

Bioinformatics tools to analyze full transcriptomes

Notis Argiriou



Outline

RNA sequencing overview

Available Bioinformatic tools in the Institute

Examples

- Transcriptomes
  - RNA-Seq of *Solanum eleagnifolium*
  - RNA-Seq of *Salvia pomifera*
- Metatranscriptome analysis
  - Wastes from olive oil
  - *Laurencia microcladia*

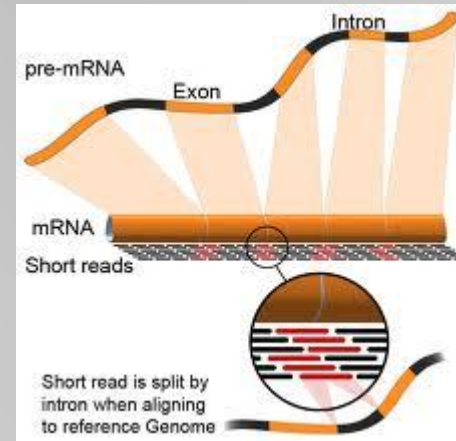
Conclusions

RNA-seq

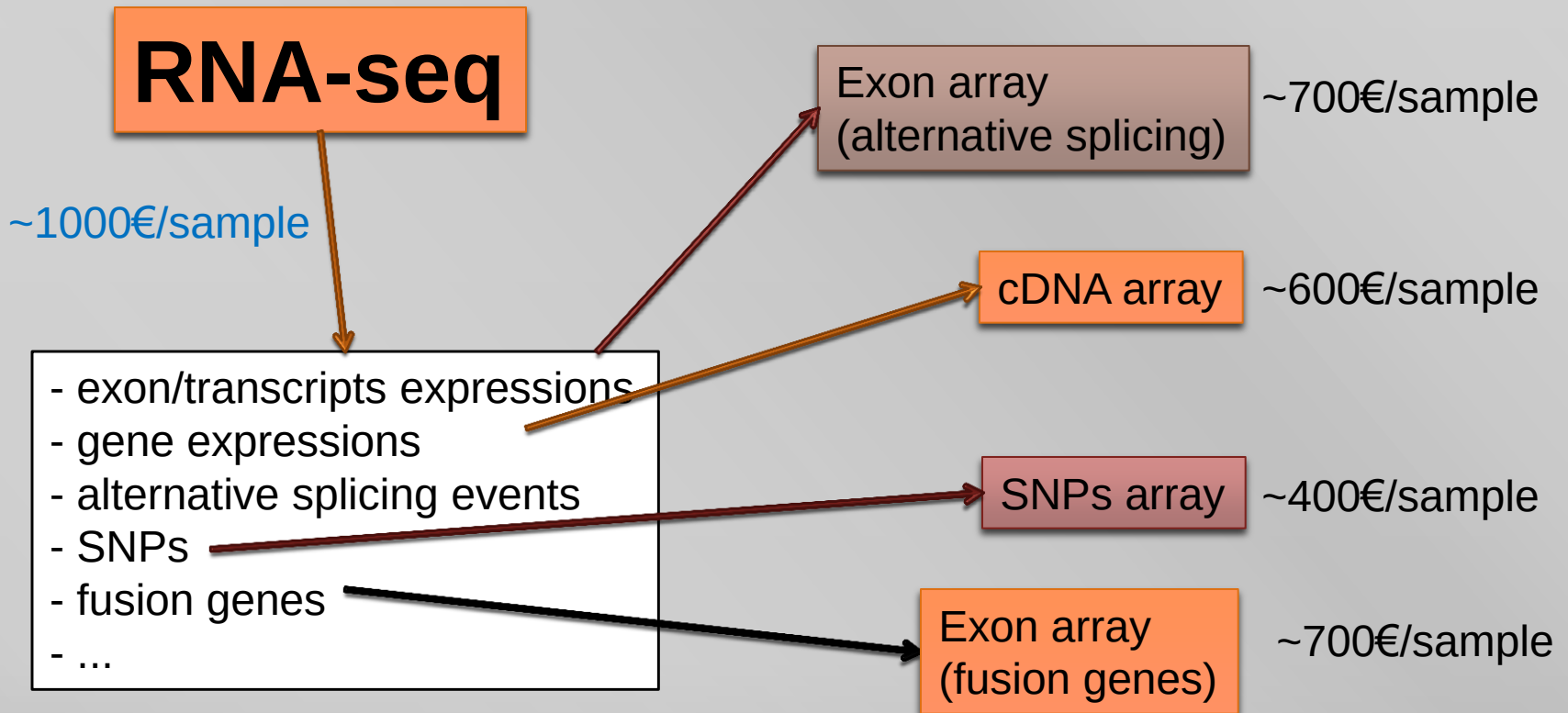
high-throughput
sequencing technology
for sequencing RNAs
invaluable tool for study
the gene content of an
organism

allows researchers to
obtain information like:

- gene/transcript/exon expressions
- alternative splicing
- gene fusions
- post-transcriptional mutations
- single nucleotide variations
- Identify organism in ecosystems
- ...



Why RNA-seq?



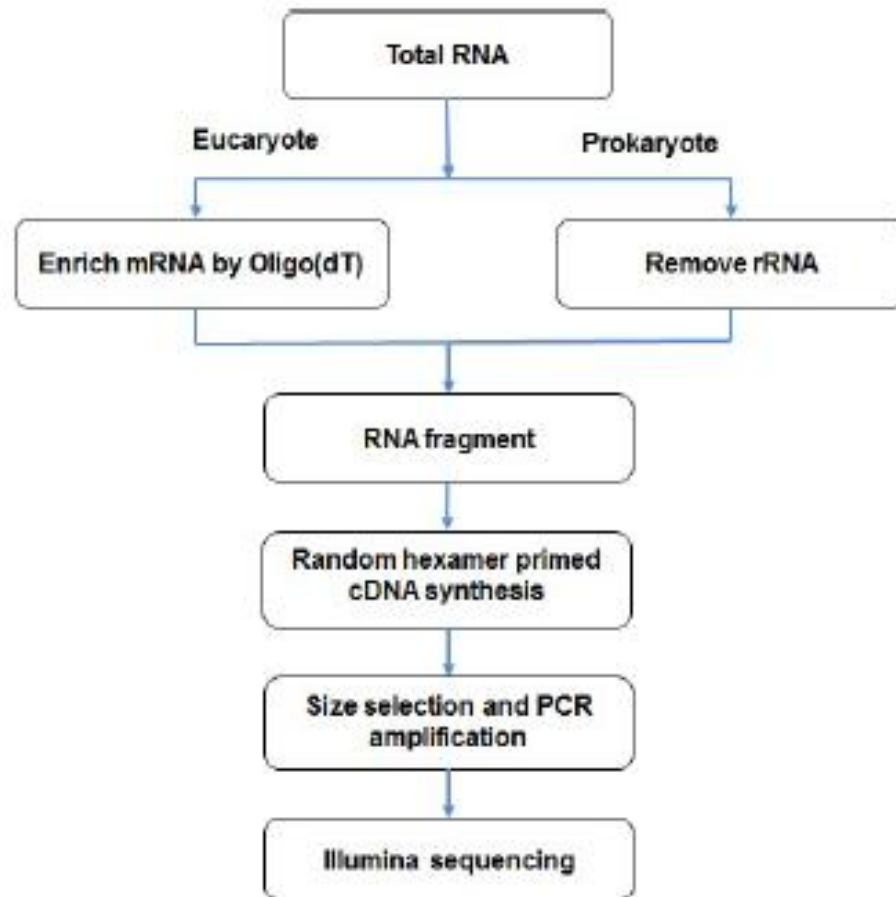
RNA-seq vs array technologies

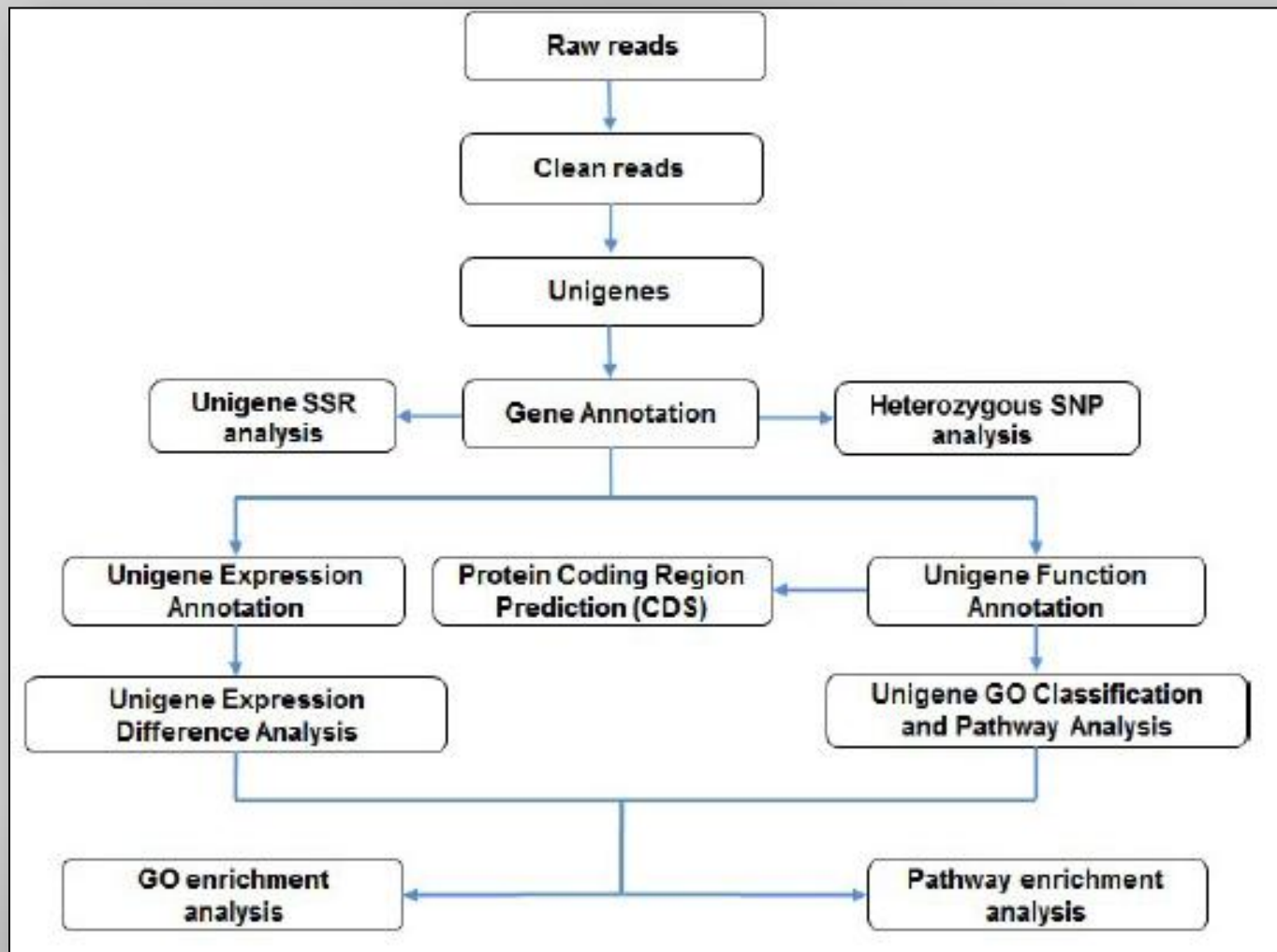
RNA sequencing overview



RNA-seq

- It reduces greatly the variability between experiments compared to other established measurement technologies like microarrays, exon arrays, etc.
- Due to the small size of the read (cDNA is fragmented before sequencing) the bioinformatics analysis is challenging, e.g.
  - *de novo* assembly
  - aligning of sequenced reads
  - computation of gene/transcript/exon expressions







General steps of RNA-seq analysis

Filtering of short reads

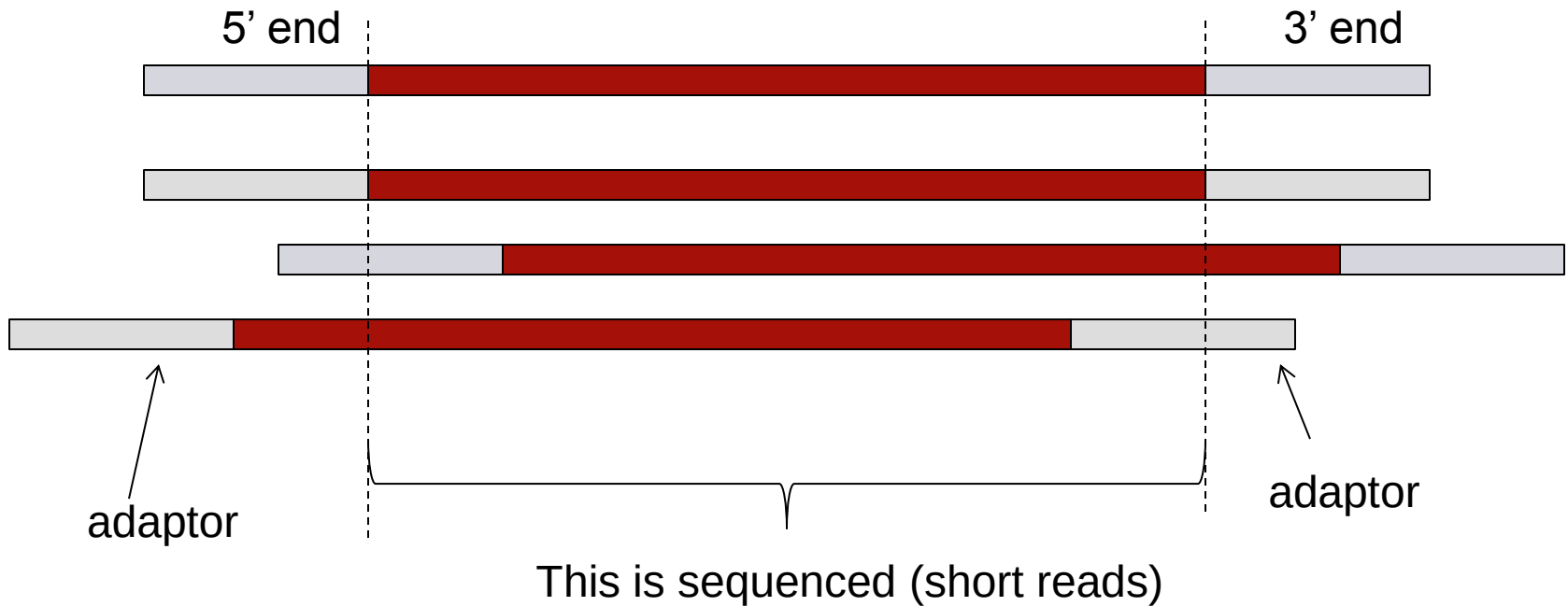
Aligning the reads against a reference (if any)

Computationaly analyse the reads' alignments

- compute the gene/transcript/exon expressions
- find new/known alternative splicing events
- find new/known fusion genes
- find new/known SNPs

Visualization

Filtering of short reads



Adaptor and reads in RNA-seq

Data file format

FASTQ format

```
@SRR047027.1 GFTKDI01EMEVN length=68
CACGGAATTAGCCGGTGCTTATTCTTCAGGTACCGTCATCAGCTGTCGATA
+SRR047027.1 GFTKDI01EMEVN length=68
FFFFGGFFFEEDDGGFFFFFHFFFFF@??FFCCDFFFFFFFDDDF
@SRR047027.3 GFTKDI01ETSLN length=54
CACGTAGTTAGCCGTGGCTTCTGCTTAGATACCGTCAAGGCACACAGGA
+SRR047027.3 GFTKDI01ETSLN length=54
IIIIIIIIIIIIIIHHIIIIIIIIIIHHIIIIIIIFA999;
@SRR047027.5 GFTKDI01ANQX8 length=44
CACGTATTAGCCGTCACCTTCTCTGTTGGTACCGTCAITTTT
+SRR047027.5 GFTKDI01ANQX8 length=44
IIIIIIHHIIIIIIIIIIIIIIFFFFBA>?=-,,,,,
@SRR047027.7 GFTKDI01D721L length=215
```

1. Tag of eliminate adaptors
2. Quality check
  - Base quality
  - Presence of repeats
  - Poly-A tag or trim

FASTA format

```
>9181024
AGCCGCGCCGAATTCTGGGCACGCTGGTTCGAAGGCATCGTCAGCTTCACGCCCGCGGCGCTTCTCGCGGCCAACCCGAGCGCGCAGTTCCAGCTGGGG
>9181029
TGTTCCGGCCGATTGGTGATTTCGAGATCAGCATCAAGCTCTCTCGAAACGGGATCCCTTCTCGTCATGGGAGATGAGCCTAGCGACAG
>9181034
TGCGCAATTGCCGTGGCAGCAGATAGGCCAGGATCACCCGCGCCGTGGCGCTGTGGAACAGGTCCATGGCGCTGCCGCGCCCATAGTTCAGGCG
>9181039
CAGGCCGATACCCGATATCCAGGCGAAATACCACATGATGTTCTCTCGTTGCGATTTGCGATTGCGCGTAATCAATACGCCGTGTGTTTCGT
>9181044
CAGCAGCACGCGGCCACACCGCTGCGGGCCGCGGCACTGGTGGCTTCGTCGCGCTGTAATAGTCGCCCAGCAGGCGCTGCAG
>9181049
CGATGCAATGAGGCAGGTTATGGAGGCGACGAGTGTGGAGAGGCAGATGAAGAGCTTTGAGGCGCGAAGGAGAGGAAAGGCTGAGAAGGGCTTCCG
>9181054
TTGAATCATTATTGACCGCAGAGCAGATTCACTTGGTGTTATCCGCTGCTACTATGGAGGAACTTAGTCAAATATGCAAAACGTTTCTACGCAATTTG
>9181059
GAAGAGGTGGGCTTTGAGCAGCCGCTTAAACATGGATGGGGAAGACTATTCCACAATGTTGGGGCTAAACACGCGAAGGCACGGTCACTTTAGATTACAGGTAGAGCATGG
>9181064
CTCGCCGTGATAGATGTTGTACCCAGAATCAGTGCCGAGTTCTCGCCGTTGATGAAGTCGCGGCCGATGATGAAGGCTTGTGCCAGACCA
>9181069
AGGATTGGCAACTCCTAGTGATCTTAGGAATGCATTGATGGGACCATCTGGGCATTGTAAAGCACACTCCACACCAGAAACGGCAACCAAACCTG
```

Fastq file structure

How to filter the reads?

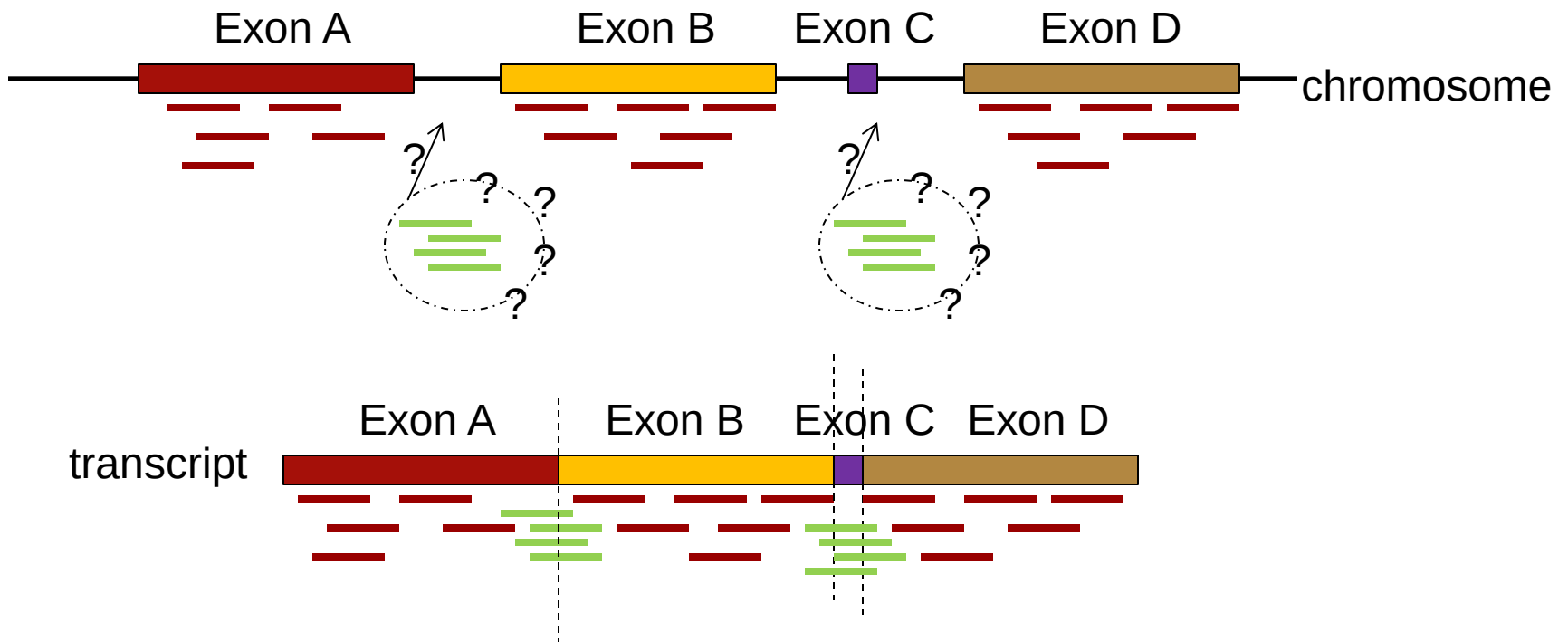
Using scripts (Shell, Perl, Python etc)

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTTTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTTCTAGAATAAGGGTAACTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAAACTTGTTTGTGATGTGTGTATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACCTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

Nucleotide sequence

Quality score in ASCII

Aligning the reads against a reference (if any)



Reads' mappings at chromosome and transcript level

$2 \times 10^9 / 100 =$
20.000.000
reads

$1 \times 10^9 / 250 =$
4.000.000
reads

$6 \times 10^{10} / 100 =$
6.000.000.000
reads

Output of main NGS platform

Table 1 Technical specifications of Next Generation Sequencing platforms utilised in this study

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	\$128 K	\$80 K**	\$695 K	\$256 K	\$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	\$502	\$1000 (318 chip)	\$2000	\$148	\$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

Available Bioinformatics tools in the Institute for RNA-Seq

■ Assemblers

- Phred/phrap/consed
- MIRA
- VELVET
- Abyss
- Soapdenovo
- Soapdenovo Trans
- Trinity

■ Tools

- Emboss
- GALAXY
- MEGAN
- BLAST
- HMMER
- TopHat
- Cufflinks
- MG-RAST

• Computer power

• Testing Server

- 8 cores RAM 14 Gb
- HD 3.5 Tb

• Working cluster (In progress)

- 36 Cores
- RAM 384 Gb
- HD 12 Tb High Speed SSD

• OS

- Ubuntu Server 12.04

• Databases

- nt
- nr
- PFAM
- 16S
- Custom databases



Assemblers

- **Velvet**: Sequence assembler for very short reads
- **Assembly By Short Sequences** - a de novo, parallel, paired-end sequence assembler
- **MIRA 3** - Whole Genome Shotgun and EST Sequence Assembler for Sanger, 454 and Solexa / Illumina
- **SOAPdenovo** and **SOAPdenovo Trans** is a novel short-read assembly method that can build a de novo draft assembly for the human-sized genomes. The program is specially designed to assemble Illumina short reads.
- **Trinity** represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data



Tools

- **EMBOSS** : A high-quality package of free, Open Source software for molecular biology (160.40.70.143/emboss-explorer)
- **Galaxy** is an open, web-based platform for data intensive biomedical research.
- **MEGAN 4** - MEtaGenome Analyzer: Software for analyzing metagenomes.
- **MG-RAST** – Metagenomic web tool

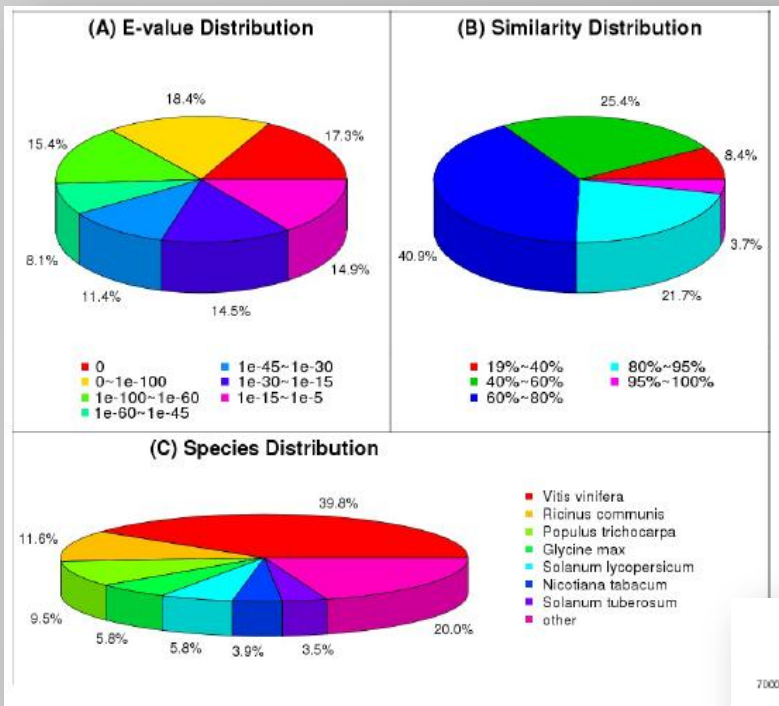
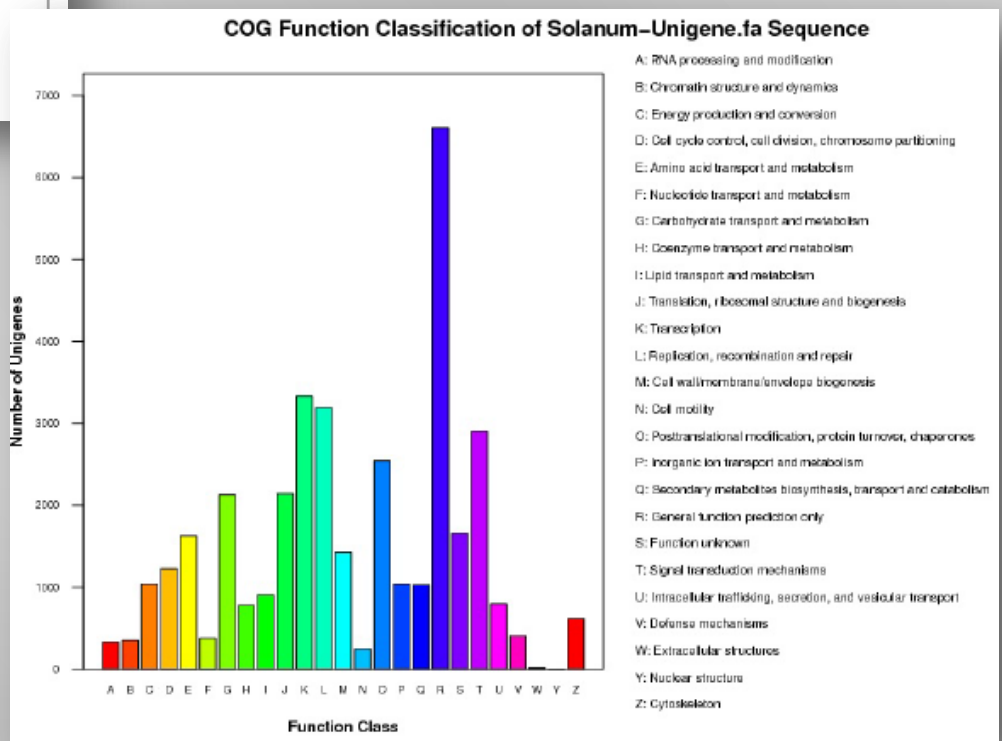


Conclusions

- RNA-seq data analysis:
  - is computational intensive (when compared to, for example, microarray analysis)
  - needs very good filtering criteria, which are based on biology mathematics, in order to improve the quality of the results (i.e. low number of false positives)
  - there is not only one established way of doing it
  - many tools used for analysis, e.g. aligners, samtools, etc., are still work in progress
- Visualization:
  - multiple facets, i.e. read coverage, fusion genes, etc.
  - depends on the user profile:
    1. biologist/medical doctor
    2. bioinformatician

RNA-Seq of *Solanum eleagnifolium*





RNA-Seq of *Salvia pomifera*



Output statistics of sequencing

Samples	Total Raw Reads	Total Clean Reads	Total Clean Nucleotides (nt)	Q20 percentage	N percentage	GC percentage
Salvia_pomifera	54,000,000	45,887,274	4,129,854,660	95.64%	0.00%	49.37%

* Total Clean Nucleotides = Total Clean Reads1 x Read1 size + Total Clean Reads2 x Read2 size

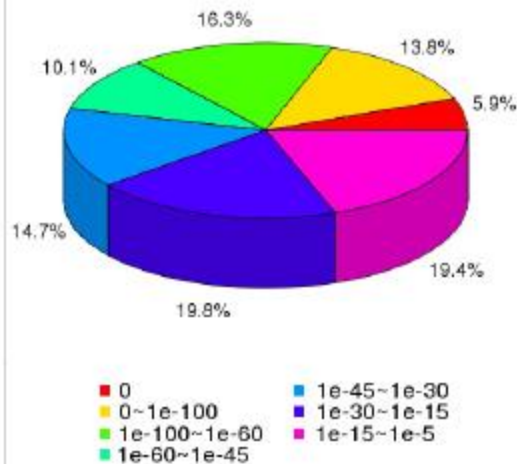
Statistics of assembly quality

	Sample	Total Number	Total Length(nt)	Mean Length(nt)	N50	Total Consensus Sequences	Distinct Clusters	Distinct Singletons
Contig	Salvia_pomifera	125,371	40,286,050	321	531	-	-	-
Unigene	Salvia_pomifera	66,051	42,298,081	640	891	66,051	22,395	43,656

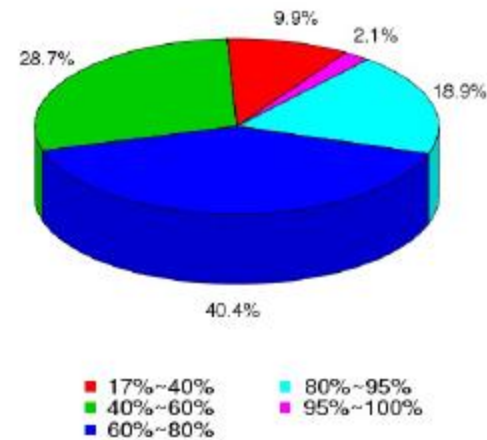
Statistics of annotation results

Sequence File	NR	NT	Swiss-Prot	KEGG	COG	GO	ALL
Salvia_pomifera-Unigene.fa	46,270	32,857	29,395	27,679	17,303	33,332	47,687

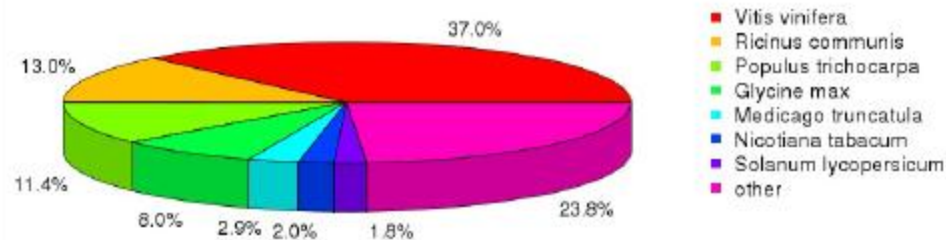
(A) E-value Distribution



(B) Similarity Distribution



(C) Species Distribution





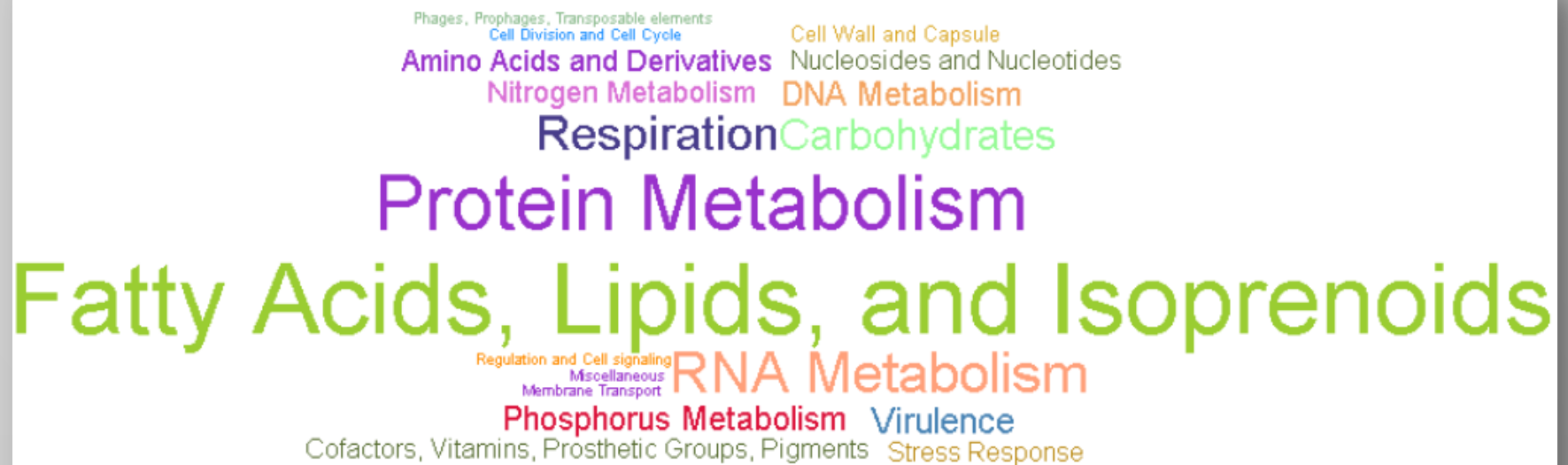
RNA-Seq of ecosystems

- Metagenome analysis of Table Olive Processing Wastewater
  - IonTorrent sequencing
  - 3 runs
- RNA-Seq of the red algae *Laurencia microcladia*
  - Illumina sequencing
  - 1 run 4G

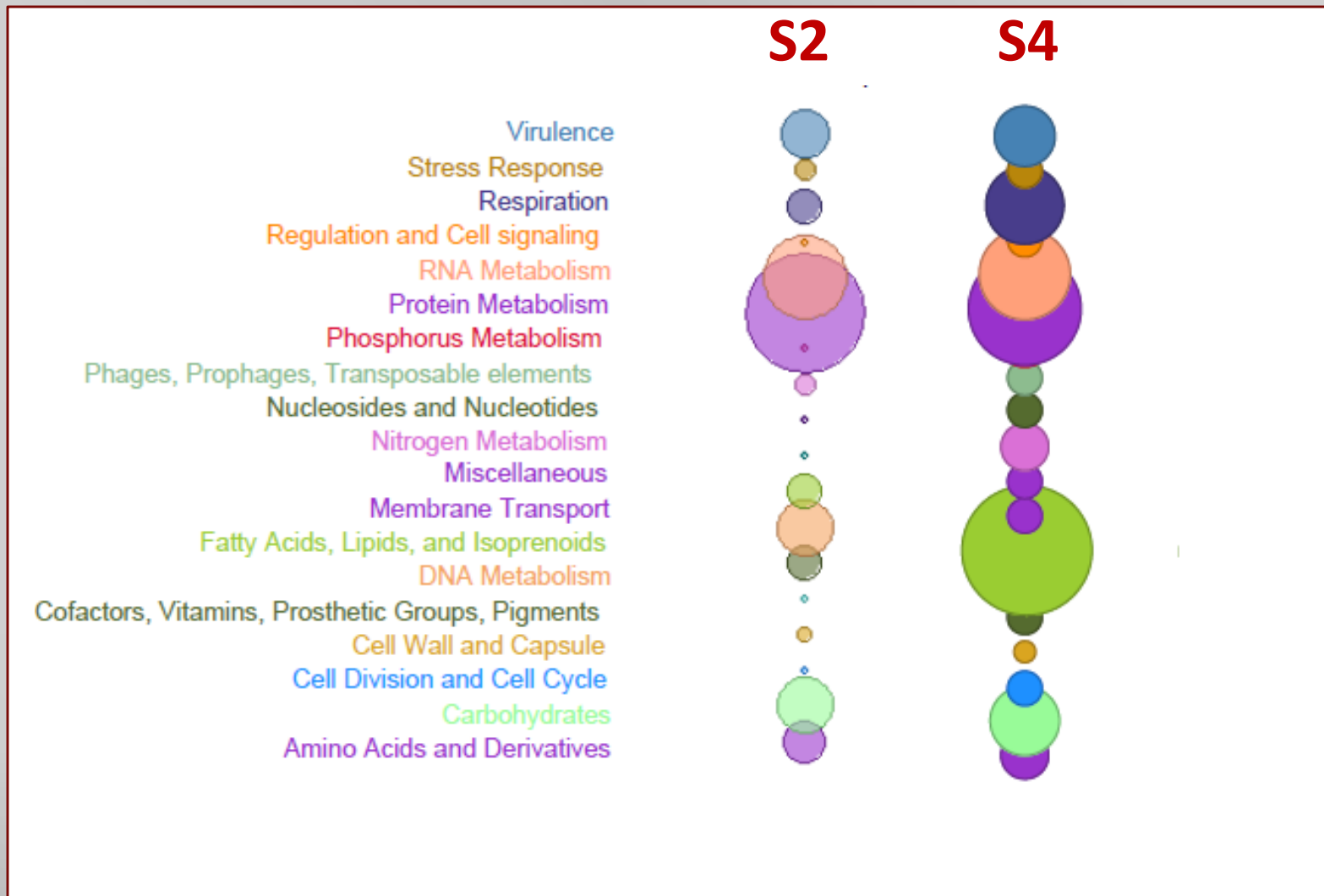
Table Olive Processing Wastewater

Barcode Name	Sample	Output	%>=Q20	Reads	Mean Read Length	Read Length Histogram	BAM
No barcode	3MBR-19-9-12	8.4M	7.2M	74712	113 bp		
RNA_Barc...	3MBR-19-9-12	43.6M	38.7M	422754	103 bp		
Barcode Name	Sample	Output	%>=Q20	Reads	Mean Read Length	Read Length Histogram	BAM
No barcode	2MBR8-8-12	6.8M	5.7M	63607	108 bp		
RNA_Barc...	2MBR8-8-12	36.2M	30.8M	340066	106 bp		
Total Number of Bases [Mbp]			26.94				
► Number of Q20 Bases [Mbp]			24.02				
Total Number of Reads			330,484				
Mean Length [bp]			81				
Longest Read [bp]			376				

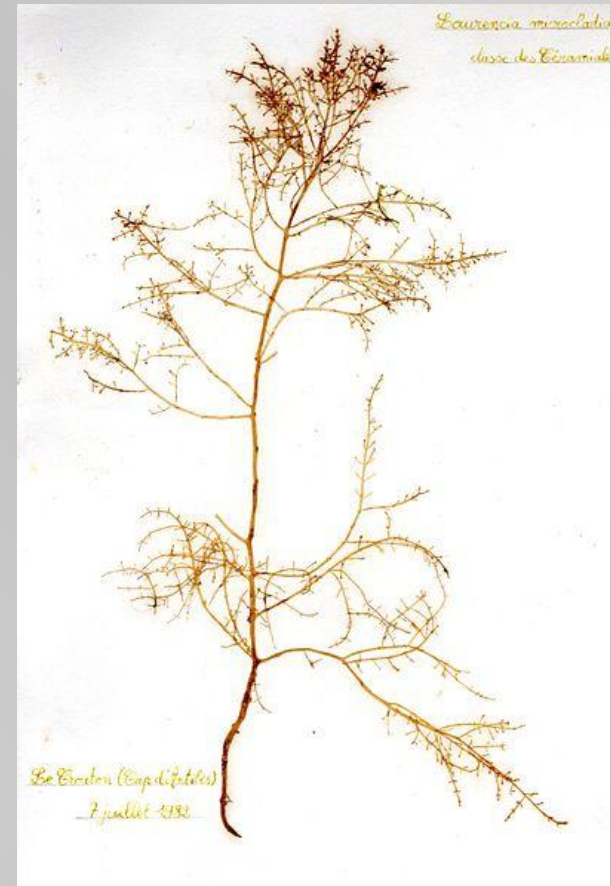
SEED Profile of Sample 4



SEED Profiles Comparison



Laurencia microcladia



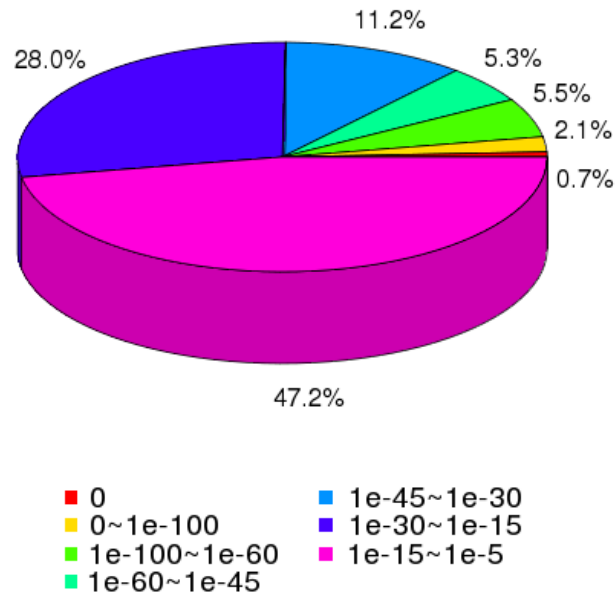
Sequencing Results

Samples	Total Raw Reads	Total Clean Reads	Total Clean Nucleotides (nt)	Q20 percentage	N percentage	GC percentage
MPO_290	49,191,548	45,811,836	4,123,065,240	97.06%	0.00%	32.05%

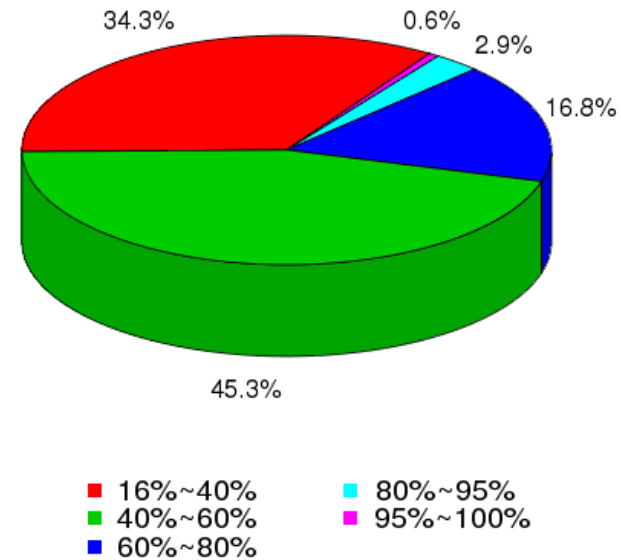
Assembly Results

	Sample	Total Number	Total Length(nt)	Mean Length(nt)	N50	Total Consensus Sequences	Distinct Clusters	Distinct Singletons
Contig	MPO_290	309,845	51,640,879	167	147	-	-	-
Unigene	MPO_290	40,746	19,262,933	473	533	40,746	11,556	29,190

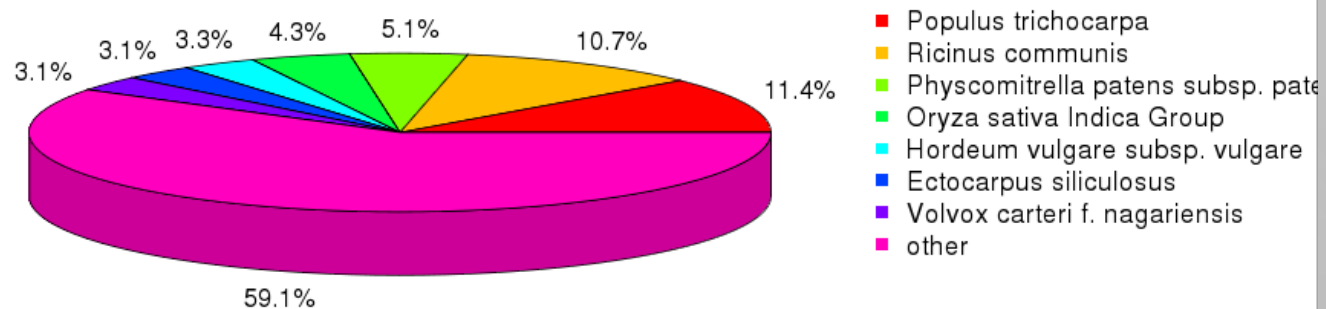
(A) E-value Distribution



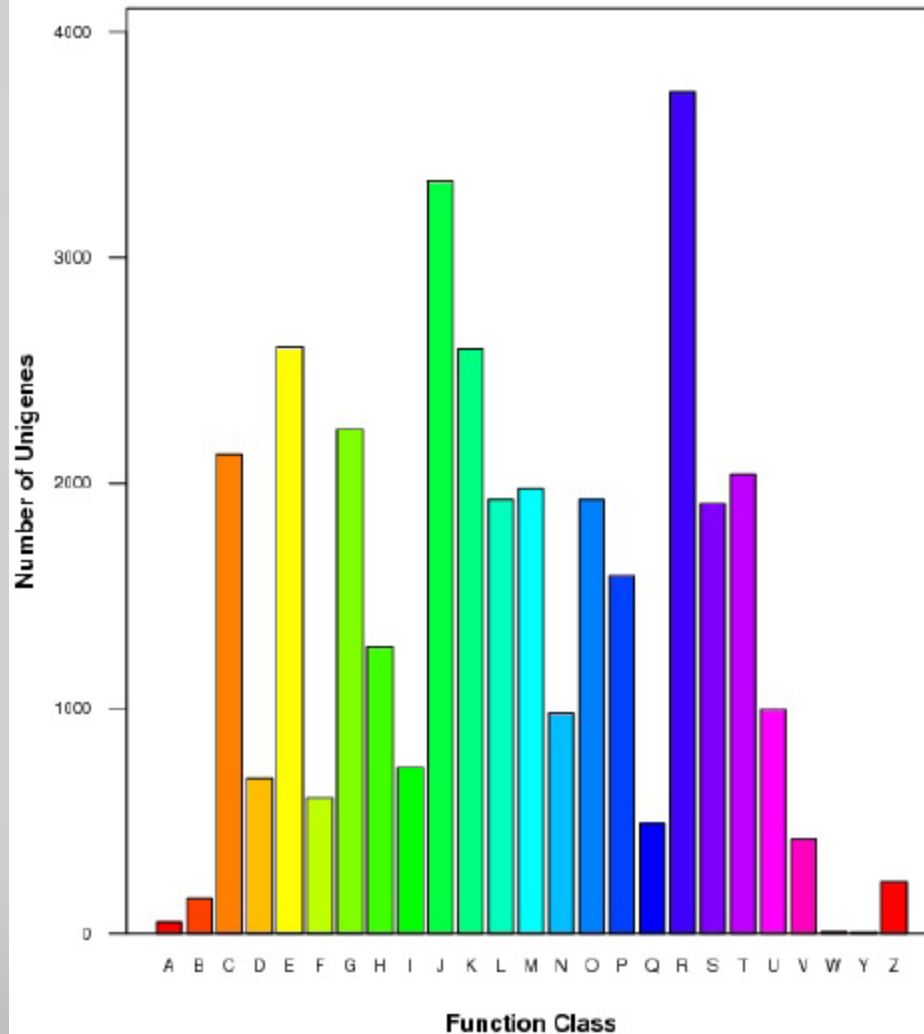
(B) Similarity Distribution



(C) Species Distribution

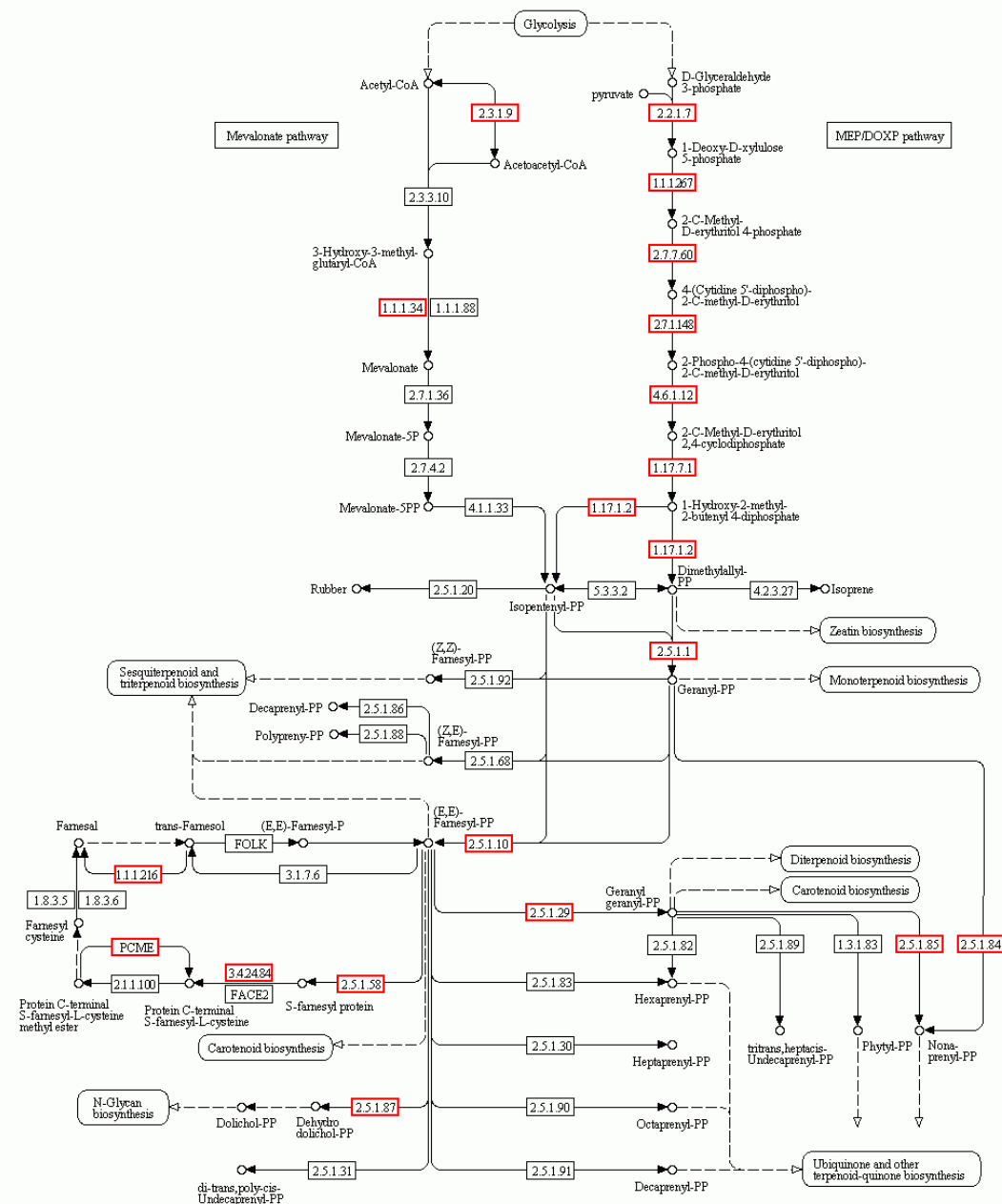


COG Function Classification of MPO_290-Unigene.fa Sequence



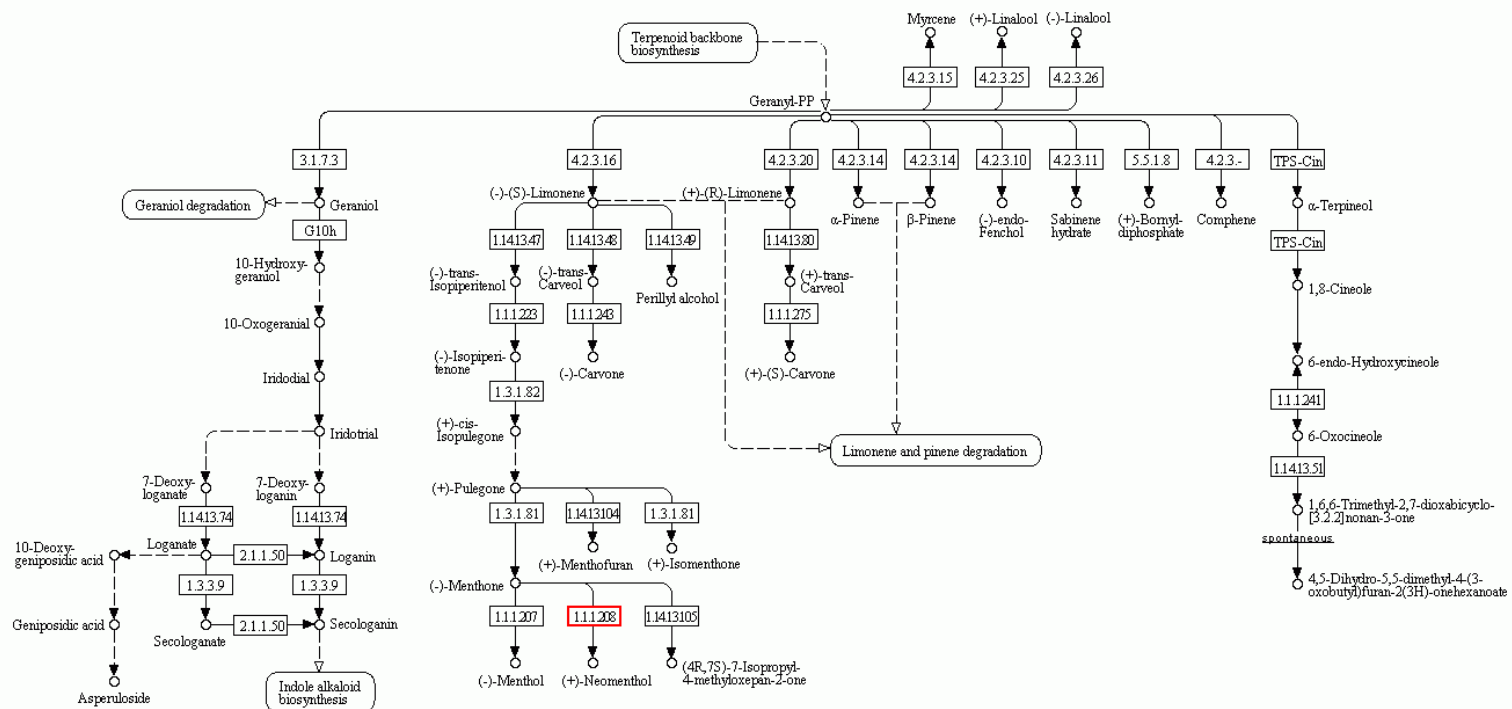
- A: RNA processing and modification
- B: Chromatin structure and dynamics
- C: Energy production and conversion
- D: Cell cycle control, cell division, chromosome partitioning
- E: Amino acid transport and metabolism
- F: Nucleotide transport and metabolism
- G: Carbohydrate transport and metabolism
- H: Coenzyme transport and metabolism
- I: Lipid transport and metabolism
- J: Translation, ribosomal structure and biogenesis
- K: Transcription
- L: Replication, recombination and repair
- M: Cell wall/membrane/envelope biogenesis
- N: Cell motility
- O: Posttranslational modification, protein turnover, chaperones
- P: Inorganic ion transport and metabolism
- Q: Secondary metabolites biosynthesis, transport and catabolism
- R: General function prediction only
- S: Function unknown
- T: Signal transduction mechanisms
- U: Intracellular trafficking, secretion, and vesicular transport
- V: Defense mechanisms
- W: Extracellular structures
- Y: Nuclear structure
- Z: Cytoskeleton

TERPENOID BACKBONE BIOSYNTHESIS

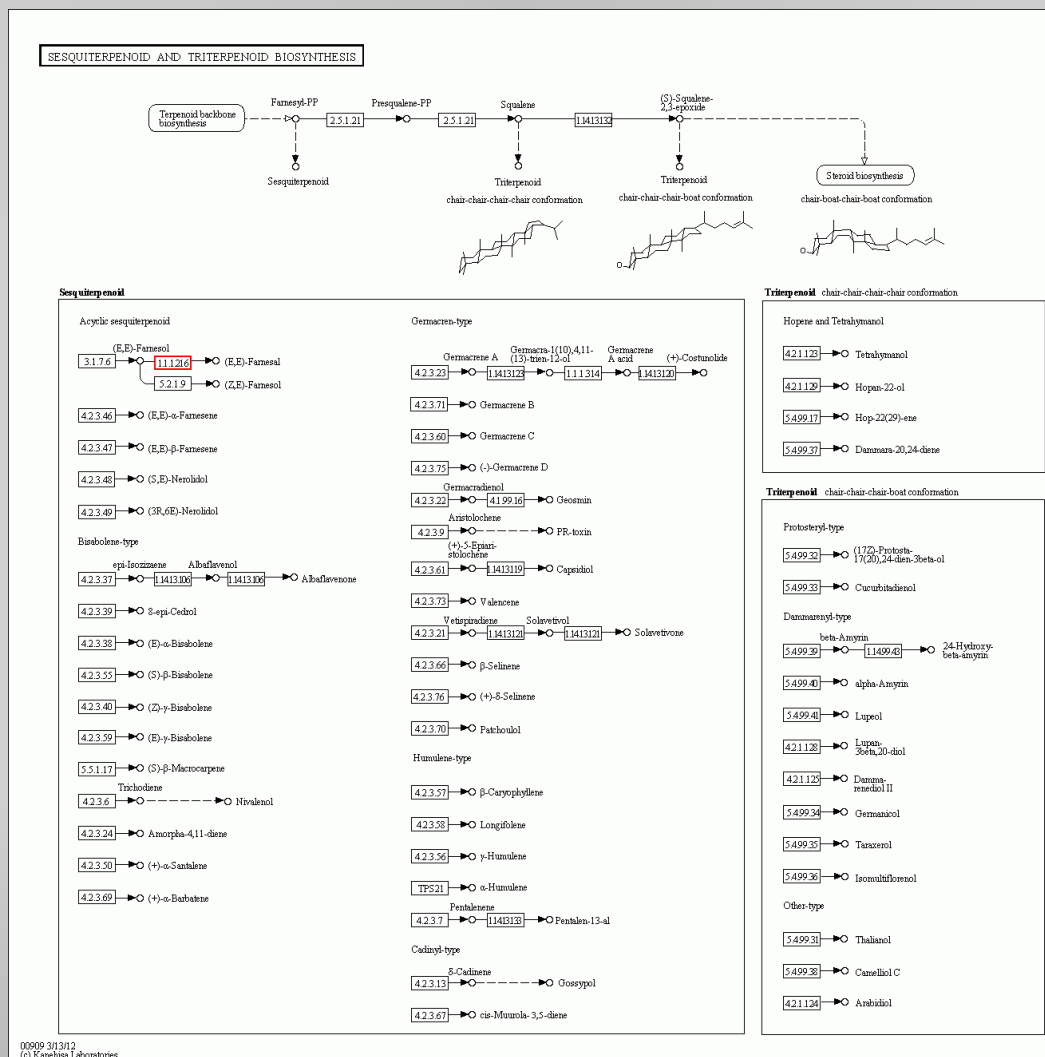


Monoterpenoid biosynthesis

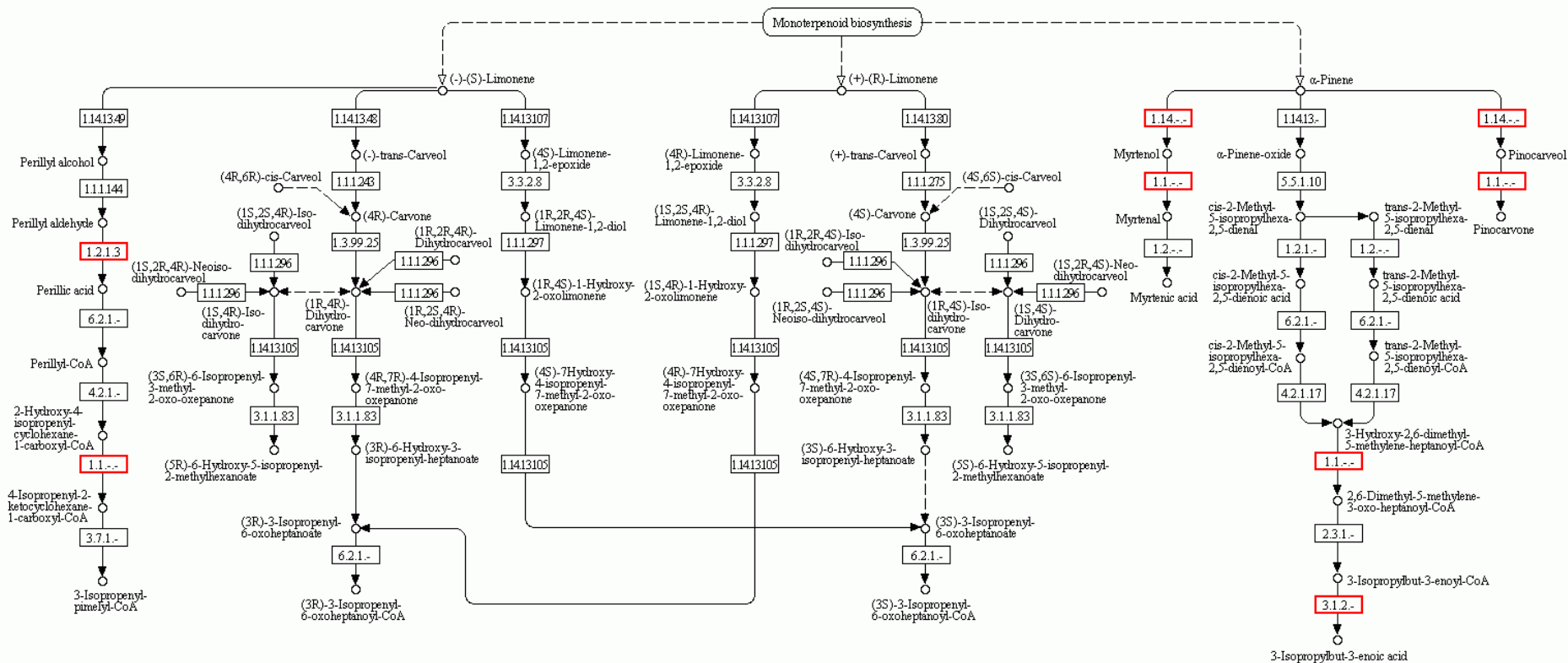
MONOTERPENOID BIOSYNTHESIS

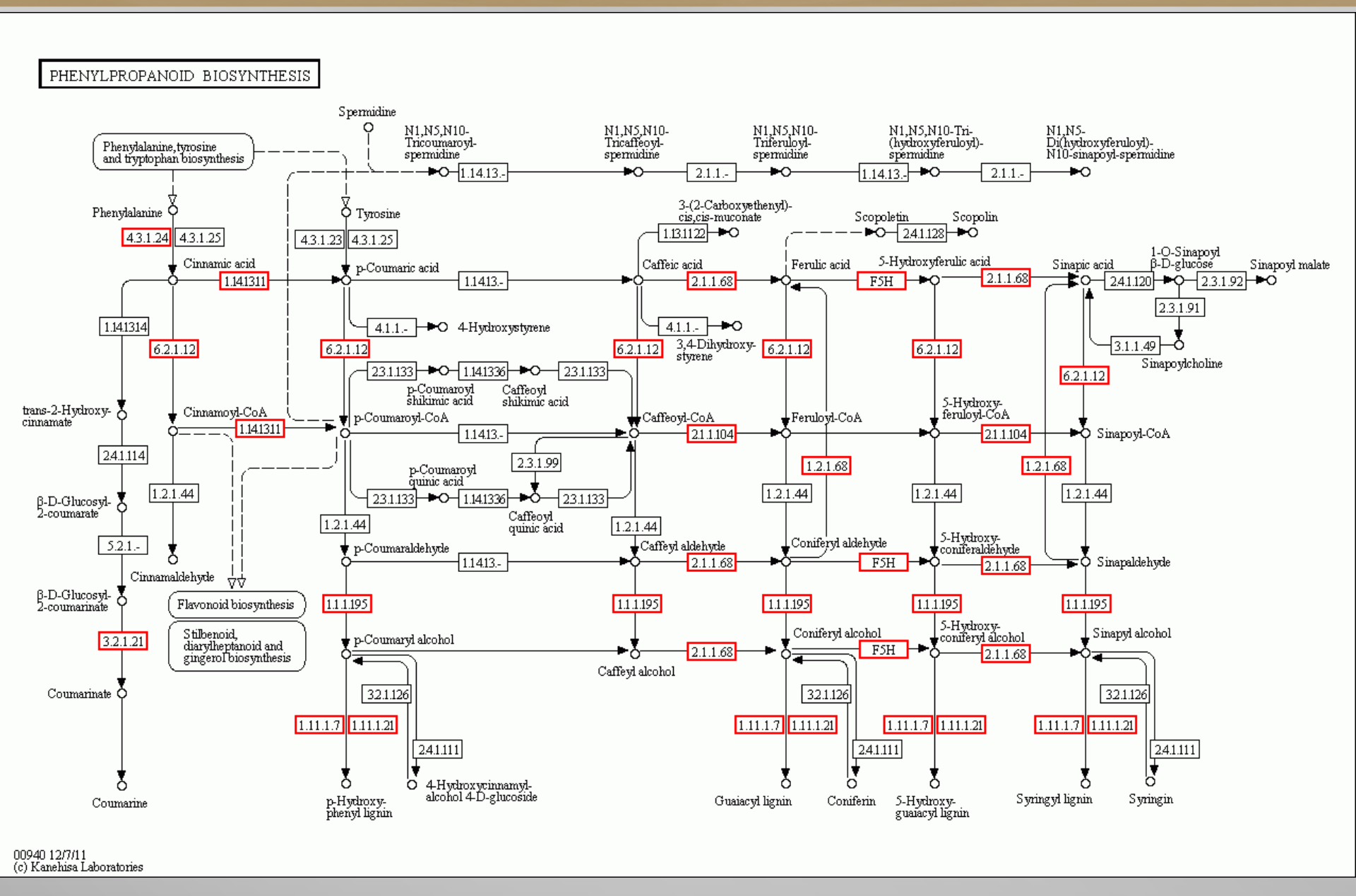


Sesquiterpenoid and triterpenoid biosynthesis

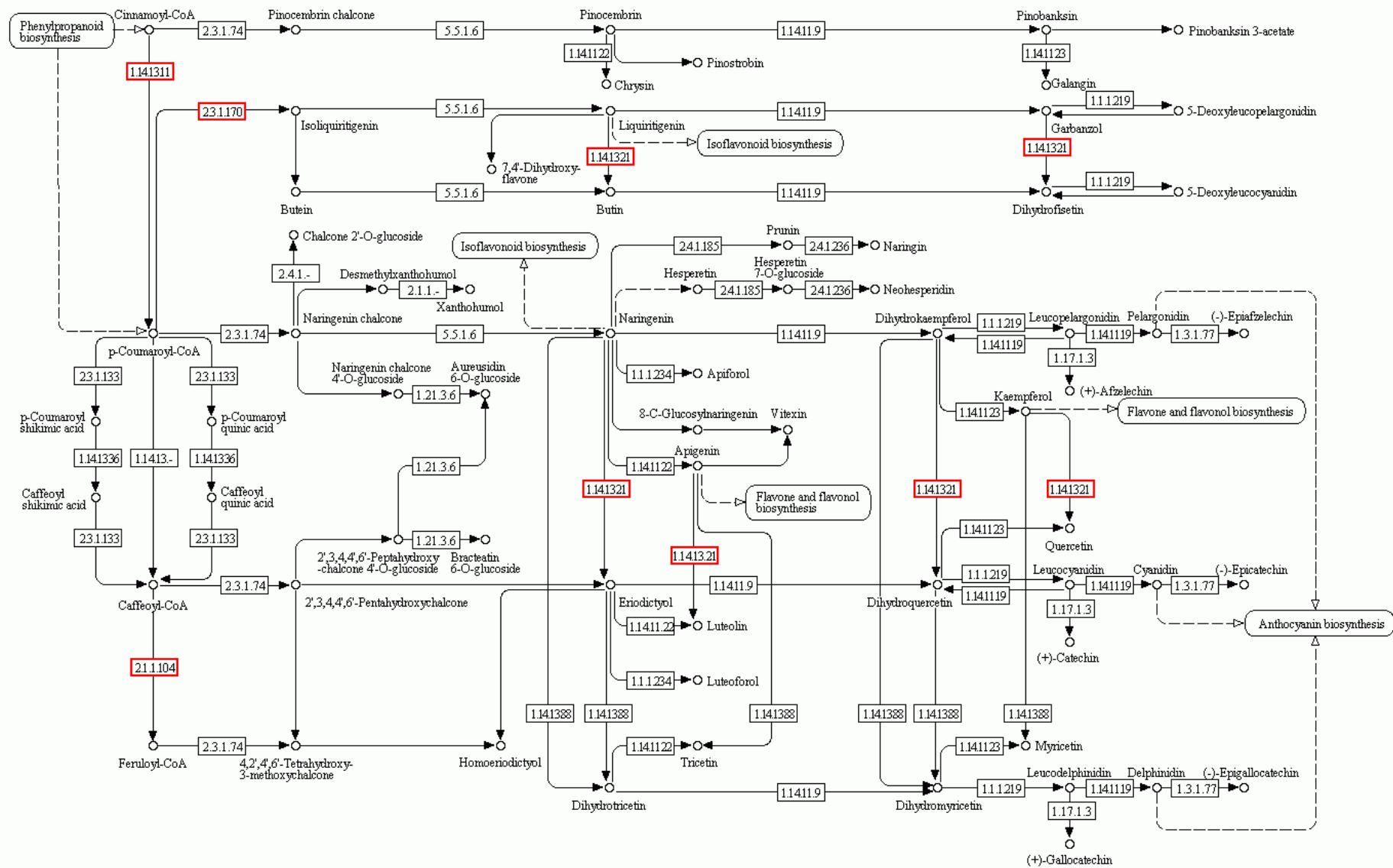


LIMONENE AND PINENE DEGRADATION

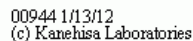




FLAVONOID BIOSYNTHESIS

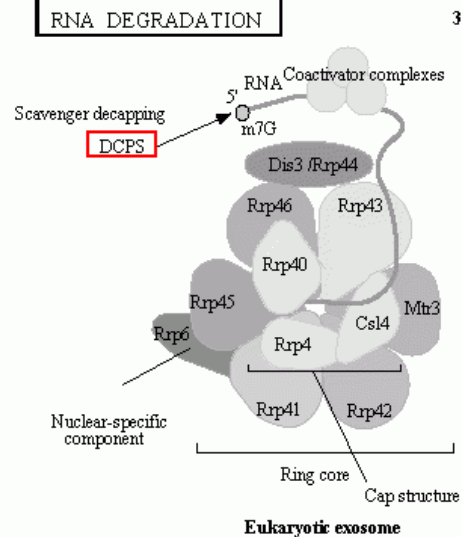


4'-O-Methylapigenin



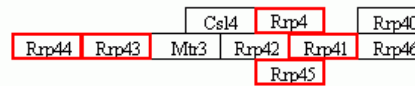
Eukaryotic RNA degradation

RNA DEGRADATION



3' → 5' decay

Eukaryotic core exosome



Exosome associated factors

Nuclear exosome cofactors

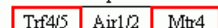


Cytoplasmic exosome cofactor

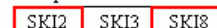


Exosome coactivator complexes

TRAMP complex

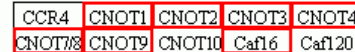


Ski complex



Cytoplasmic deadenylation

Ccr4-NOT complex



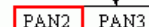
RHAU

PARN

TOB

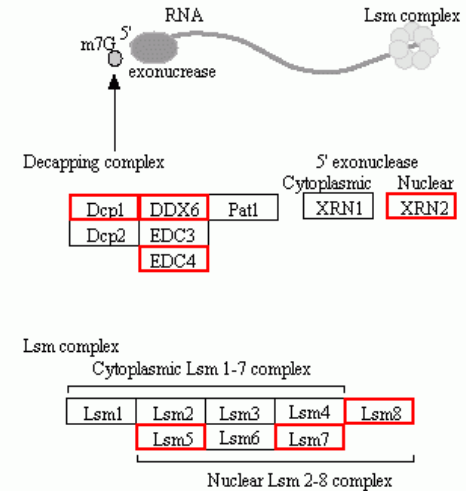
PABP1

Pan complex

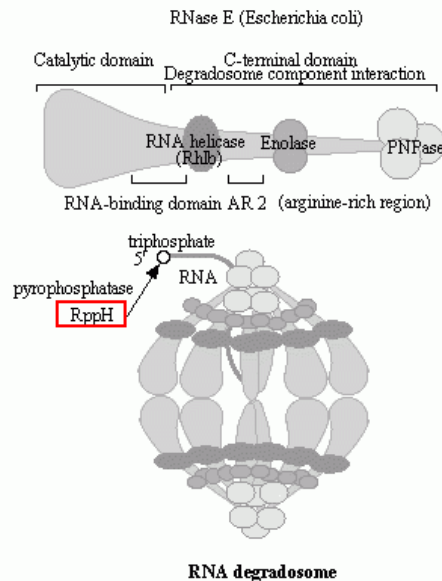


mRNA surveillance pathway

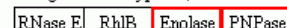
5' → 3' decay



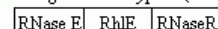
Bacterial RNA degradation



RNA degradosome type A (Escherichia coli)



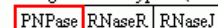
RNA degradosome type B (Pseudomonas)



RNA degradosome type C (Rhodobacter)



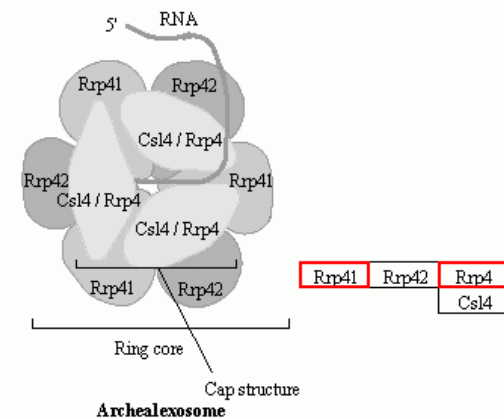
RNA degradosome type D (Bacillus subtilis)



Associated proteins



Archaeal RNA degradation



Thank you