Expression Data Types and Their Use in Annotation

Shu Ouyang
souyang@jcvi.org

TIGR Rice Annotation Workshop
May 24, 2007

Gene Expression

- Gene expression is a reflection of the rate of transcription and the turnover of the mRNA: collectively this is the mRNA accumulation
- Correlations are made between mRNA accumulation and physiological conditions such as stages of development OR between normal and mutant cells
- Three patterns of expression:
  - Up regulation
  - Down regulation
  - Constitutive
- Similar patterns of expression indicate coordinated transcription which reflects similarities in gene regulation

What is Expression Data?

- Data captured that represent the transcript population and transcript structure in a cell
  - Sequence of the transcript
  - Frequency of the transcript in a mRNA population
  - Pattern of expression of the transcript
  - Modified transcripts (alternative splice forms)

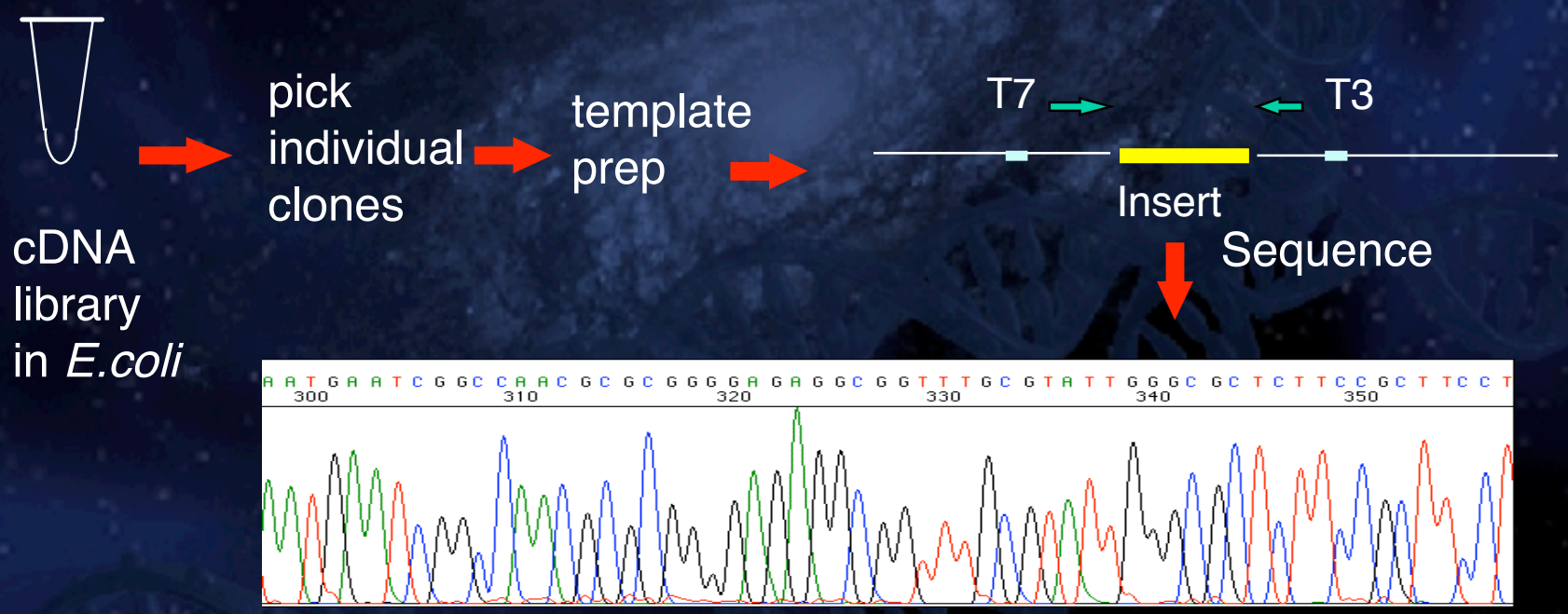
Expression Data Types

- ESTs = Expressed Sequence Tags
- Full length cDNAs
- MPSS = Massively Parallel Signature Sequencing
- SAGE = Serial Analysis of Gene Expression
- Microarrays

* All have utility in annotation, either structural or functional. Some data types are more powerful than others.

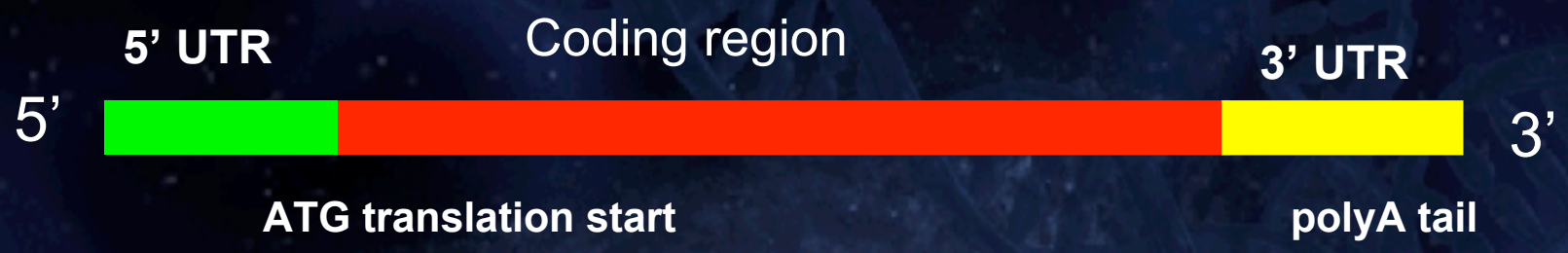
Expressed Sequence Tags (ESTs)

- What is an EST?
  - single pass sequence from cDNA
  - specific tissue, stage, environment, etc.
  - Make a cDNA library
  - End sequence (mostly one end, sometimes both)



Expressed Sequence Tags (ESTs)

- 5’ versus 3’ ends



- Most ‘regular’ cDNA libraries are NOT FL and thus will be partial cDNA clones, thus 5’ has a higher chance to reveal coding regions as these are more conserved
- Yet 3’ ends will reveal divergence and allow for separation of gene family members (paralogs)
- 3’ end sequencing technically more difficult than 5’ (typically get shorter read lengths)
- Costs are prohibitive to most plant EST projects to do both 5’ and 3’ sequencing.

Status of EST Sequencing Projects

Homo sapiens (human)	8,109,026
Mus musculus + domesticus (mouse)	4,840,638
Danio rerio (zebrafish)	1,350,105
Bos taurus (cattle)	1,318,108
Arabidopsis thaliana (thale cress)	1,276,690
Xenopus tropicalis	1,256,244
Oryza sativa (rice)	1,211,418
Zea mays (maize)	1,161,241
Triticum aestivum (wheat)	1,050,203
Rattus norvegicus + sp. (rat)	871,163
Ciona intestinalis	686,396
Xenopus laevis (African clawed frog)	677,784
Sus scrofa (pig)	646,392
Gallus gallus (chicken)	599,330
Drosophila melanogaster (fruit fly)	532,557
Hordeum vulgare + subsp. vulgare (barley)	437,713
Salmo salar (Atlantic salmon)	432,630
Glycine max (soybean)	371,817
Canis familiaris (dog)	365,909
Caenorhabditis elegans (nematode)	346,064

dbEST release
May 11, 2007

Number of public entries: 43,205,713

6 out the top 20 are plant species

Top plant/animal species:
Arabidopsis thaliana (1,276,690) vs Human (8,109,026)

http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

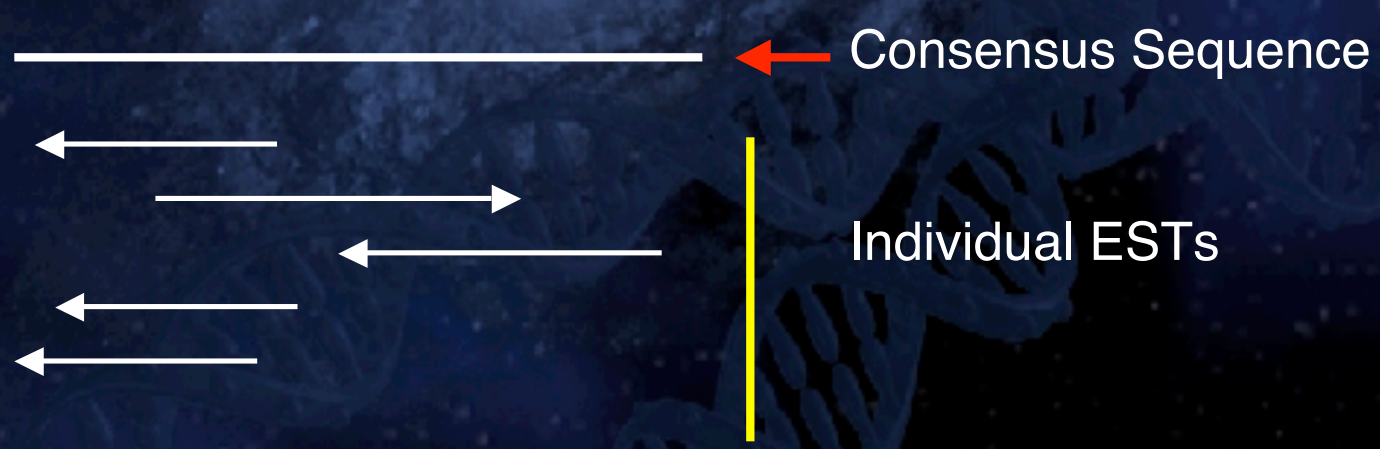
Uses of EST sequencing:

- Gene discovery
- Digital northerns/insights into transcriptome
- Genome analyses, especially annotation of genomic DNA

Issues with EST sequencing:

- Inherent low quality due to single pass nature
- Not 100% full length cDNA clones
- Redundant sequencing of abundant transcripts

Address through clustering/ assembly to build consensus sequences = Gene Index and TA



Current Rice EST Status

DFCI Rice Gene Index

About OsGI Gene Index

Development and Goals	Background Information about OsGI
Release Summary	display a statistical summary of all OsGI releases
Category Comparison	display estimated number of genes among all plant releases

Sequence Similarity Search

BLAST	search TC sequences based on sequence similarity

Sequence Reports

Identifiers or Keywords	search TC reports using TC identifiers, GB accessions or keywords
TC Annotator	list all TC annotation
EST Annotator	list all EST annotation
Libraries	search EST libraries by keywords or tissue origins
CAT# Download	download EST and TC sequences originating from one library

Functional Annotation and Analysis

Alternative Splice Forms	prediction of alternative splice variants
EST Expression	compare EST expression between different libraries or tissues
Gene Ontology	classification of TCs by GO vocabularies
Metabolic Pathways	association of TCs with metabolic and signaling pathways
Oligomer Prediction	list all 70-mer oligo predictions

Release 17.0 (June 20, 2006)

Input Sequences	
ESTs	1163134
ETs	114986
Output Sequences	
TC sequences	77158
singleton ESTs	85212
singleton ETs	19426
Total unique: 181796	

Current release of DFCI Rice Gene Index (former TIGR Gene Index):

1,163,134 ESTs + 114,986 ETs cluster and assemble
into 77,158 TCs, 85,212 singleton ESTs, and 19,426 singleton ETs
resulting in total of 181,796 unique sequences

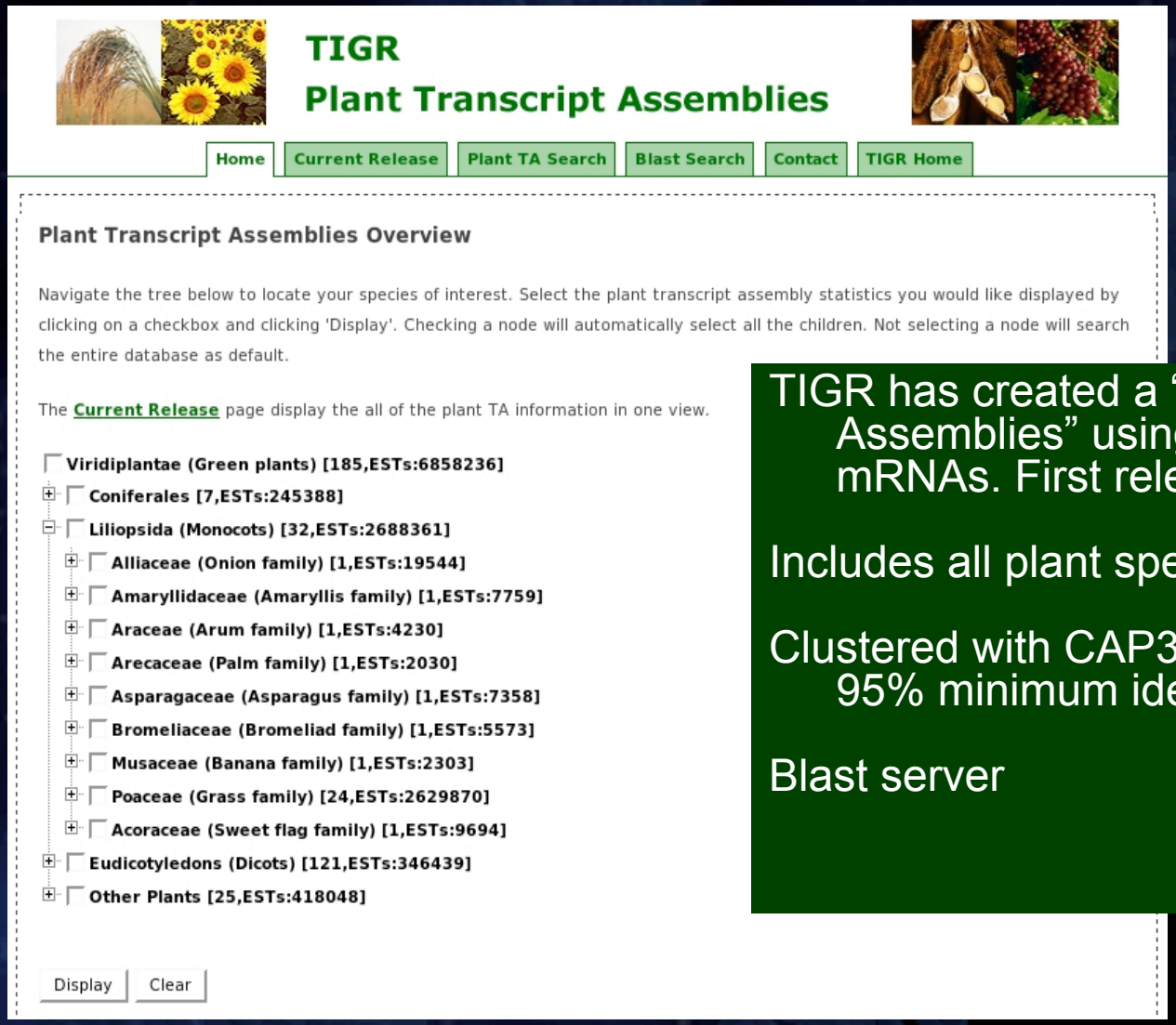
Gene Index Issues

- Collect ESTs, FL-cDNAs, mRNAs and any CDS in GenBank (includes genome annotation predicted genes).

- Thus, the gene index is not entirely an experimentally derived transcript assembly.

- Contains all *Oryza sativa* sequences, not just ssp. japonica or indica

Plant Transcript Assemblies



TIGR has created a “Plant Transcript Assemblies” using just ESTs, FL-cDNAs, mRNAs. First released early last year.

Includes all plant species with > 1,000 ESTs

Clustered with CAP3 (50 bp minimum match, 95% minimum identity)

Blast server

Rice Transcript Assembly

Total unique sequences

Taxon ID	Scientific Name	Common Name	EST Retrieval Date	Release	Transcript Assemblies			Transcript Assembly Components				Download FASTA
					Assemblies	Singletons	Total	EST	fl-cDNA	mRNA	Trash Count	
3702	*Arabidopsis thaliana*	Thale-cress	2006-06-05	2	27983	120385	148368	616064	65976	2204	0	Download
4530	*Oryza sativa*	Rice	2006-06-05	2	49870	197646	247516	1169591	34559	888	0	Download
4113	*Solanum tuberosum*	Potato	2006-06-05	2	26280	54792	81072	219485	988	608	0	Download
4565	*Triticum aestivum*	Bread wheat	2006-06-05	2	62121	257828	319949	840871	1832	554	0	Download
4577	*Zea mays*	Maize	2006-09-28	3	64026	220306	284332	1014701	4289	10132	129260	Download

Rice TA:

Note the higher number of singleton ESTs in the Rice TA as we did not use the genomic predictions to link ESTs together.

How to Use ESTs in Annotation

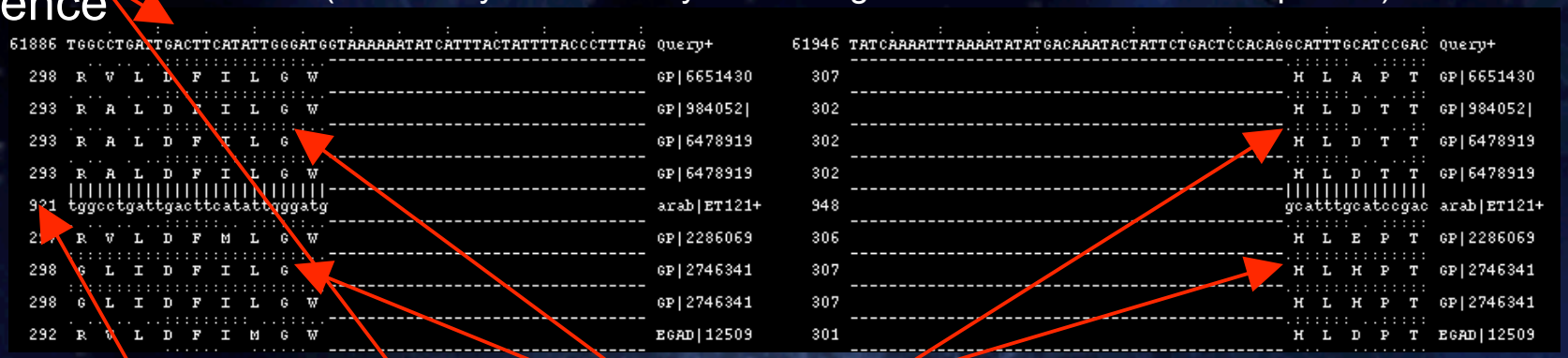
- Provide gene structure evidence (intron/exons)
- Provide expression evidence (hypothetical to expressed)
- Provide limited functional annotation information

How ESTs Aid in Structual Annotation

genomic sequence

GT ← intron → AG

(defined by discontinuity between genomic and cDNA/a.a. sequence)



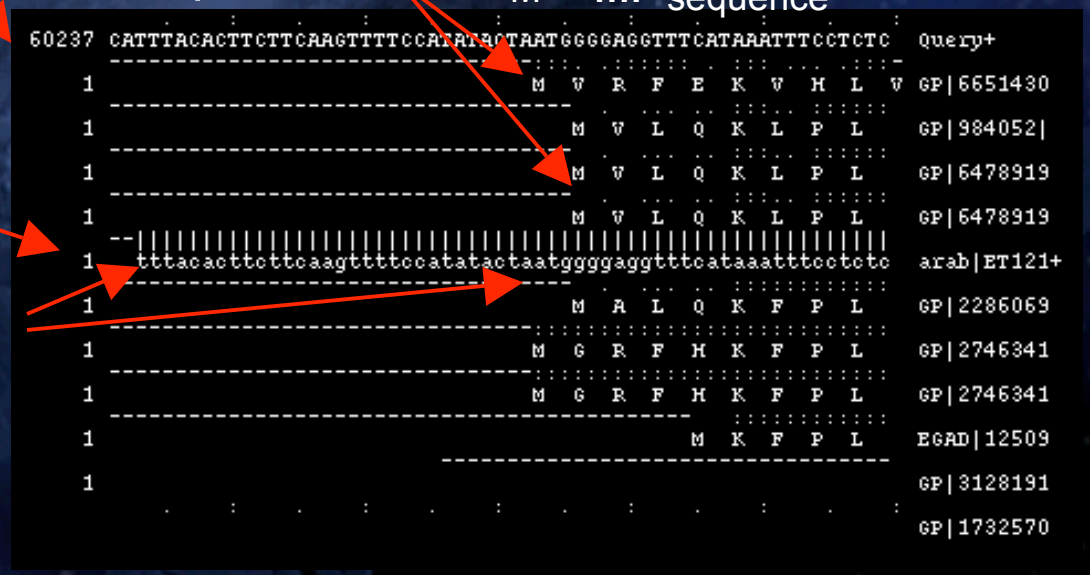
protein sequences

ATG....
: : :
M ....

Initiation codon defined by alignment between genomic and protein sequence

EST sequence

EST extends beyond → ATG 5' UTR



What are Full Length cDNAs?

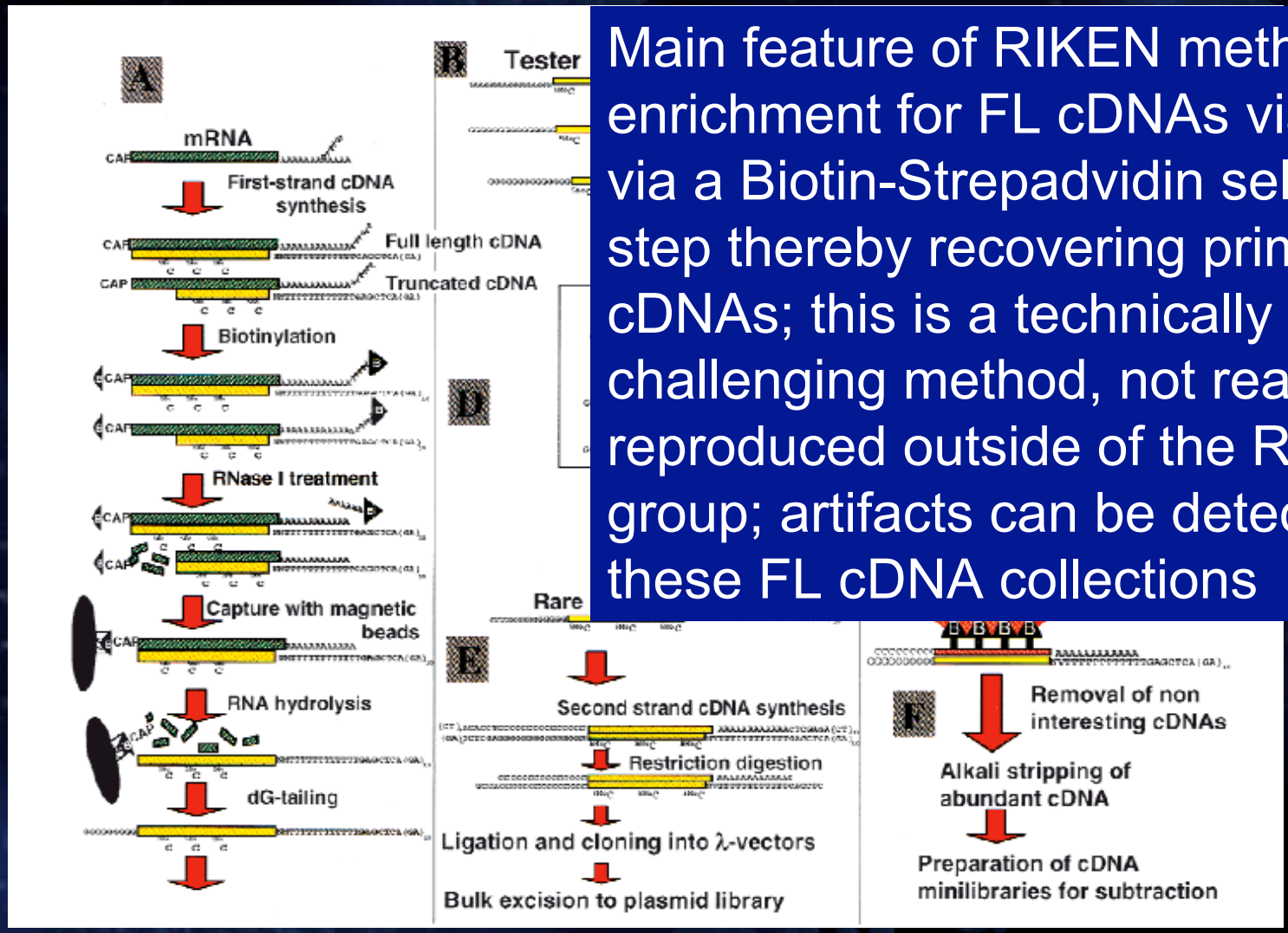
- Complete sequences of the transcript, include 5’ UTR and 3’ UTR
- More difficult to obtain, thus not as prevalent for a genome
- Fl-cDNA collections being developed for Arabidopsis and rice
- Clearly, MUCH more valuable in annotation than ESTs as the entire transcript is present

Uses of Full Length cDNAs

- Improve structural annotation
- Identify alternative splice variants
- Identify genes not present in annotation

Predominant FL cDNA Efforts

- Biotinylated CAP trapper method by the RIKEN group in Japan (mouse, Arabidopsis, rice).



Main feature of RIKEN method is enrichment for FL cDNAs via 5' CAP via a Biotin-Strepadvidin selection step thereby recovering primarily FL cDNAs; this is a technically challenging method, not readily reproduced outside of the RIKEN group; artifacts can be detected in these FL cDNA collections

- Carnici et al. Genome Research 10:1617-1630

Using FL cDNAs to Improve Annotation

Castelli et al. Genome Research 2004

- Used Arabidopsis FL cDNA data and rice genome sequence to update Arabidopsis genome annotation

- Sequenced 31,558 cDNA clones from normalized cDNA libraries

- Generated FL sequence for 21,572 clones

- Mapped to current Arabidopsis genome annotation

- Able to improve annotation (1,931), identify new genes (326)

- Also utilized comparative genomics with rice using "evolutionarily conserved regions (ecores)" to define genes

Using FL cDNAs to Improve Annotation

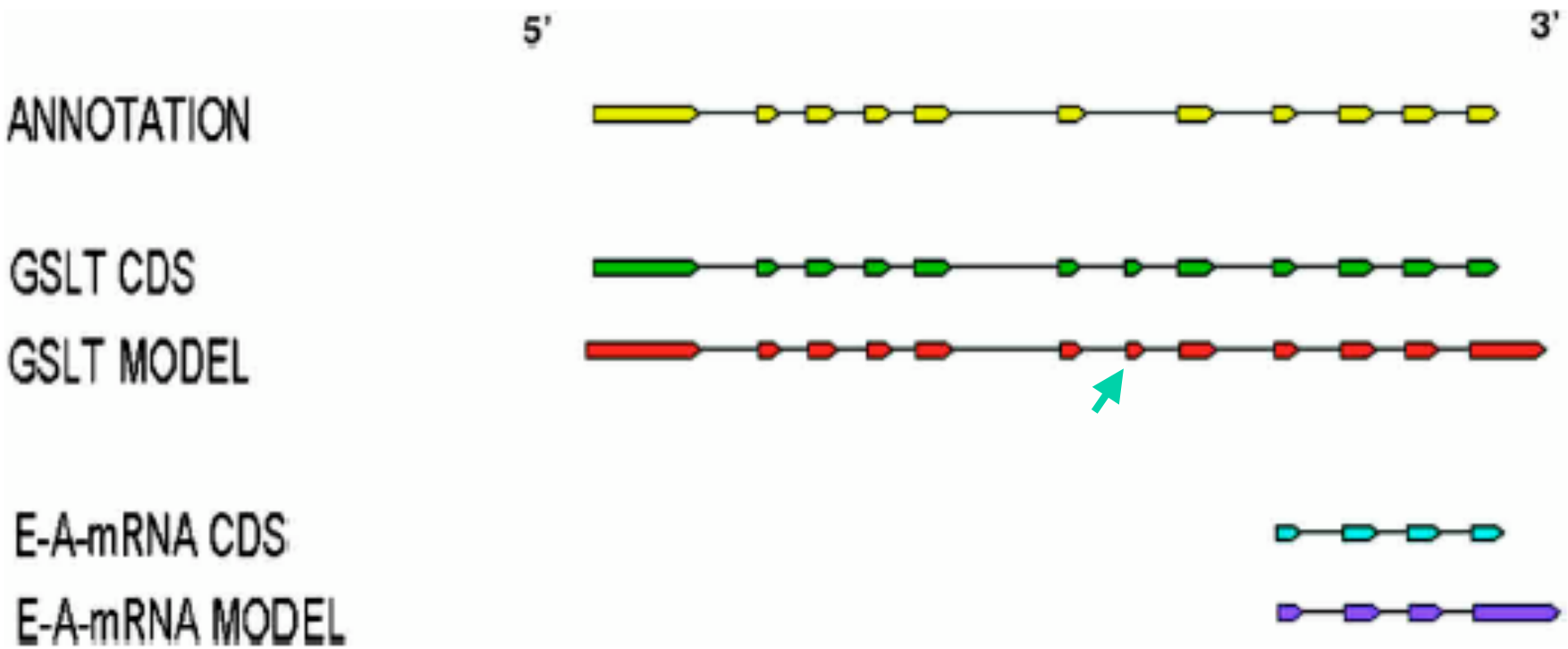


Figure 3 An example of 5′ extension detected by the GSTL resource. In this example, the gene structure of At3g58760 can also be corrected for a missing exon located between exons 6 and 7 of the annotated gene, due to longer cDNAs present in the GSLT resource.

Castelli et al. Genome Research 2004

Using FL cDNAs to Improve Annotation

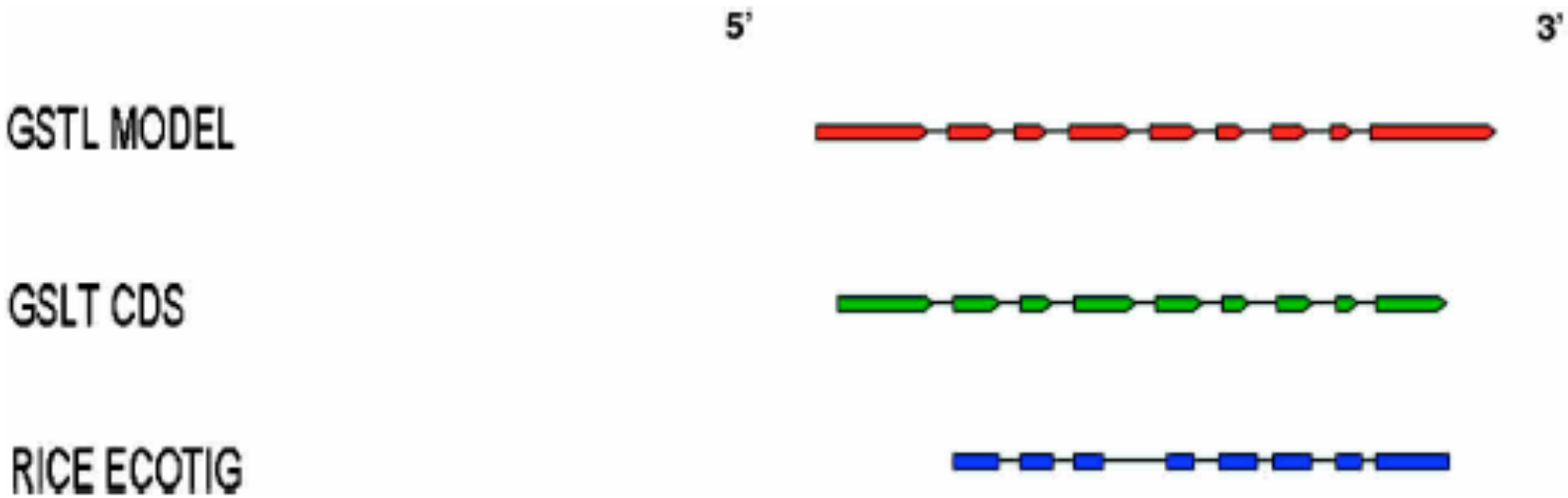


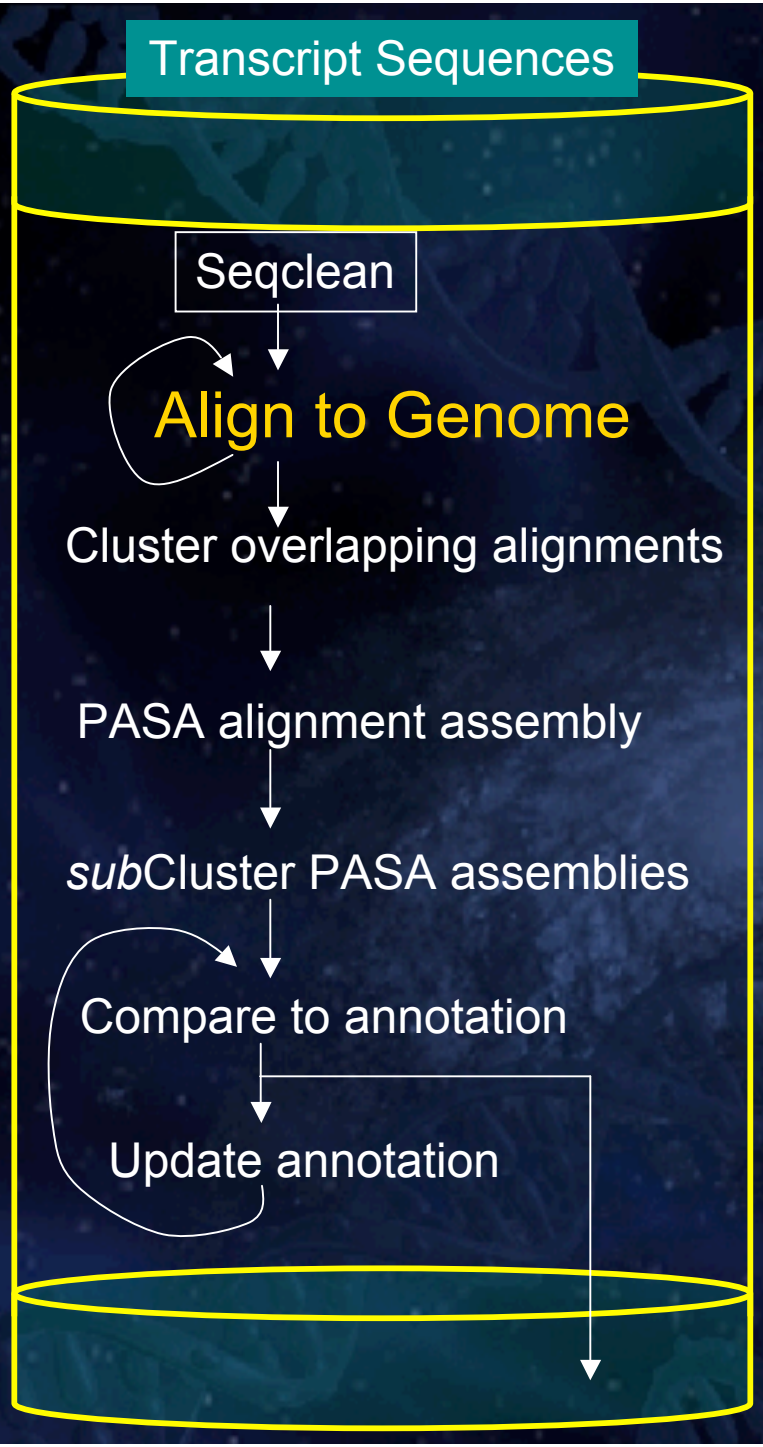
Figure 7 Novel gene detected by an ecotig and a GSLT cDNA sequence (GSLTF85ZE11, accession no. BX819512).

Castelli et al. Genome Research 2004

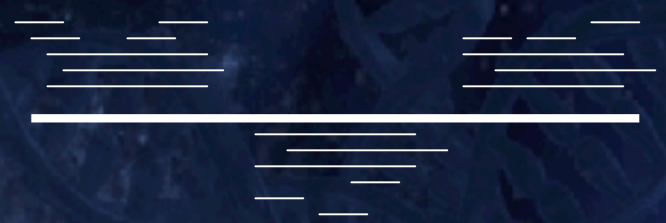
Rice FL cDNA Collections

- The Rice Full-Length cDNA Consortium, Science 2003
  - Generated sequences from 28,469 FL cDNA clones from a range of cDNA libraries of *Oryza sativa* spp *japonica* var Nipponbare (same variety as International Rice Genome Sequencing Project)
  - Available via GenBank and KOME = Knowledge-based Oryza Biological Encylcopedia (http://cdna01.dna.affrc.go.jp/cDNA)
  - Mapped 28,469 to rice genome to 3 versions of the rice genome sequence
    - Indica draft from BGI
    - Japonica Nipponbare draft from Syngenta
    - Public IRGSP BAC/PAC draft
  - Revealed between 15,523 and 19,036 nonredundant transcript units
  - Identified 5,045 transcription units with alternative structures
    - Initiation site
    - Internal exons
    - Termination site
    - Splice acceptor/donor site

PASA Pipeline



GMAP and sim4 spliced alignments



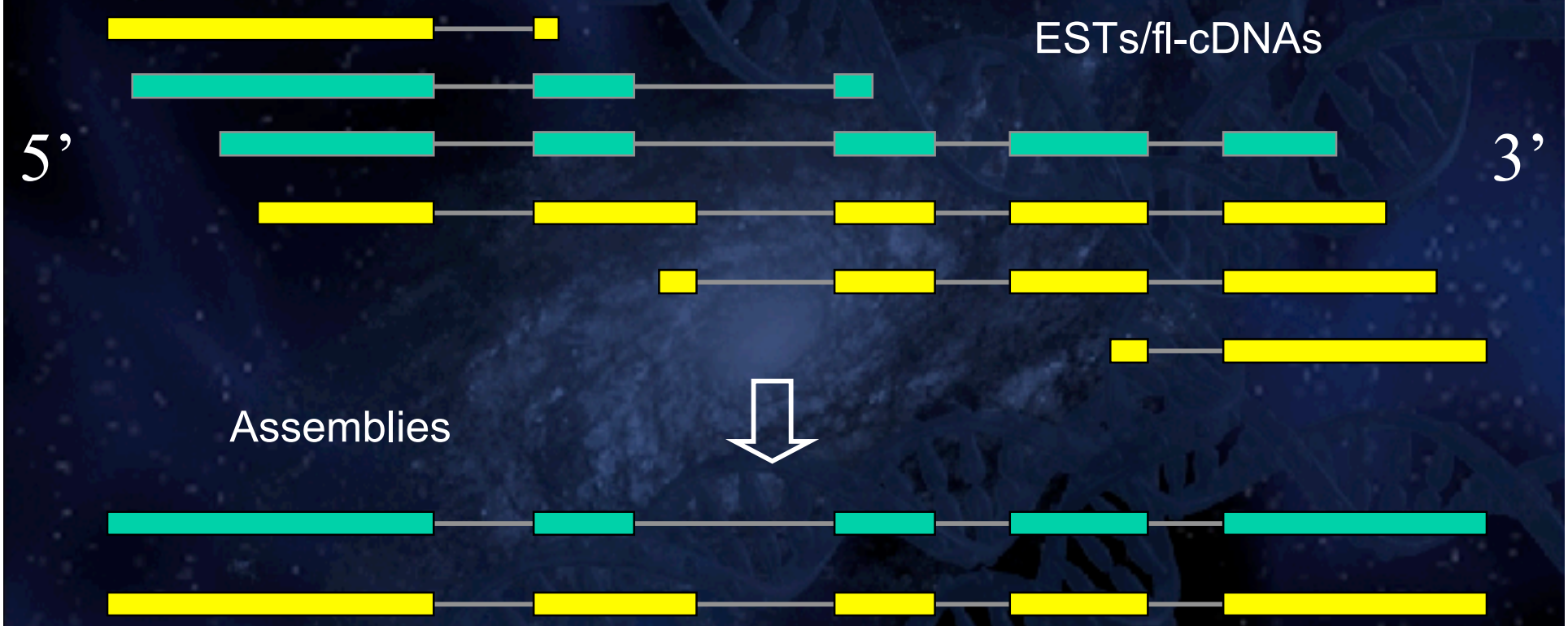
Valid alignment criteria:

- min *95%* Identity
  min *90%* transcript length aligned
  (both configurable parameters)
- consensus splice sites
  - (GT,GC) **donors**
  - AG acceptor

Assign Transcribed Orientations
- Splice sites
- Polyadenylation sites

B. Haas

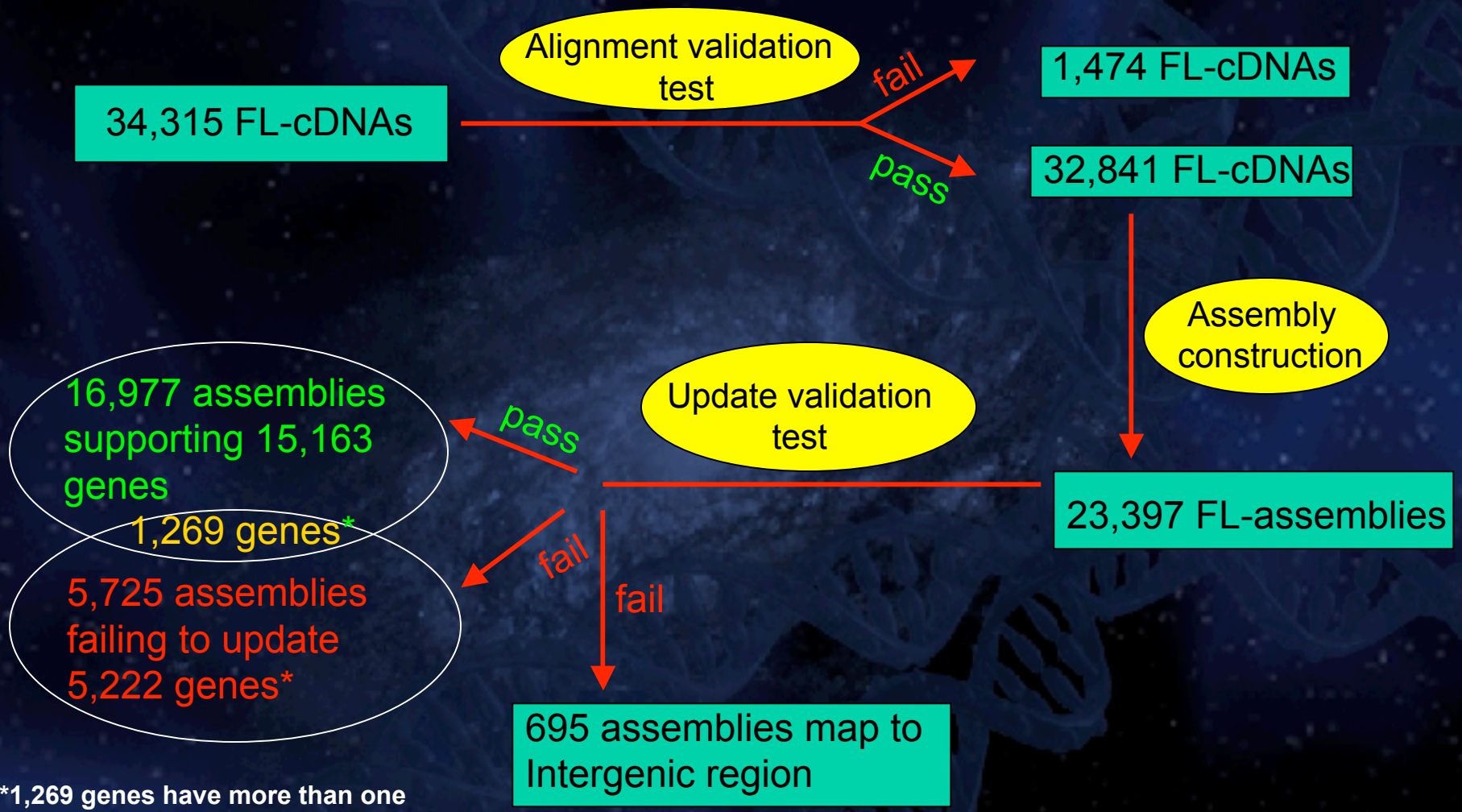
Alignment Assembly Using PASA: Program to Assemble Spliced Alignments



Maximally Assemble Compatible Alignments

B. Haas

Where do FL-cDNAs go in PASA?



***1,269 genes have more than one splice forms, with at least one splice form successfully updated and at least one failed to be updated**

* Based on TIGR rice version 3

A. Wang

SAGE: Serial Analysis of Gene Expression

- Serial analysis of gene expression (SAGE) is a method for comprehensive analysis of gene expression patterns.

- Developed in 1995
  - Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995. Serial analysis of gene expression. Science 270, 484-487.

  - Involves construction of cDNA, restriction with Type II restriction enzymes (cleave ~20 bp from their recognition site) yielding short “tags” which are concatenated together and sequenced

SAGE: Serial Analysis of Gene Expression

- Three principles underlie the SAGE methodology:

  - A short sequence tag (10 - 14 bp) contains sufficient information to uniquely identify a transcript provided that the tag is obtained from a unique position within each transcript

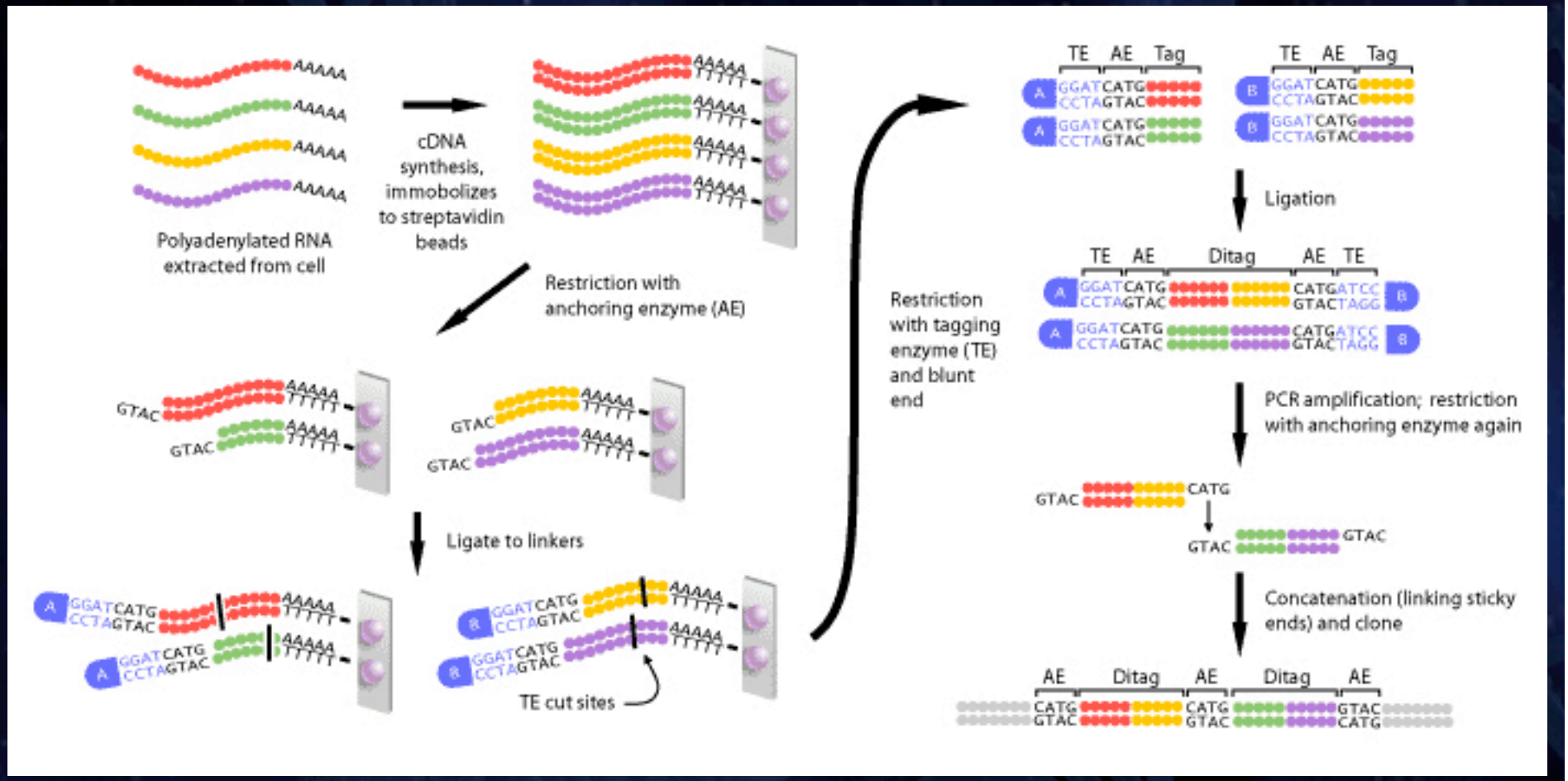
  - Sequence tags can be linked together to from long serial molecules that can be cloned and sequenced

  - Quantitation of the number of times that a particular tag is observed provides the expression level of the corresponding transcript.

- Improvements in SAGE technology has allowed for SAGE libraries to be constructed from small tissues samples (as few as 5000 cells (microSAGE)

http://www.sagenet.org/findings/index.html

SAGE: Serial Analysis of Gene Expression



Anchoring enzyme: usually *Nla*III; also use *Sau*3A and *Rsa*I
Tagging enzyme: typeIIS restriction enzymes, typically *Bsm*FI (14 bp tags); *Mme*I for long SAGE

http://bioteach.ubc.ca/MolecularBiology/PainlessGeneExpressionProfiling/index.htm

Downfalls of SAGE

- Tags are short, difficult to investigate
- Tags could be shared by multiple genes
- Type IIS restriction endonucleases could yield tags of various lengths
- Some genes don’t have the anchoring enzyme recognition sequence

SAGE: Serial Analysis of Gene Expression

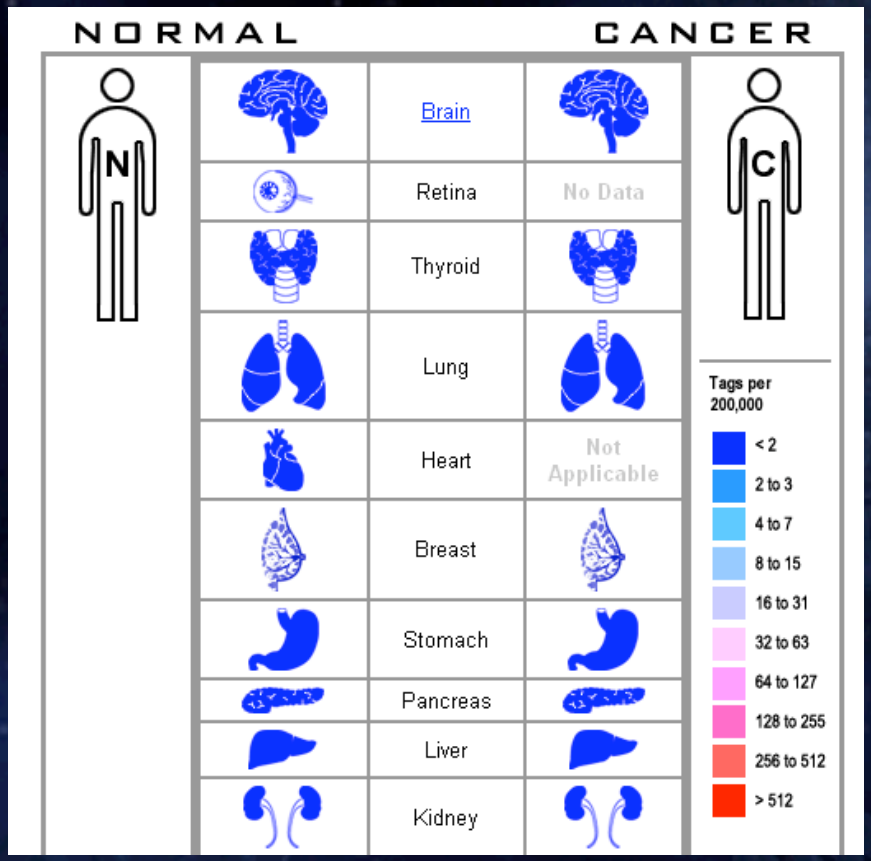
- Due to the ability to generate 100,000s of SAGE tags with little cost, SAGE data is HIGHLY quantitative

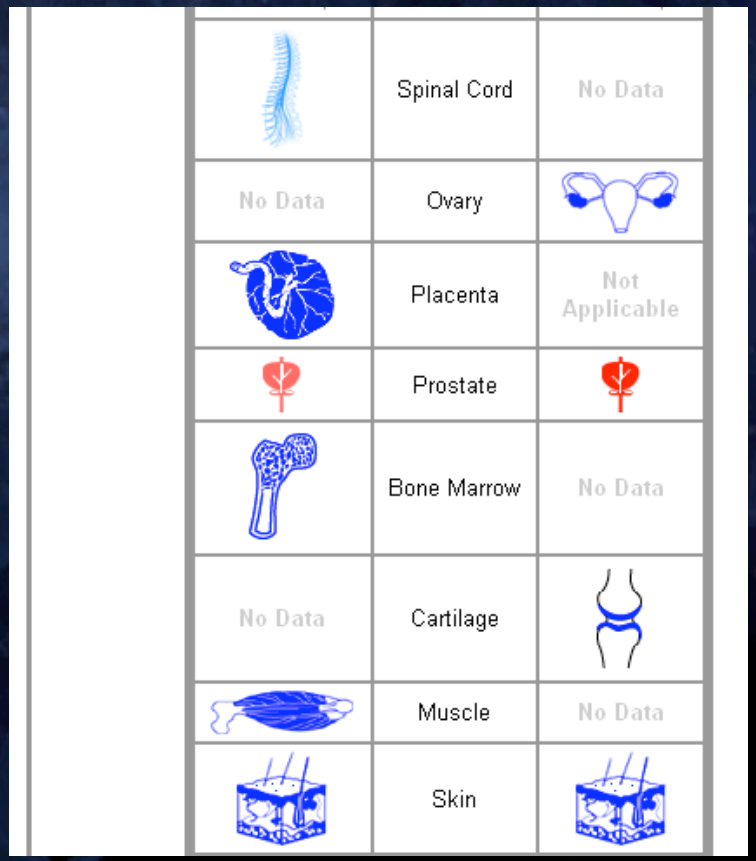
- National Cancer Institute has the Cancer Gene Anatomy project in which SAGE tags are used to assess gene frequency in cancer tissues

Search query: GGATGGGGAT, Tissues only

Colored organ image is hyperlinked to Digital Northern. "Brain" label is hyperlinked to expanded anatomic view of the brain.

SAGE GENIE
Query: PSA





SAGE: Serial Analysis of Gene Expression

Digital Northern Results

Search query: GGATGGGGAT , prostate , normal , Tissues only

Color Code										
Tags per 200,000	<2	<4	<8	<16	<32	<64	<128	<256	<512	>512

Library	Total Tags in Library	Tags per 200,000	Color Code
SAGE_Prostate_normal_B_2	64058	346	
SAGE_Prostate_normal_MD_PR317	59277	334	

Digital Northern Results

Search query: GGATGGGGAT , prostate , neoplasia , Tissues only

Color Code										
Tags per 200,000	<2	<4	<8	<16	<32	<64	<128	<256	<512	>512

Library	Total Tags in Library	Tags per 200,000	Color Code
SAGE_Prostate_adenocarcinoma_MD_PR317	64951	889	
SAGE_Prostate_carcinoma_B_LN-1	22599	407	
SAGE_Prostate_carcinoma_B_pool2	66034	302	

Uses of SAGE

- Evidence a gene is expressed

- Functional annotation of tissue and frequency at which gene is expressed

- Limited information on structural annotation of gene

- Evidence of genes yet to be predicted

Rice SAGE Data

Hideo Matsumura*, Shizuko Nirasawa and Ryohei Terauchi
Iwate Biotechnology Research Center, Narita, Kitakami, Iwate 024–0003, Japan

The Plant Journal (1999) 20(6), 719–726

TECHNICAL ADVANCE

Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE)

- Isolated SAGE tags from 5 day old etiolated rice seedlings
- Sequenced 650 plasmid clones
- Got 10,122 total tags
- Got 5,921 distinct tags
- 1,367 matched EST or cDNA

Table 1. Summary of SAGE analysis in rice seedlings

Total no. of tags studied[a]	No. of different tags (genes)	No. of tags matched with sequences in the database[b] (%)	No. of tags appearing more than once
10,122	5921	1367 (23.1)	5593

[a]Tag was extracted from the sequence data as 9 bp sequence adjacent to an *Nla*III site (CATG).
[b]Determined by searching previously known rice cDNA and EST databases with the 13 bp tag sequence.

Done in 1999 !!!

Rice SAGE Data: MGOS

SAGE project associated with rice blast disease

Four libraries (challenged with *Magnaporthe grisea*) were sequenced

A total of 152,816 tags (28,848 tags with multiple occurrence)

21 bp tags

Access tags and frequency through project website

www.mgosdb.org

MPSS Data and Others

- Massively Parallel Signature Sequencing

MPSS quantifies gene expression by simultaneously counting and identifying all mRNA species in a sample. Typically, in a single experiment, at least 1 million mRNAs are counted.

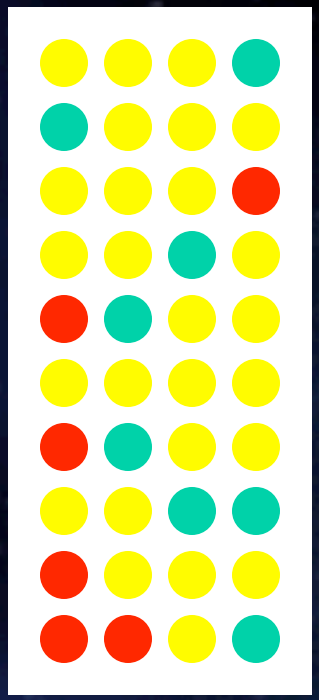
Individual mRNAs are identified through the generation of a 17- to 20-base signature sequence, immediately adjacent to the 3' end of the 3'-most *Sau*3A restriction site (GATC, also *Dpn*II) in cDNA sequences.

- 454's Massively Parallel Pyrosequencing Platform
- ABI's***S***upported ***O***ligonucleotide ***Li***gation and ***D***etection (***SOLiD***) Platform
- Solexa's ***S***equencing ***B***y ***S***ynthesis (***SBS***)Platform

Entire lecture/lab on MPSS and others in the afternoon by Dr. Kan Nobuta of Univ. of Delaware

What is a Microarray?

- Large number of probes at high density
  - Allows for the detection of expression of thousands of genes in a single assay



Spotted DNA on substrate = probes as you know the sequence/identity of these features

Labeled mRNA = targets as these are unknown

Why Use Microarrays?

- Allows for the analysis of expression of the whole genome in one time
- Identify co-regulation of different genes
- Identify genes that are involved in the process of interest
- Identify gene function
- An expression profile is also a phenotype
- Find the effect of a treatment or mutation on the complete transcriptome

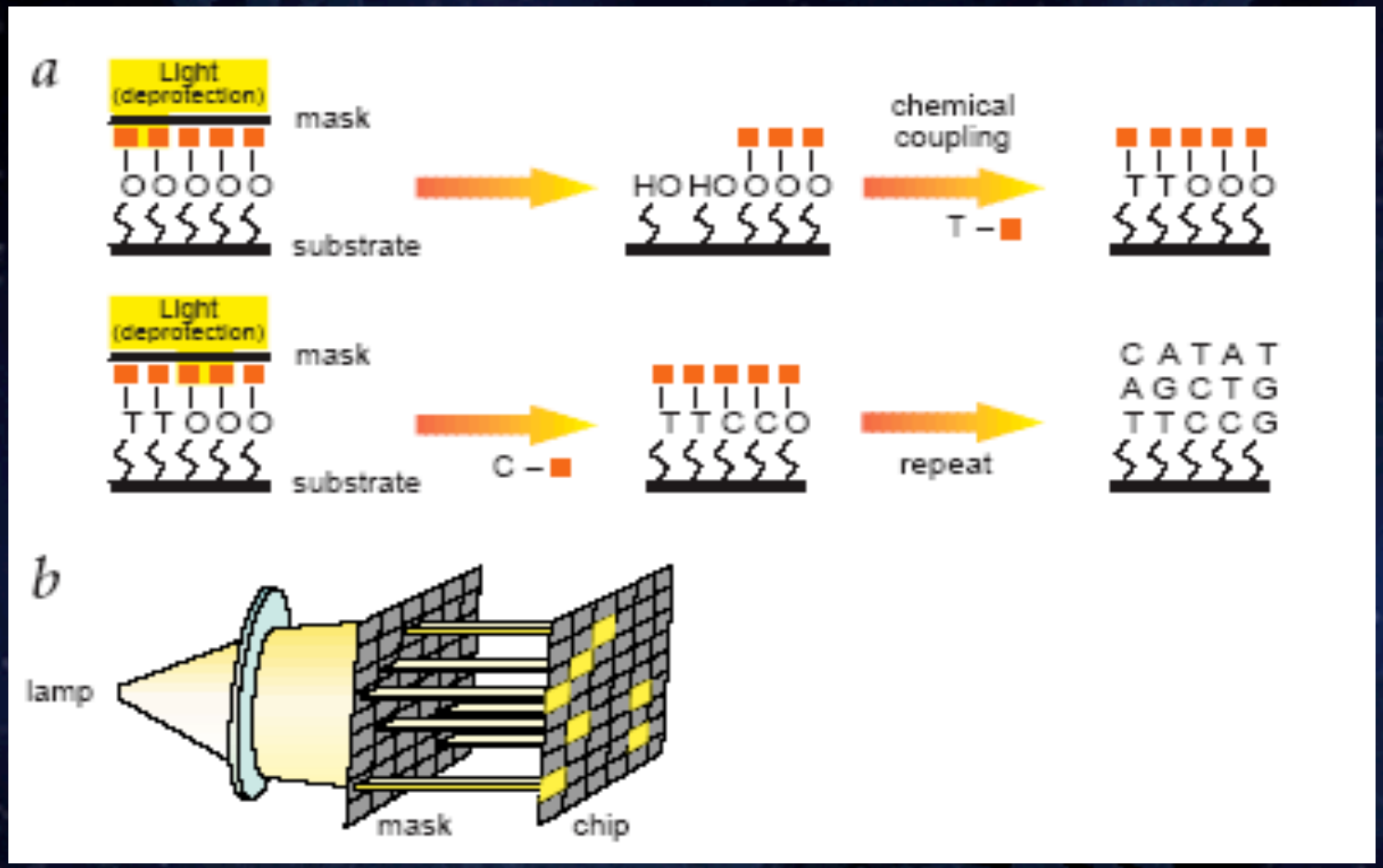
Microarray Types

- Different types of microarrays
  - Short oligo probes
    - Short stretch of synthesized DNA 25 bases
  - Long oligo probes
    - 50 – 70 bases of DNA
  - cDNA probes
    - PCR amplified DNA
- Single or two colored arrays

Production of Microarrays

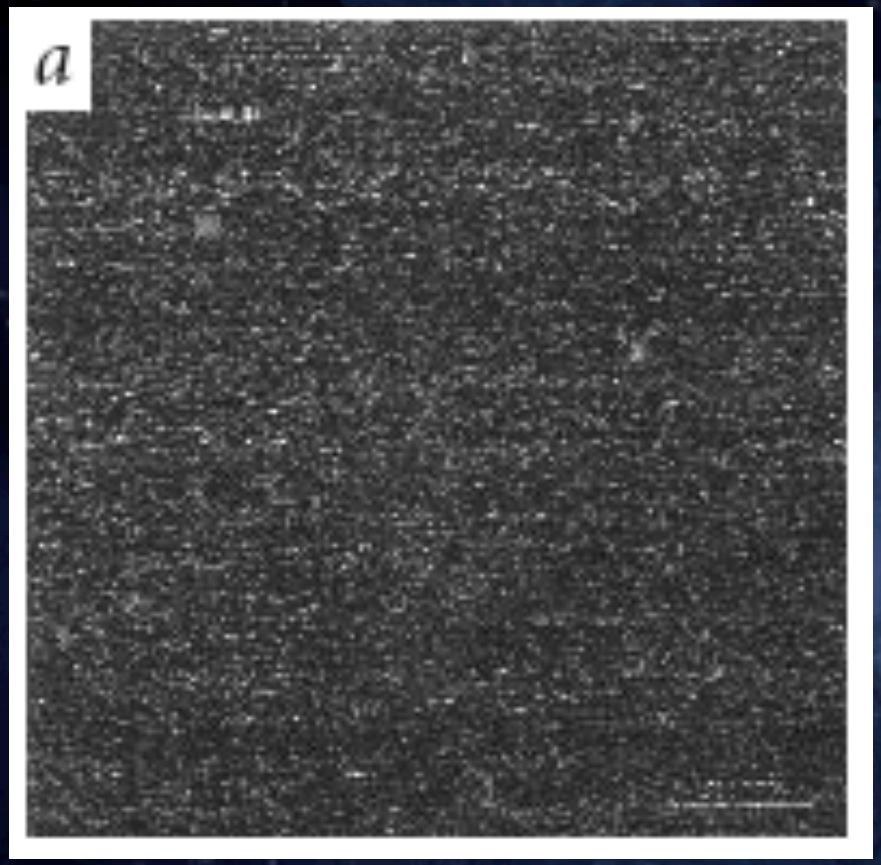
- On slide synthesis of probes (Nimblegen and Affymetrix)
  - Provides a high degree of specificity as the oligomer can be as short as 25 nt or as long as 70 nt
  - Cost per slide is high as requires large instrumentation only available at companies
  - Affymetrix and Nimblegen are typically only single channel hybridizations (half the amount of data)

Photolithography (Affy Technology)



Light directed synthesis of oligos on array. Light is directed through a mask to deprotect and activate select oligos. Chemical coupling then adds the requisite nucleotide. Process repeated to synthesize desired sequence and length of oligos

Affymetrix Technology



Gene Expression Array for Humans

- 1.28 x 1.28 cm array

- probe sets for 40,000 human genes and ESTs

- features are < 22 x 22 um

- 11-20 probe pairs per gene/EST

Lipshutz et al. Nature Genetics 1999

Affymetrix Technology

- Probes are chosen on unique DNA composition and thermodynamic design rules

- Probes enriched for 3' end of gene (more unique), labeled targets enriched for 3' sequences due to partially degraded mRNA

- Perfect Match (PM) and MisMatched (MM) probe pairs reduces the contribution of background

Lipshutz et al. Nature Genetics 1999

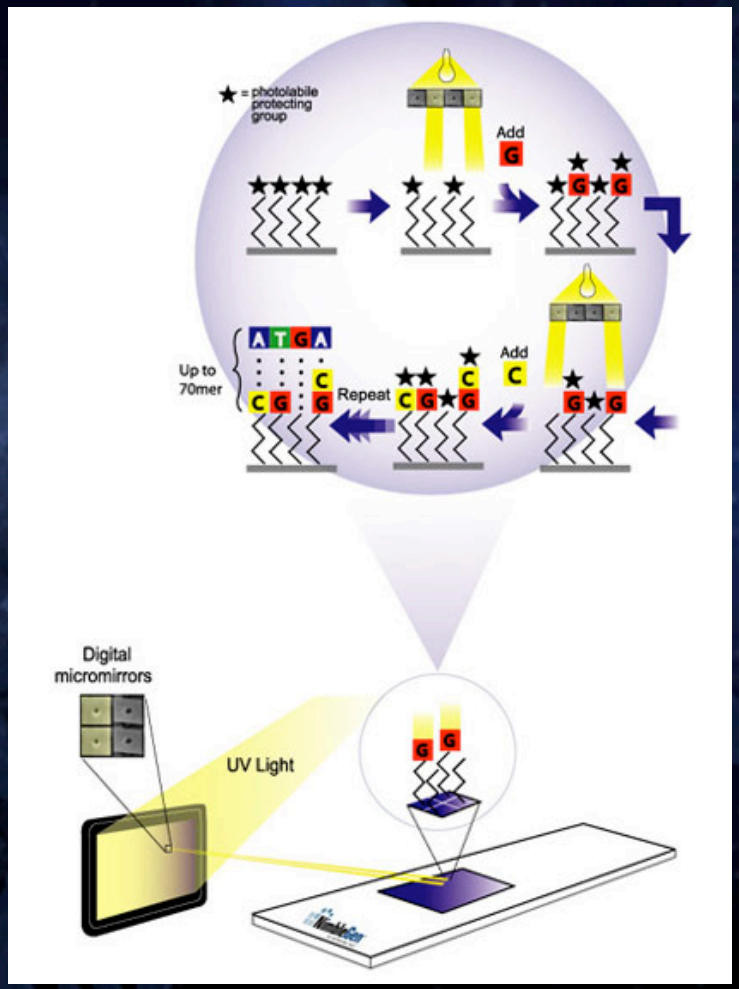
Nimblegen Technology
Maskless Array Synthesizer (MAS)

Similar to Affy yet uses glass slides and Digital Micromirror Device (DMD) instead of a mask

<u>Advantages:</u>
Cheap, flexible

<u>Disadvantages:</u>
1) Not readily accessible; hybridizations done in Iceland
2) Expensive

<u>Optimal use:</u>
1) Design of oligos for an array
2) Pilot experiments or production experiments if total need of arrays < 1000



www.nimblegen.com

Spotted cDNA Microarrays

1) Clone selection from cDNA collection

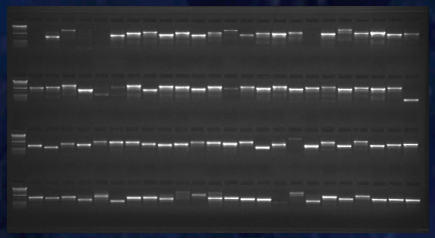
2) Resequencing (5' and 3')

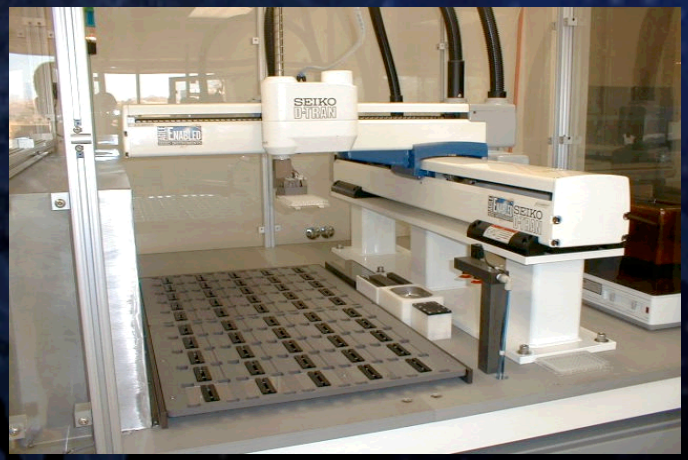
3) PCR amplification of insert

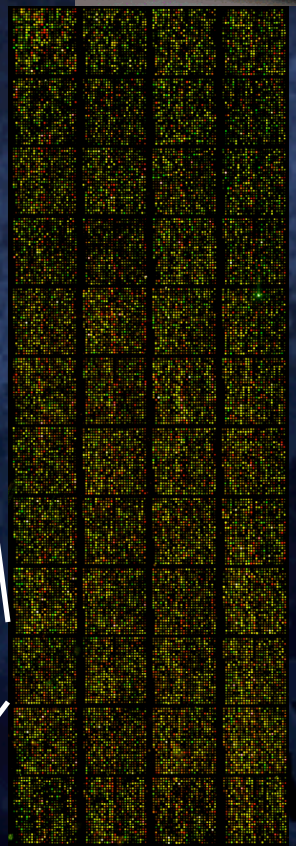
4) Gel verification

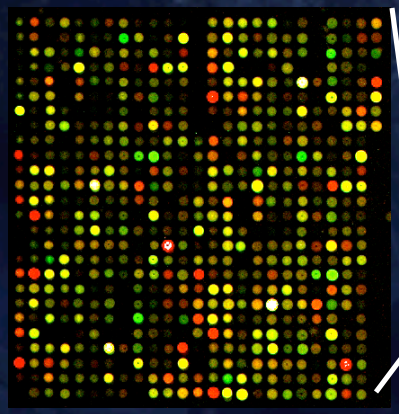
5) Arraying

6) Hybridization

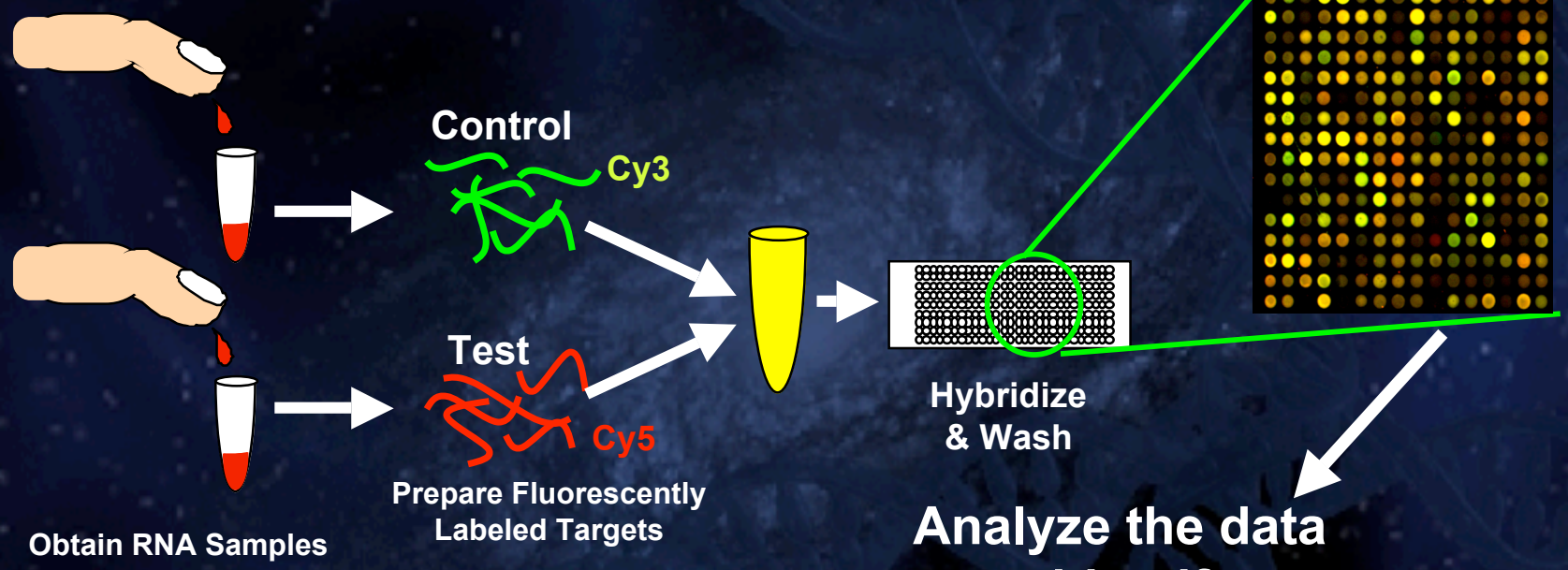








Microarray Assay



Measure Fluorescence
Two channels:
Red & Green

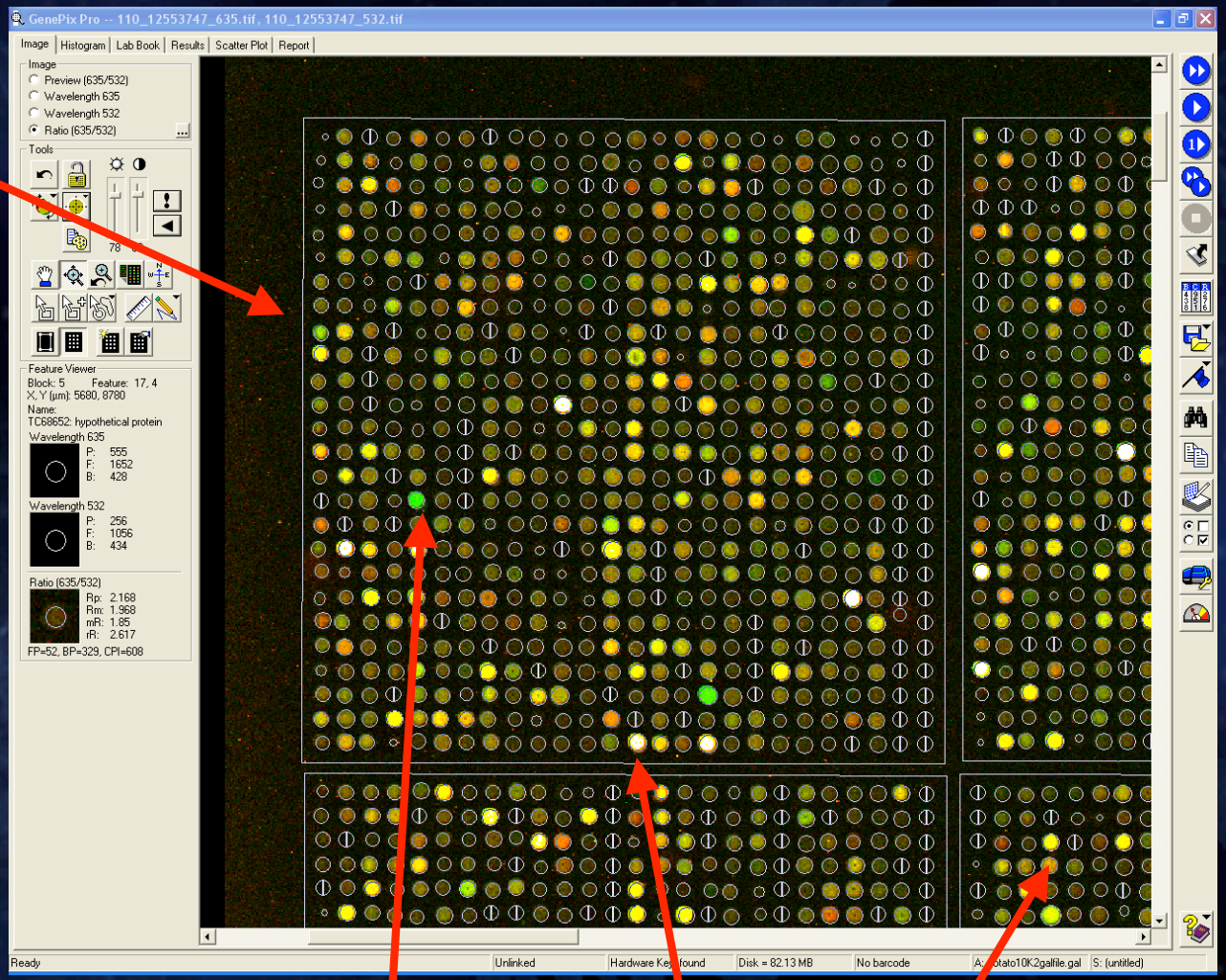
Analyze the data to identify differentially expressed genes

Image analysis example

One block of several (in this case 48 blocks on array)

Imaging program generates grid to quantitate spots

This is a Cy3 (green) vs Cy5 (red) scan, intensity value for each channel



Cy3 OverExpressed

Saturated Spot

Equally expressed

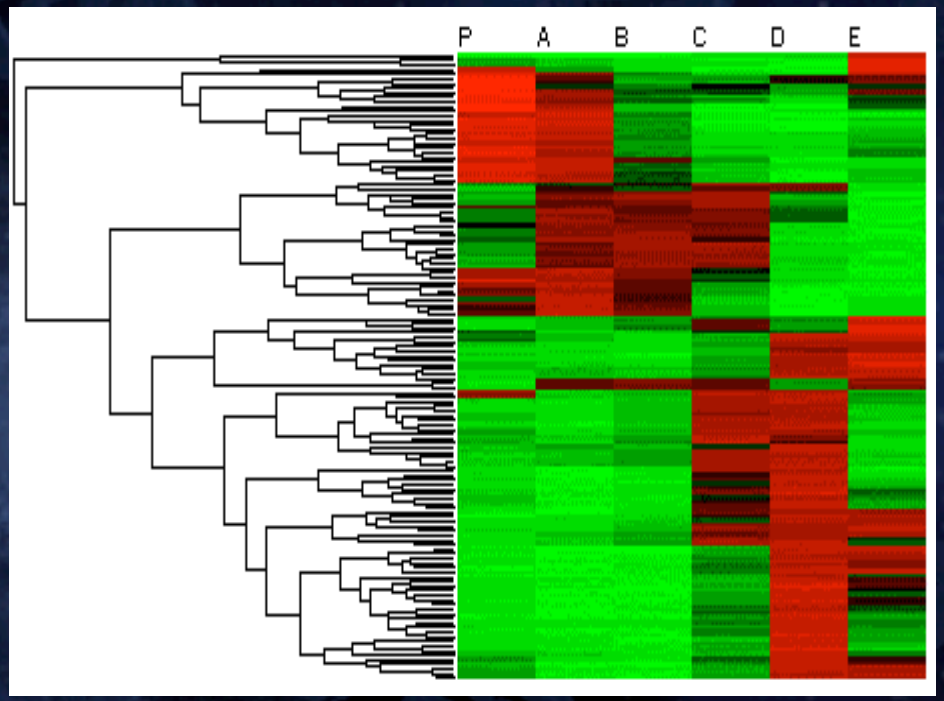
Genepix Microarray Scanner

Microarray data processing

- The result of a microarray experiment is an image (typically a TIFF file)
- On this image spots are identified and the pixel intensity is determined
- These intensities are the expression values for that spot
- Need to do normalization of intensities as there a large number of factors affecting the hybridization (labeling, washing, decay, detection, etc)

Microarray data processing

- Identify genes differentially expressed
- Perform global clustering of expression patterns
- Identify co-regulated genes
- Identify conserved regulatory regions of co-regulated genes



Hierarchical clustering

Integrating Array Data into Annotation

- Limited primarily to functional annotation
  - “salt stress regulated gene XXXX”
  - “Gene XXX, expressed in leaf primordia”
  - Future work will result in identification of conserved regulatory regions/motifs
  - More elaborate array uses such as genome tiling paths will yield data on transcriptomes such as exon/intron boundaries, novel transcripts, etc

Chromosome/Genome Tiling Arrays

- Short oligo arrays provide the opportunity to array the ENTIRE genome sequence on a slide
- Thus, one can assess the entire nucleic acid sequence of an organism for transcript potential
- Done with Affy technology (very expensive) and recently with Nimblegene technology (much less expensive)
- Isolate mRNA, label, hybridize, quantitate intensities, align with genomic sequence, compare with current annotation

Chromosome/Genome Tiling Arrays

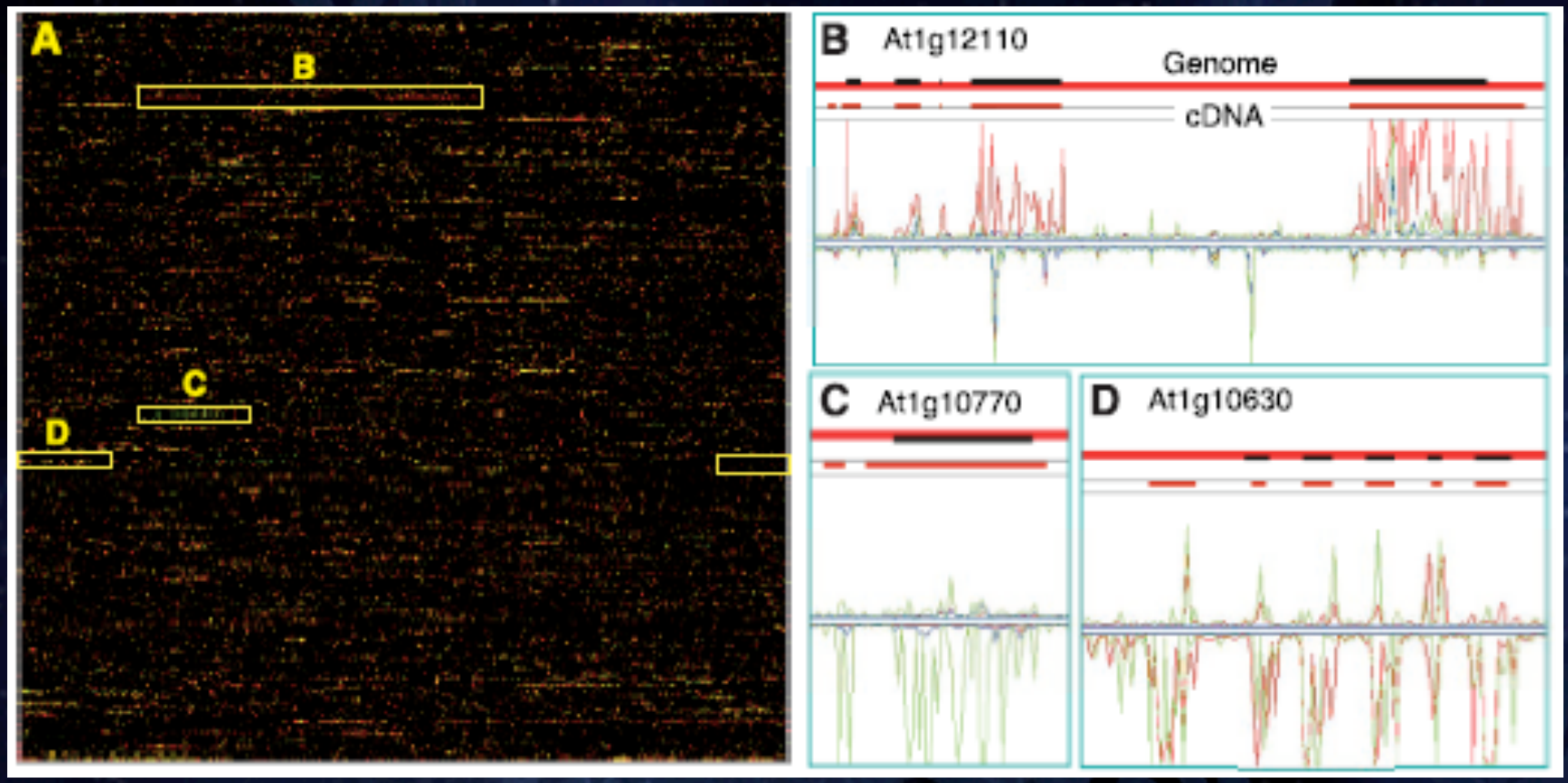
- Design involves tiling short oligos along the array in a fashion such that all, or almost all, of the genome/chromosome is represented. Sometimes the oligos overlap, sometimes they are abutted, sometimes there is a gap between them

- Overlapping Head to Tail Gap

Arabidopsis Genome Tiling Arrays

- Yamada et al. Science 2003
- Affy based arrays
- Used 4 mRNA populations to assess the transcriptome and ORFeome
- This manuscript reports on whole genome arrays and FL cDNAs
- Whole genome arrays: Arabidopsis genome is represented on 12 oligonucleotide arrays. Each array contains 834,000 25-mer oligonucleotides.
- Note this manuscript compares their data to the 2000 Arabidopsis annotation Version 1

Chromosome/Genome Tiling Arrays

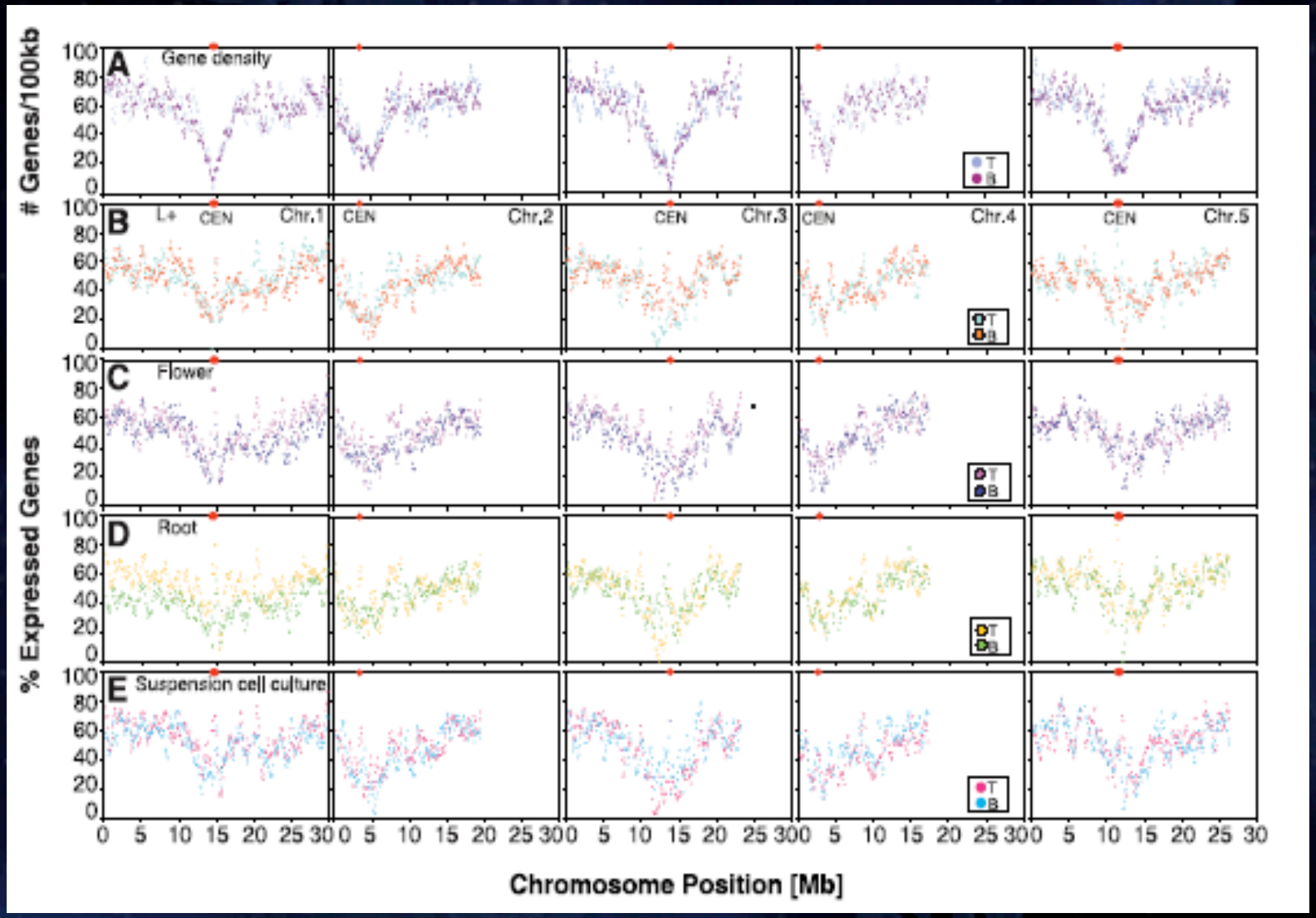


Array Image

Alignment of transcripts with annotation

Yamada et al. Science 2003

Chromosome/Genome Tiling Arrays



gene density

seedlings

flowers

roots

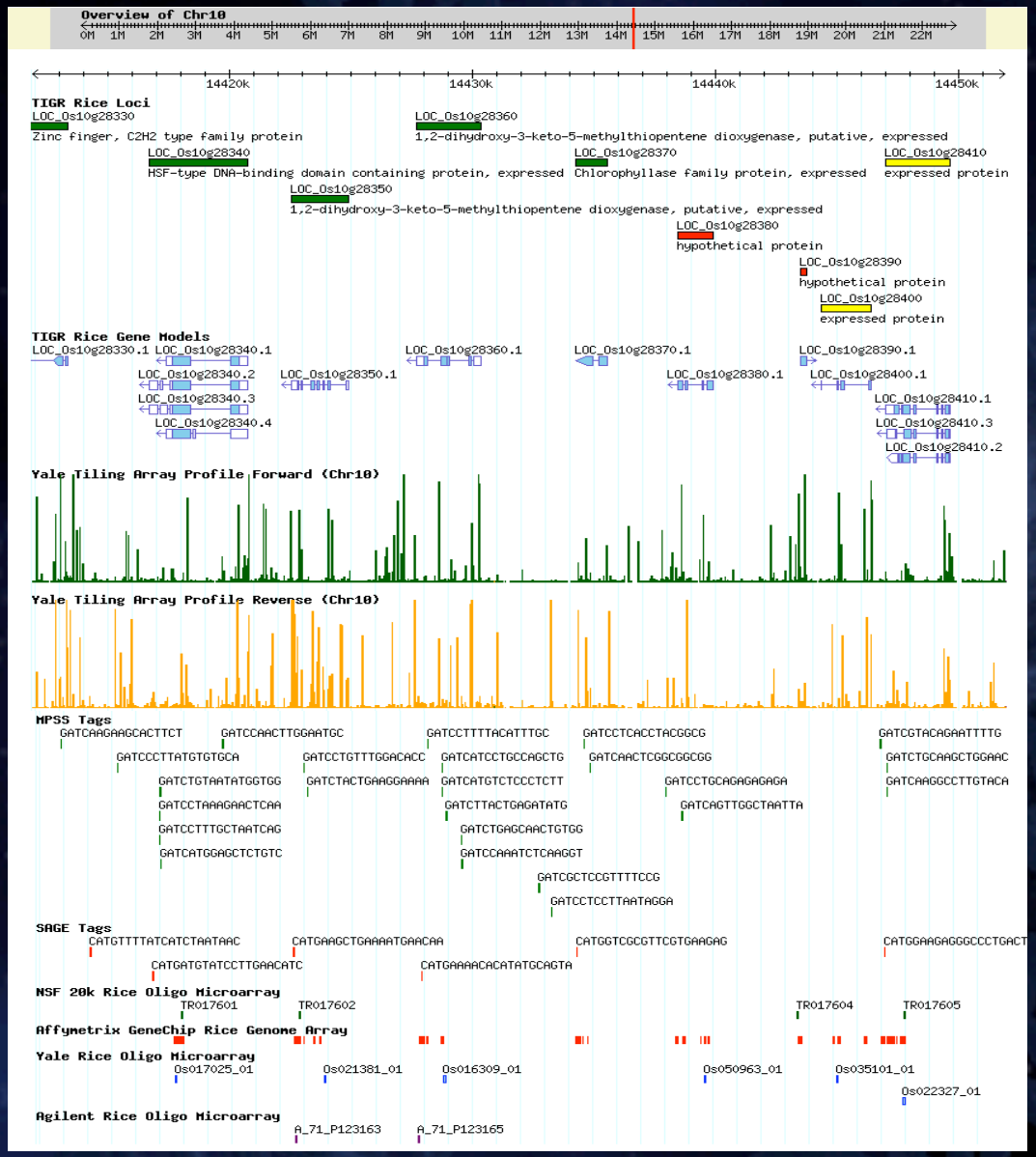
suspension cel culture

Gene density/transcript activity vs chr position
Yamada et al. Science 2003

Available Rice Microarray Platforms

Platform	Probe	Probe number	Designing base d on
NSF 20K Oligo Array	50-70mer	20,230	TIGR annotation, chloroplast, mitochontrion, japonica
NSF 45K Oligo Array	50-70mer	43,482	TIGR annotation, chloroplast, mitochontrion, japonica
Affymetrix GeneChip Rice G enome Array	25mer, ~11 oligos/set	631,066 oligos, 55,515 probe sets	UniGene, mRNAs, and TIGR annotation (v2), japonica/indica
Agilent Rice Oligo Microarray	60mer	21,495	28 K fl -cDNA, japonica
BGI/Yale 62K Oligo Array	70mer	58,258	BGI annotation, fl -cDNA, japonica/indica
Yale/NimbleGen Oligo Tiling Microarray	36mer (10 nt interval)	japonica and indica genome s	TIGR/BGI assemblies, japonica/indica

Rice Genome Annotation: Expression Data



Genome Browser alignment of array probes from 6 platforms:

NSF 20K Array
NSF 45k
Yale 60K Array
Affy Array
Agilent Array
Yale Chr 10 Tiling Array
MPSS and SAGE

http://rice.tigr.org/tigr-scripts/osa1_web/gbrowse/rice

NSF Rice Array Project

www.ricearray.org

2nd version of ~43 k oligos for ~45 k models

- PI: P. Ronald
- coPIs: R. Buell, P. Schnable, HH Chou, D. Rocke

Goals:

- Deliver a long oligo array to the public that represents the rice genome
- Provide a rice gene expression database to the public
- Link expression to rice genome annotation
- Public Hyb Service
- Array Training

Studies Available at NSF Rice Array Database

Number of results found: **7**

Click on a study's Title to view more information about the study. Click on a column header to sort results by that column.

	ID	Title	Investigator	Hybs	Platform	FTP Download	PDF Description	Summary Files
1.	4	Analysis of rice cellular expression	Nelson, Tim	72	Yale 60k 1A Yale 60k 1B [Link]			A B
2.	10	Temperature testing in Light vs Dark condition	Pamela Ronald	12	NSF 20k (Davis)		–	
3.	14	Rice Transcriptome	Xingwang Deng	30	Yale 60k 1A Yale 60k 1B [Link]		–	–
4.	17	Abiotic stress	Ju-Kon Kim	12	Operon BGI 60k		–	–
5.	18	Expression data from rice under salinity stress	Timothy Close	24	Affymetrix		–	–
6.	20	Transcriptomic adaptations in rice suspension cells under sucrose starvation	Huei-Jing Wang	12	Agilent 22K		–	–
7.	21	Bacterial lipopolysaccharides induce defense responses associated with Programmed cell death in rice cell	Hanae Kaku	4	Agilent 22K		–	–

7 studies, 166 hybs, from 4 different platforms

Multi-platform Rice Microarray Search Tool

http://www.ricearray.org/matrix.search.shtml
http://rice.tigr.org/ricearray/matrix.search.shtml

Rice Multi-platform Microarrary Search

The Rice Multi-platform Microarrary Search page allows you to perform a singleton/batch search of rice oligo microarray probes from multiple platforms such as NSF 20K Rice Oligo Microarray, Affymetrix GeneChip Rice Genome Array, Yale Rice Oligo Microarray and Agilent Rice Oligo Microarray. These probes have been mapped to the TIGR Rice Genome Annotation gene models (release 3), KOME full-length cDNA, or the TIGR Rice Gene Index release 15 .

You can either upload a file containing a list of accessions/oligo names (e.g LOC_Os03g52660.1" for a rice model, "gi|32970393"/"AK060375.1" for full length cDNA, "TC253764"/"NP919509" for Rice Gene Index, "TR006054" for NSF 20K oligo array, "probe:Rice:Os.2405.1.S1_at:997:261" for Affymetrix GeneChip Rice Genome Array, "Os014986_01" for Yale Rice Oligo Microarray and "A_71_P119956" for Agilent Rice Oligo Microarray) in plain text format, or paste a list of identifiers in the text box directly.

After making your selection, click the 'Submit' button. If you have selected both a file upload and pasted a list of search terms, only the file will be used.

The entire search matrix tables are available for download:

Rice Multi-platform Table (zip, 11M)
Rice Multi-platform Table (with coordinates) (zip, 21M)

Enter the name of the file containing a list of identifiers

Browse...

OR

Enter a list of identifiers terms in the text area below

Return results as HTML Plain text

Submit Clear

Multi-platform Rice Microarray Search Tool

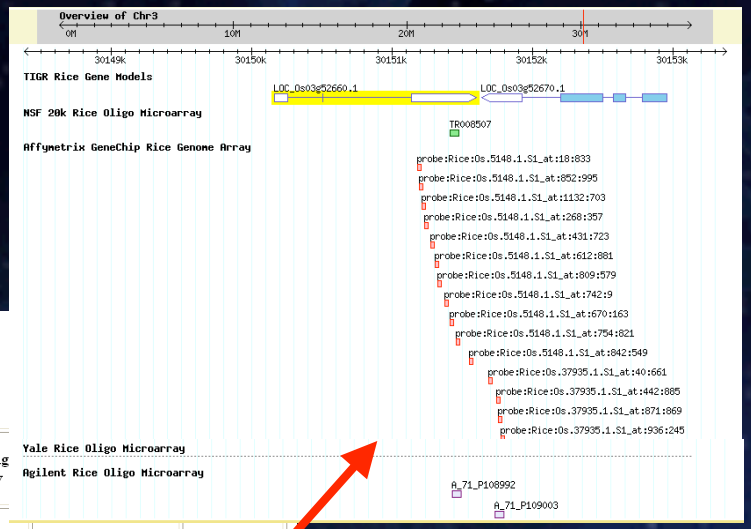
Multi-platform Rice Microarray Search Results

Number of matches found: 1

Accession	NSF 20K Rice Oligo Microarray	Affymetrix GeneChip Rice Genome Array	Yale Rice Oligo Microarray	Agilent Rice Oligo Microarray	Oligo Mapping View
LOC_Os03g52660.1	TR008507	probe:Rice:Os.5148.1.S1_at:18:833; probe:Rice:Os.5148.1.S1_at:852:995; probe:Rice:Os.5148.1.S1_at:1132:703; probe:Rice:Os.5148.1.S1_at:268:357; probe:Rice:Os.5148.1.S1_at:431:723; probe:Rice:Os.5148.1.S1_at:612:881; probe:Rice:Os.5148.1.S1_at:809:579; probe:Rice:Os.5148.1.S1_at:742:9; probe:Rice:Os.5148.1.S1_at:670:163; probe:Rice:Os.5148.1.S1_at:754:821; probe:Rice:Os.5148.1.S1_at:842:549;	n/a	A_71_P108992	View

Integrating Array Data and Annotation

Link to TIGR Osa1 Annotation Manatee Page for Locus



Multi-platform Rice Microarray Search Results

Number of matches found: 1

Accession	NSF 20K Rice Oligo Microarray	Affymetrix GeneChip Rice Genome Array	Yale Rice Olig Microarray		
LOC_Os03g52660.1	TR008507	probe:Rice:Os.5148.1.S1_at:18:833; probe:Rice:Os.5148.1.S1_at:852:995; probe:Rice:Os.5148.1.S1_at:1132:703; probe:Rice:Os.5148.1.S1_at:268:357; probe:Rice:Os.5148.1.S1_at:431:723; probe:Rice:Os.5148.1.S1_at:612:881; probe:Rice:Os.5148.1.S1_at:809:579; probe:Rice:Os.5148.1.S1_at:742:9; probe:Rice:Os.5148.1.S1_at:670:163; probe:Rice:Os.5148.1.S1_at:754:821; probe:Rice:Os.5148.1.S1_at:842:549;	n/a	A_71_P108992	View

Link to TIGR Genome Browser with probe mappings

Oryza sativa spp japonica cv. Nipponbar LOC_Os03g52660 Annotated by TIGR

Download sequence · Show genomic region on LOC_Os03g52660

Gene Identification

Gene Product Name: **expressed protein**
Locus Name:
Alternative Splice Form: LOC_Os03g52660.1
Comment: EST AU173863,AU173864 from this gene

Gene Ontology Classification |

None found

Attributes

Chromosome: 3
Coordinates (5' - 3'): 30150165 - 30151597 on assembly LOC_Os03g52660
Nucleotide length: 630
Predicted protein length: 210
Predicted molecular weight: 0.00
Predicted pI: 0.00

Links to array projects

Query Microarray Data

First | < Prev | | Page Jump | | Next > | End

TR002000 - Showing results **1 - 12** - Page **1** of **1**

	Study ID	Study Name	Study Investigator	Hyb ID	Hyb Name	Hyb Platform	Oligo ID	Fold Change (log2)	Normalized Query Intensity	Normalized Reference Intensity	Export [Check All] [Clear All]
1.	10	Temperature testing in Light vs Dark condition	Pamela Ronald	172	Leaf_2weeks_42oC_2510	NSF 20k (UCDavis)	TR002000 LOC_Os01g44410.1	-0.286	274	333	
		Protein kinase, putative							Source_tissue_shoot_primeshoot Lightleaf_4	Source_tissue_shoot_primeshoot Darkleaf_4	
2.	10	Temperature testing in Light vs Dark condition	Pamela Ronald	175	Leaf_2weeks_46oC_2495	NSF 20k (UCDavis)	TR002000 LOC_Os01g44410.1	-0.415	173	232	
		Protein kinase, putative							Source_tissue_shoot_primeshoot Lightleaf_7	Source_tissue_shoot_primeshoot Darkleaf_7	
3.	10	Temperature testing in Light vs Dark condition	Pamela Ronald	169	Leaf_2weeks_42oC_2454	NSF 20k (UCDavis)	TR002000 LOC_Os01g44410.1	N/A	0	0	
		Protein kinase, putative							Source_tissue_shoot_primeshoot Darkleaf_1	Source_tissue_shoot_primeshoot Lightleaf_1	
4.	10	Temperature testing in Light vs Dark condition	Pamela Ronald	170	Leaf_2weeks_42oC_2491	NSF 20k (UCDavis)	TR002000 LOC_Os01g44410.1	N/A	0	0	
		Protein kinase, putative							Source_tissue_shoot_primeshoot Darkleaf_2	Source_tissue_shoot_primeshoot Lightleaf_2	

Mining Rice EST and Array Data

Oligo and EST Anatomy Viewer

What is OEAV?

The Oligo and EST Anatomy Viewer (OEAV) is a tool we have developed at TIGR to visualize transcriptome data for rice. This tool provides quantitative data on rice transcript frequency based on ESTs (digital or electronic Northern) and microarray data (in development). This is modeled after the NCI SAGE Genie.

The purpose of this tool is to provide array users with a quick reference for comparison of array expression data with digital northern data as determined through EST frequency. However, users should be cautious about the data as there are limitations on its use. For example, some of the EST libraries were normalized or sequenced in a small number, thus, the data generated for these libraries may not be robust.

The data in OEAV can be searched based on keyword, oligo ID, accession number, rice model, or gene ontology. The gene ontology (GO) assignments are provide the deeper level as shown on the GO site and with plant GO Slim identifiers th provide a higher view of the ontologies. Please visit the gene ontology web site more on the GO identifiers

Use the Oligo and EST Anatomy Viewer

- **Find the oligo and its matching sequence**

Search item ◉ containing ○ matching

by Keyword (e.g. carboxylase)

Oligo and EST Anatomy Viewer

EST based (currently)
Also available for maize and wheat

Oligo and EST Anatomic Viewer Search Results

Search query: aquaporin

Number of matches found: 1

Below are the oligonucleotides that we have on the rice array that are associated with your search term. The TIGR oligo id, a unique identifier for each oligonucleotide, is hyperlinked to further annotation for this oligonucleotide. The accessions are the sequences in the TIGR Rice model, TIGR rice Gene Index, full length cDNA that match this oligonucleotide at 100 % identity over 100 % length. The putative annotation is the gene name assignment for this sequence. Two outputs are provided for the EST Based Digital Northern: the Frequency via Library and the Anatomy Viewer. The Frequency via Library is a tabular format for the data while the anatomy viewer provides graphical views of expression based on tissues.

TIGR Oligo ID	Accession	Putative Annotation	EST Based Digital Northern	
			Frequency via Library	Anatomy Viewer
TR006304	gi|435648|dbj|D25534.1|RICYK33 TC261902 LOC_Os03g05290.1	tonoplast intrinsic protein, gamma (gamma tip) (aquaporin-tip)	TC261902	TC261902

Search by:
Keyword
Oligo ID
Accession
Plant GOSlim

http://www.ricearray.org/rice_digital_northern_search.shtml

Mining Rice EST and Array Data

Digital Northern by Library Results

Color Bar Legend:

Color Code								
No. EST	1	2	3-10	11-20	21-40	41-100	101-500	>500

Query Sequence: TC261902

Putative Annotation: tonoplast intrinsic protein, gamma (gamma tip) (aquaporin-tip)

Associated Oligo ID: TR006304

Total 154 ESTs Found in 31 Libraries

[Sort the result by Library Cat#, No. EST found and Frequency]

Library Cat#	Tissue	Color Code	Total EST in Library	No. EST Found	Frequency (%)
#9DK	Leaf		5615	4	0.071
#9IU	Endosperm		9990	7	0.07
#9IV	Stem		2830	2	0.071
#9J0	Leaf		15139	13	0.086
#BMP	Pistil		852	1	0.117

Digital Northern

EST Anatomy Viewer

Color Bar Legend:

Color Code									
No. EST	0	1	2	3-10	11-20	21-40	41-100	101-500	>500

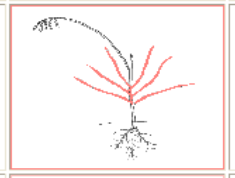
Query Sequence: TC261902

Putative Annotation: tonoplast intrinsic protein, gamma (gamma tip) (aquaporin-tip)

Associated Oligo ID: TR006304

Total ESTs Found for This Query: 154

[Sort the result by Tissue, No. EST found and Frequency]

Tissue	Plant Ontology ID	Anatomy View	Total EST from Tissue	No. ESTs Found	Frequency (%)
Leaf	PO:0009025		144677	56	0.039

Anatomy Viewer

Mining Rice EST and Array Data

Highly Expressed Gene Finder

What does this tool do?

Using EST frequency data, this tool allows users to quickly identify a set of genes that are highly expressed in certain tissue. The output is a list of oligonucleotides, their corresponding TIGR rice Gene Index accession number, a putative annotation for that sequence if stored in our database, and frequency of the sequence in the EST database.

Note: If the annotation returned as "N/A", it means that either the EST sequence did not match any known genes in the database, or the sequence was not chosen as the representative sequence for an oligo therefore its annotation was not stored in our database. Please click the TIGR oligo ID and accession number for detailed annotation for the oligo and EST sequence, respectively.

Search for:

top highly expres

in tissue Anthers

Submit Clear

Anthers
Callus
Endosperm
Flower
Immature seed
Leaf
Panicle
Phloem
Pistil
Root tip
Seed
Seedling
Sheath
Shoot
Stem
Suspension cells
Whole plant
Mixed
Unknown

http://www.tigr.org/tdb/e2k1/osa1/tissue.expression.shtml

Tissue-specific Expression

EST based (currently)
Also available for maize and wheat

Highly Expressed Rice Genes in Flower

Tissue: Flower

Total ESTs in Flower: 57600

Number of loci in display: top 5

Locus	Putative Annotation	Oligo ID	No. ESTs Found	Frequency (%)	Anatomy Viewer	Digital Northern
LOC_Os04g56160.1	plasma membrane ATPase, putative, expressed	A_71_P112005 Os.12328.1.S1_at Os018983_01 TR047214	143	0.248	view	view
LOC_Os04g56160.2	plasma membrane ATPase, putative, expressed	TR047214	102	0.177	view	view
LOC_Os04g55110.1	expressed protein	A_71_P110982 Os.53825.1.S1_at Os018867_01 TR011207 TR047137	102	0.177	view	view
LOC_Os07g09340.1	plasma membrane ATPase 1, putative, expressed	A_71_P118987 Os.5684.1.S1_at Os054790_01 TR013110 TR048114	101	0.175	view	view
LOC_Os04g55650.1	cysteine proteinase RD21a precursor, putative, expressed	Os.12701.1.S1_at Os.12701.1.S2_at Os054586_01 TR047177	90	0.156	view	view

Genome-wide Expression Methods

Method	Comprehensive	Quantitative	Sensitive	Novel gene finding
EST	no	yes	no	yes
SAGE	yes	yes	yes	yes
MPSS	yes	yes	yes	yes
Microarray	yes	yes/no	no	no

Utilizing Rice Proteomic Data

- Collect peptide sequences

  Koller *et al*., PNAS, 2002:

  2D-PAGE & MudPIT: 6,296 peptides/2,528 proteins (Syngenta fgenesh models)

  Komatsu *et al*., NAR 2004:

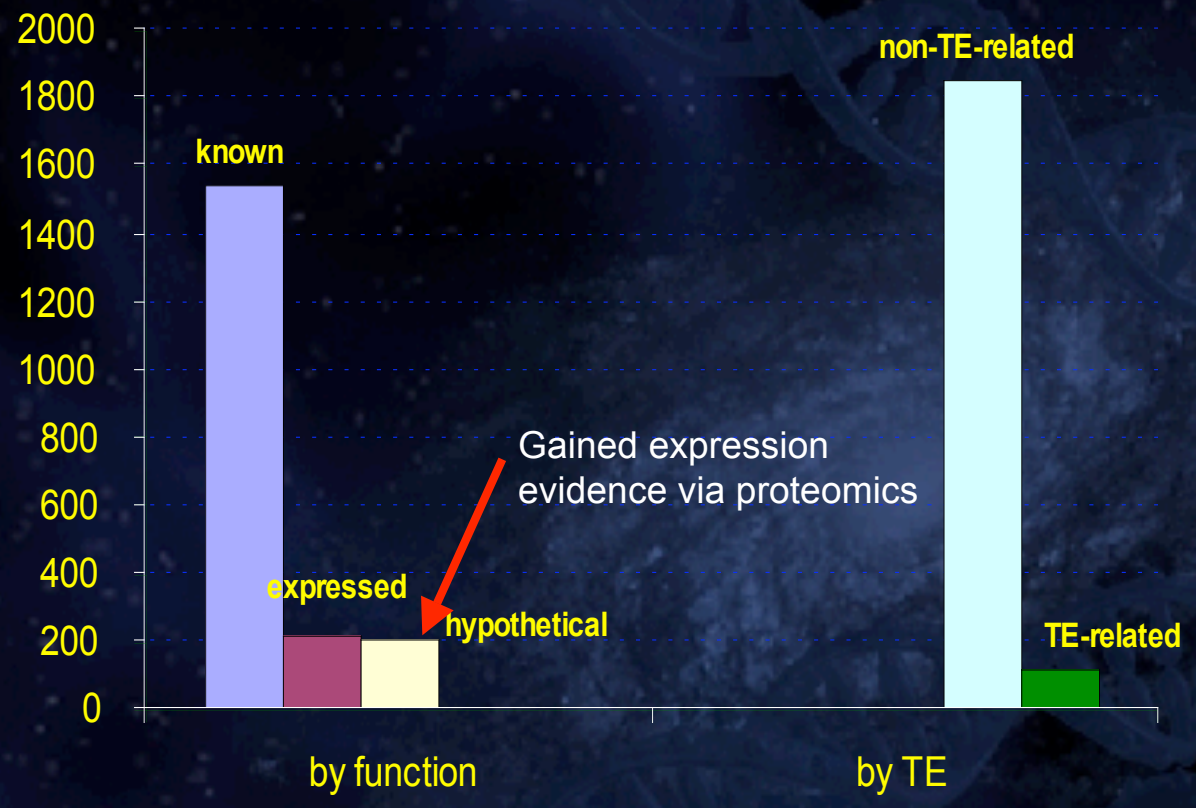
  11,941 proteins matching to 4,180 in databases

  http://gene64.dna.affrc.go.jp/RPD/main_en.html

  can search the DB, but no downloading

- Map the peptide sequences to the rice genome with blastp or tblastn

TIGR Models Matched by Syngenta Peptides



95% of Syngenta peptides were mapped to 1,965 TIGR rice models;

10.3% of the matched TIGR models were previously annotated as hypothetical proteins

blastp and tblastn, 100% identity and 100% coverage

Expression Support Tool

- Expression Evidence: EST, FLcDNA, MPSS, SAGE, Proteomic
- Report of available expression evidence for a locus

TIGR Rice Gene Expression Evidence Search Result

Locus	Model	MPSS	SAGE	Full length cDNA	Total number of mapped ESTs	Total number of mapped peptides
LOC_Os01g11880.1	12001.m07805	GATCTTGACTCTTGTTT GATCTGAGATAGAGGGA GATCTGAGATAGAGGGACTG GATCAGAGTGCATGTGACAG GATCTTGACTCTTGTTTCTG GATCAGAGTGCATGTGA	CATGACATAGTAATTCTGTGC CATGTGTAAAGAGTCGTCGTT	AK100463	22	1
LOC_Os01g01060.1	12001.m06753	GATCATCCACCTCCTCA GATCATCCACCTCCTCACCG GATCTGAGTTCTTTATG	n/a	AK059844 AK121523	77	2
LOC_Os01g01060.2	12001.m42817	GATCATCCACCTCCTCA GATCATCCACCTCCTCACCG GATCTGAGTTCTTTATG	n/a	n/a	0	2

green: MPSS tags mapped to the unique site on the genome and determined as significant tag. **This was the only dataset used in the TIGR rice annotation**.
red: MPSS tags mapped to the multiple sites and determined as significant tag.

n/a: not available

Level of Expression Support for Release 5

No. Genes/Loci Supported by Expression Evidence (EST/fl-cDNA, MPSS, SAGE and Proteomic): 26,178 (46.5%)

No. Gene Models Supported by Expression Evidence (EST/fl-cDNA, MPSS, SAGE and Proteomic): 36,509 (54.7%)

No. Gene/Loci supported **fully** by EST/FL-cDNA: 18,068

No. Gene Models supported **fully** by EST/FL-cDNA: 23,646

Expression Evidence for TIGR Rice Version 5 Models

Evidence	Locus		Model	
	count	percent	count	percent
Any expression evidence	26,178	46.52%	36,509	54.73%
PASA fully supported	18,068	32.10%	23,646	35.45%
fl-cDNA	17,096	30.38%	19,216	28.81%
EST	24,367	43.30%	33,807	50.68%
MPSS	20,423	36.29%	29,237	43.83%
SAGE	7,997	14.21%	13,052	19.57%
Proteomics	1,964	3.49%	2,983	4.47%

Papers

EST Papers:

Ronning et al. 2003. Comparative analyses of potato Expressed Sequence Tag libraries. Plant Physiology 131: 419-429.

FL cDNA papers:

Normalization and Subtraction of Cap-Trapper-Selected cDNAs to Prepare Full-Length cDNA Libraries for Rapid Discovery of New Genes; Carninci et al. ; Genome Research 10:1617–1630

Functional Annotation of aFull-Length *Arabidopsis* cDNA Collection; Seki et al.; SCIENCE VOL 296 5 2002

Collection, Mapping, and Annotation of Over 28,000 cDNA Clones from *japonica* Rice; The Rice Full-Length cDNA Consortium; 2003 VOL 301 SCIENCE

Whole Genome Sequence Comparisons and “Full-Length” cDNA Sequences: A Combined Approach to Evaluate and Improve *Arabidopsis* Genome Annotation;
Castelli et al. ; Genome Research 14:406–413

Papers

SAGE papers:

Serial analysis of gene expression; Velculescu et al.; 1995 Science 270:484-7

Analysing uncharted transcriptomes with SAGE; Velculescu et al.; TIG 2000, volume 16, No. 10 423

Microarray papers:

Expression profiling using cDNA microarrays; Duggan et al.; 1999; Nature Genetics Supp. 21:10

Exploring the new world of the genome with DNA microarrays; Brown and Botstein; 1999; Nature Genetics Supp. 21:33

High density synthetic oligonucleotide arrays; Lipshutz et al.; Nature genetics supplement volume 21 • january 1999

Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array; Singh-Gasson et al.; NATURE BIOTECHNOLOGY VOL 17 1999

Papers

Application of Arrays:

Empirical Analysis of Transcriptional Activity in the*Arabidopsis* Genome; Yamada et al. 2003 VOL 302 SCIENCE 842

Identification of Promoter Motifs Involved in the Network of Phytochrome A-Regulated Gene Expression by Combined Analysis of Genomic Sequence and Microarray Data; Hudson and Quail; *Plant Physiology*, 2003, Vol. 133, pp. 1605–1616