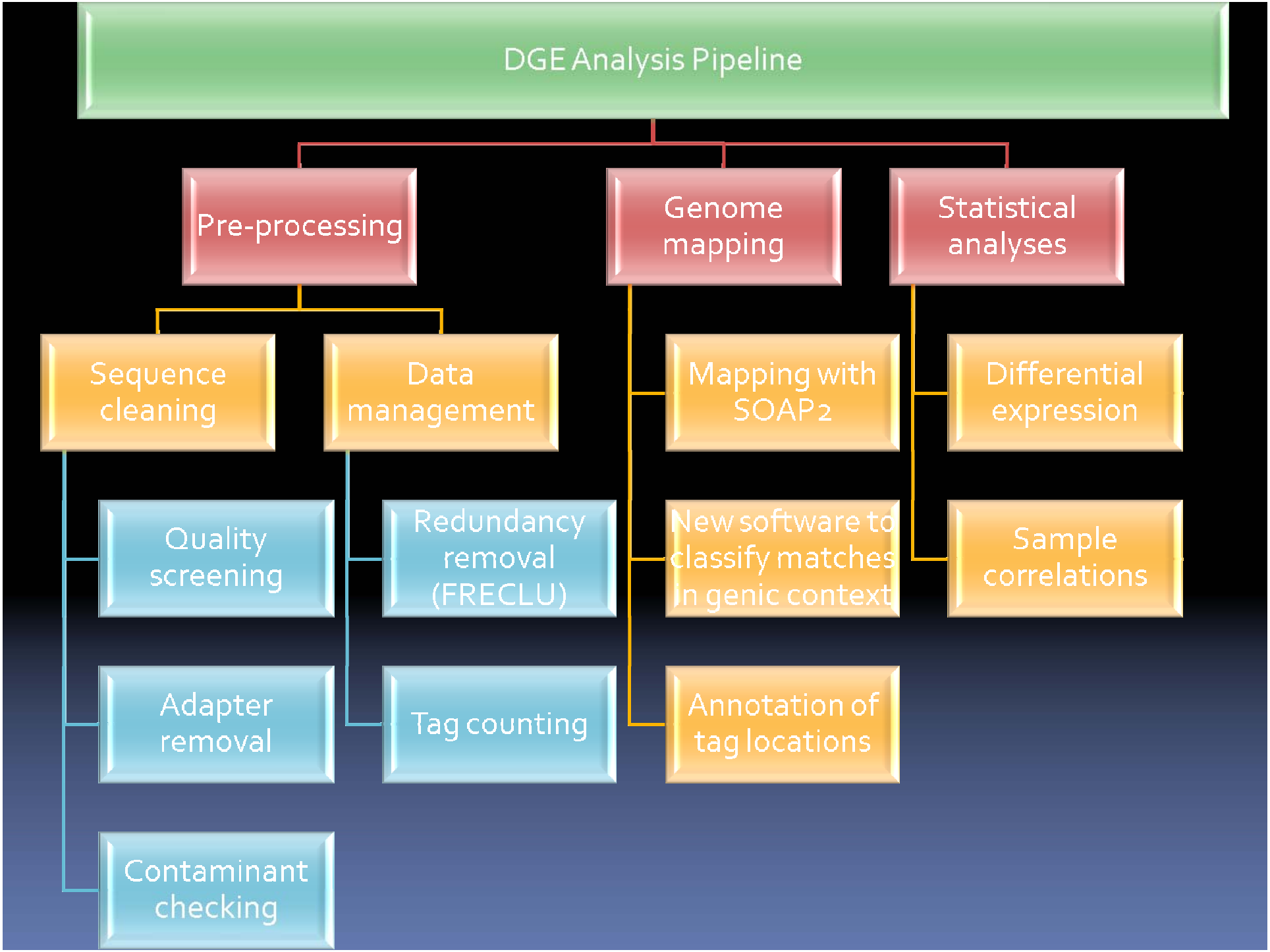
Tags for the most differentially expressed gene



Validation of expression levels between DGE and qPCR



Selected genes in 4V sample

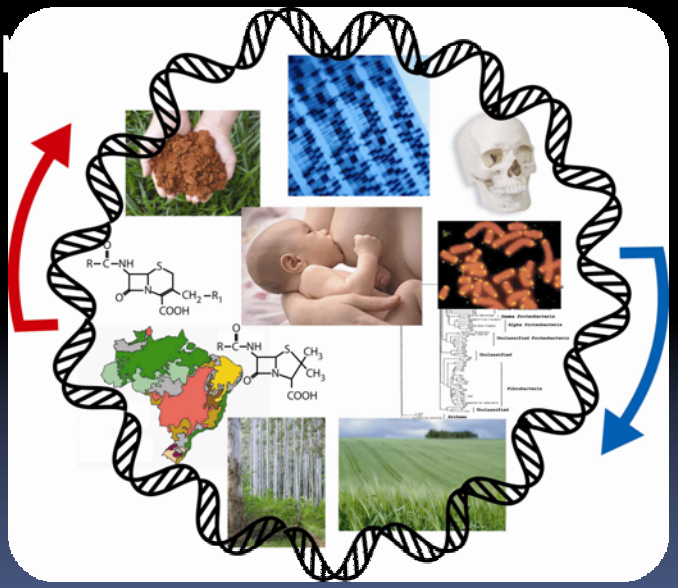

DGE Analysis Pipeline
Pre-processing
Genome mapping
Statistical analyses
Sequence cleaning
Data management
Mapping with SOAP2
Differential expression
Quality screening
Redundancy removal (FRECLU)
New software to classify matches in genic context
Sample correlations
Adapter removal
Tag counting
Annotation of tag locations
Contaminant checking

Conclusions

- Illumina GA generates DNA sequences at an unprecedented level and is taking Biology by storm
  - New genomics era
- The shear amount of data changes the perspective of how biological phenomena are analyzed
- Data analysis is the current bottleneck

Genomics Center of Distrito Federal

- Next generation sequencing platform in Brasilia, funded entirely by the state governme



Platforms

Illumina
Genome Analyzer

454 Genome Sequencer
Titanium

Participants

- Universidade Católica de Brasília
- Universidade de Brasília
- EMBRAPA
  - Recursos Genéticos e Biotecnologia
  - Agroenergia
- LACEN
- Polícia Civil do DF

Georgios Pappas Jr
gpappas@cenargen.embrapa.br

The deluge
Gustave Doré
(1832-1883)