4. Transcriptome (RNA-Seq) analysis

4.0. Introduction

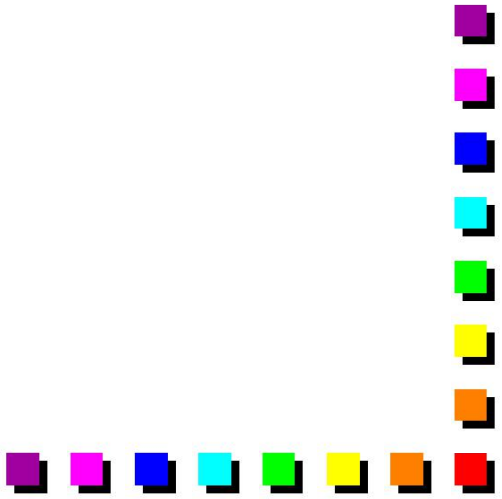
4.1. RNA-Seq and sequence annotation

4.2. Application of RNA-Seq results

0. Introduction

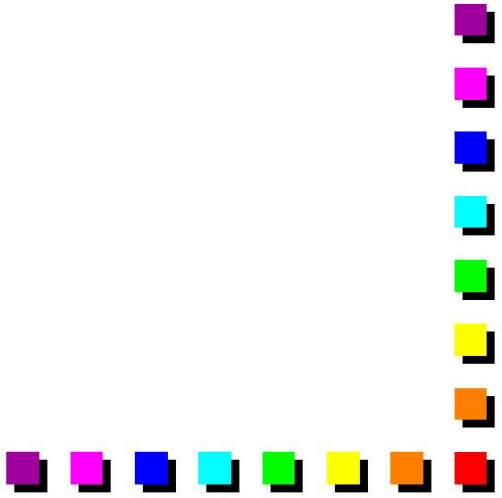
- Two main purpose:Exone sequences: exome; estimation of gene expression
- cDNA (1977-1979), EST (1991) and full-length cDNA (1982; 2003 rice)
- NGS (1996-2004) and RNA-Seq (2008)

- Microarray chip, SAGE

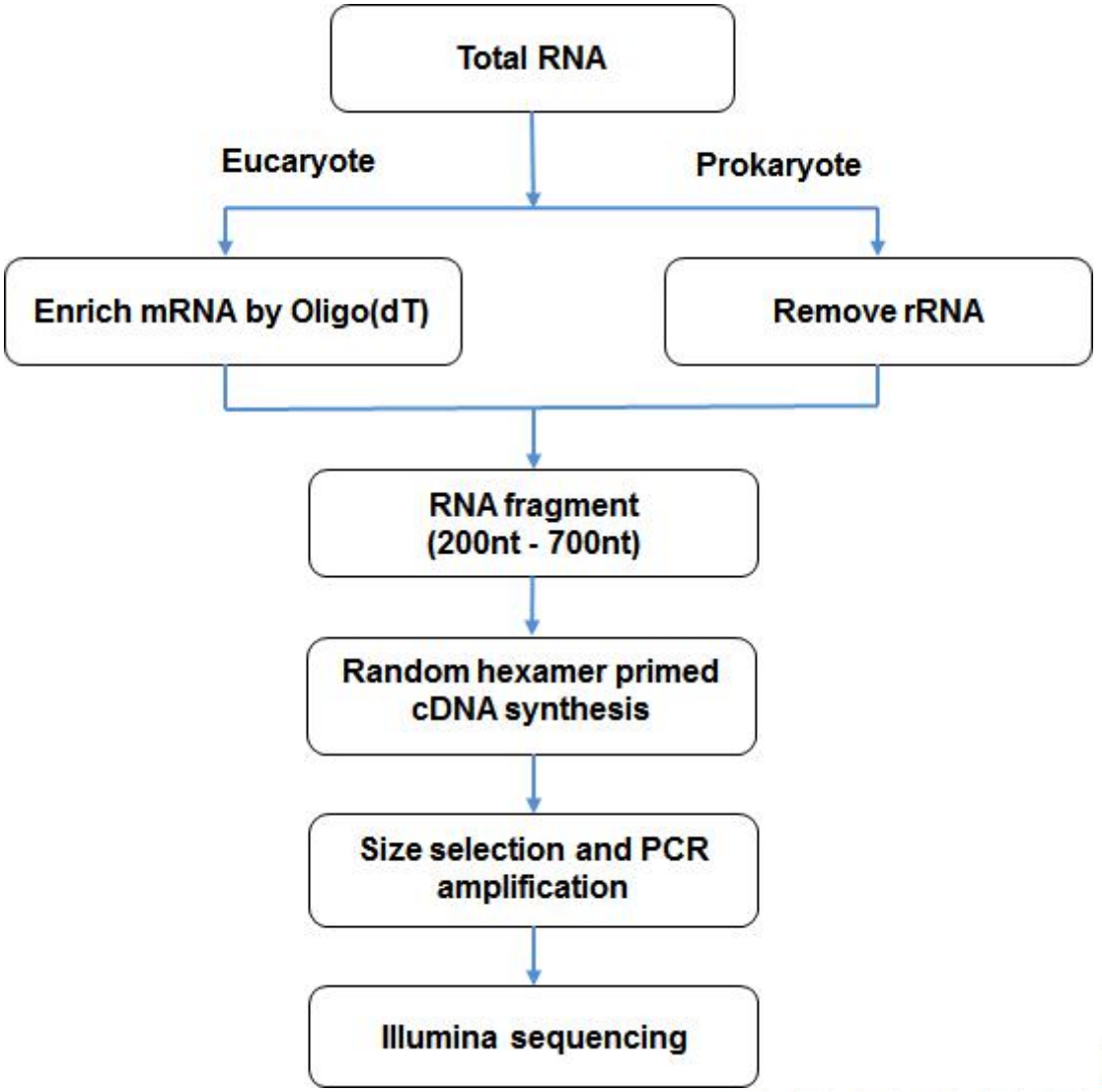


1. RNA-Seq

- also called "Whole Transcriptome Shotgun Sequencing"
- To sequence cDNA in order to get information about a sample's RNA content by the high-throughput approaches



RNA collection

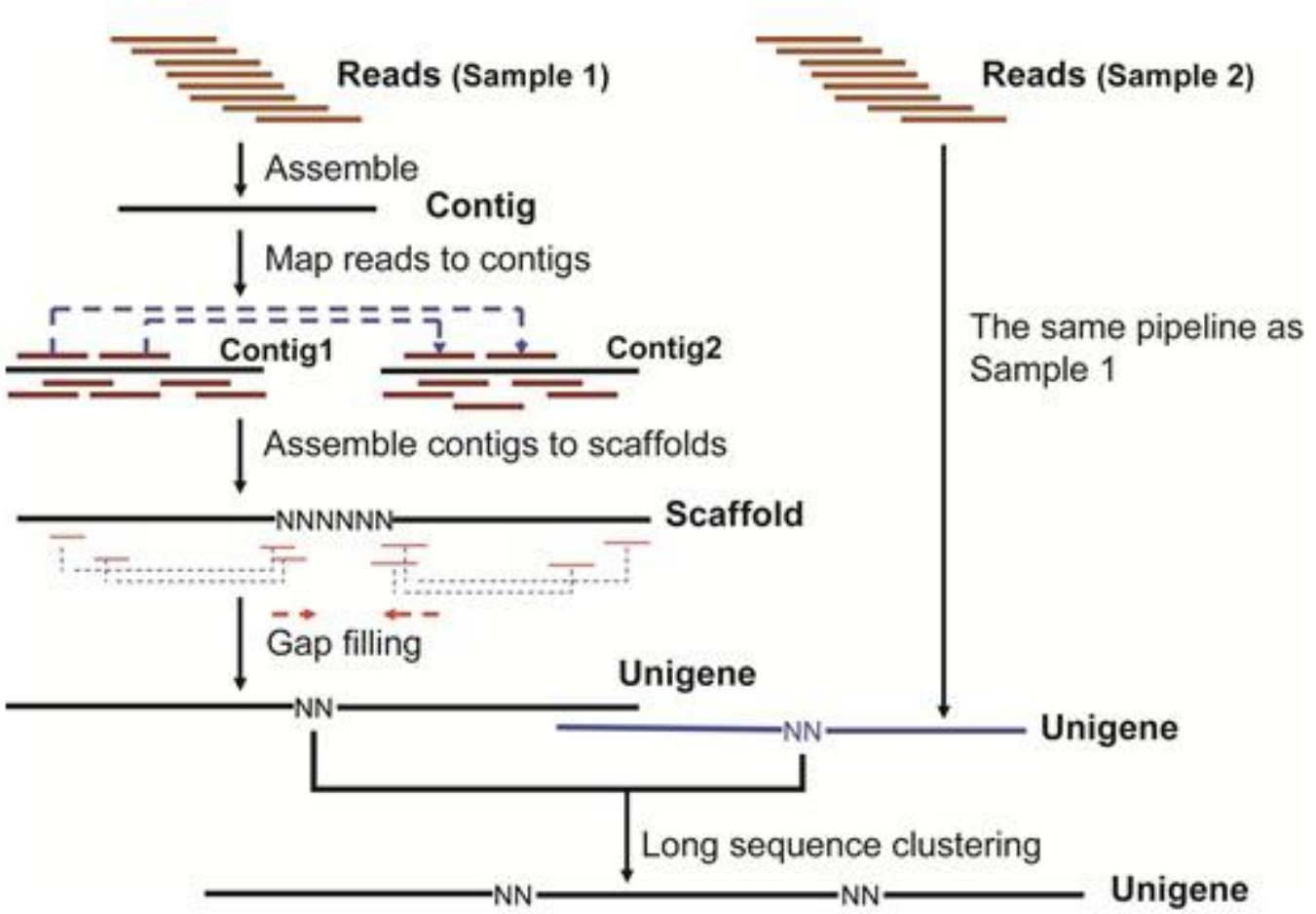


A case study of tobacco (T3)

Samples	Total Reads	Total Nucleotides (nt)	Q20 percentage	N percentage	GC percentage *
T3	27,022,226	2,432,000,340	96.15%	0.00%	44.05%

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

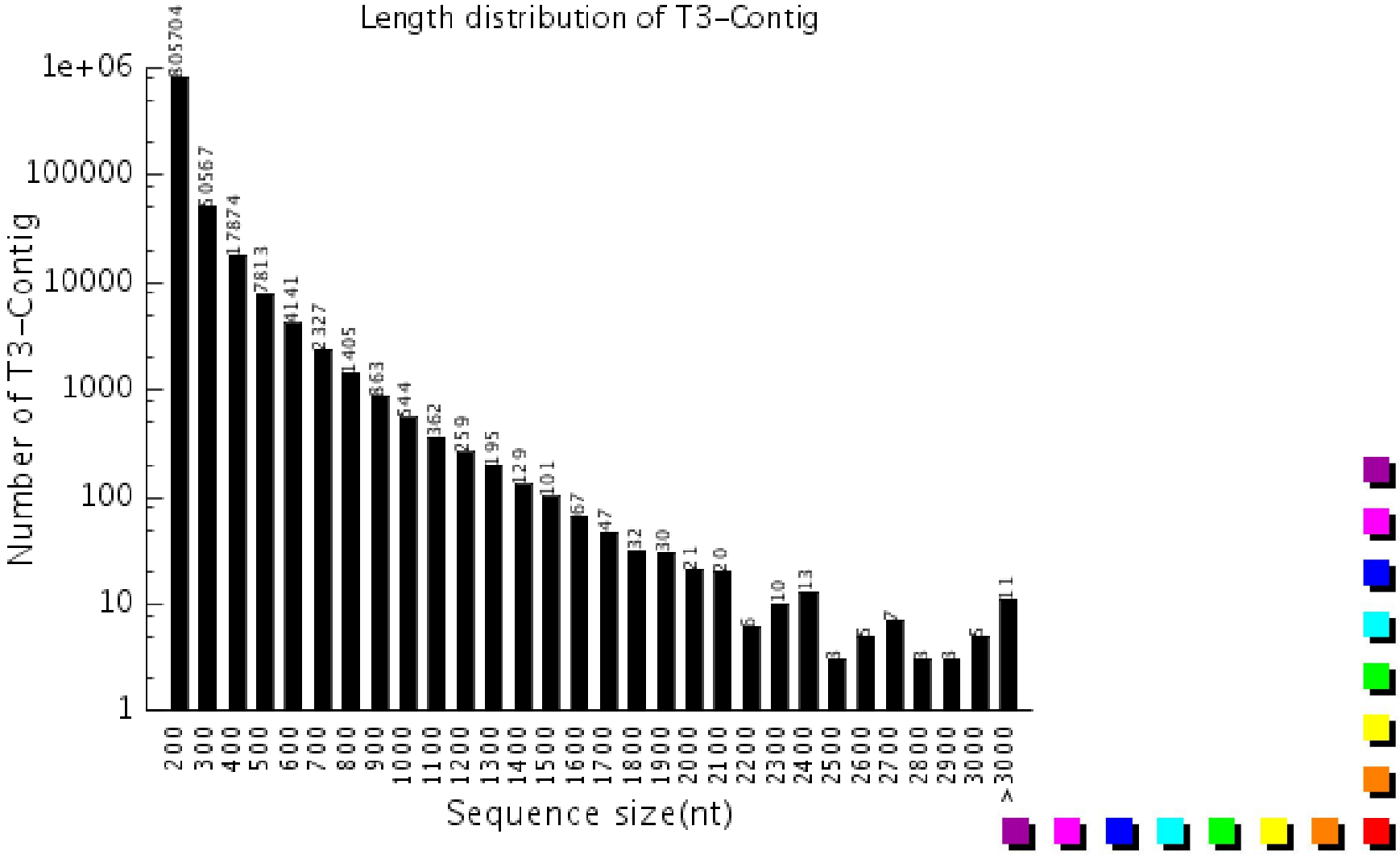
Unigene assembly



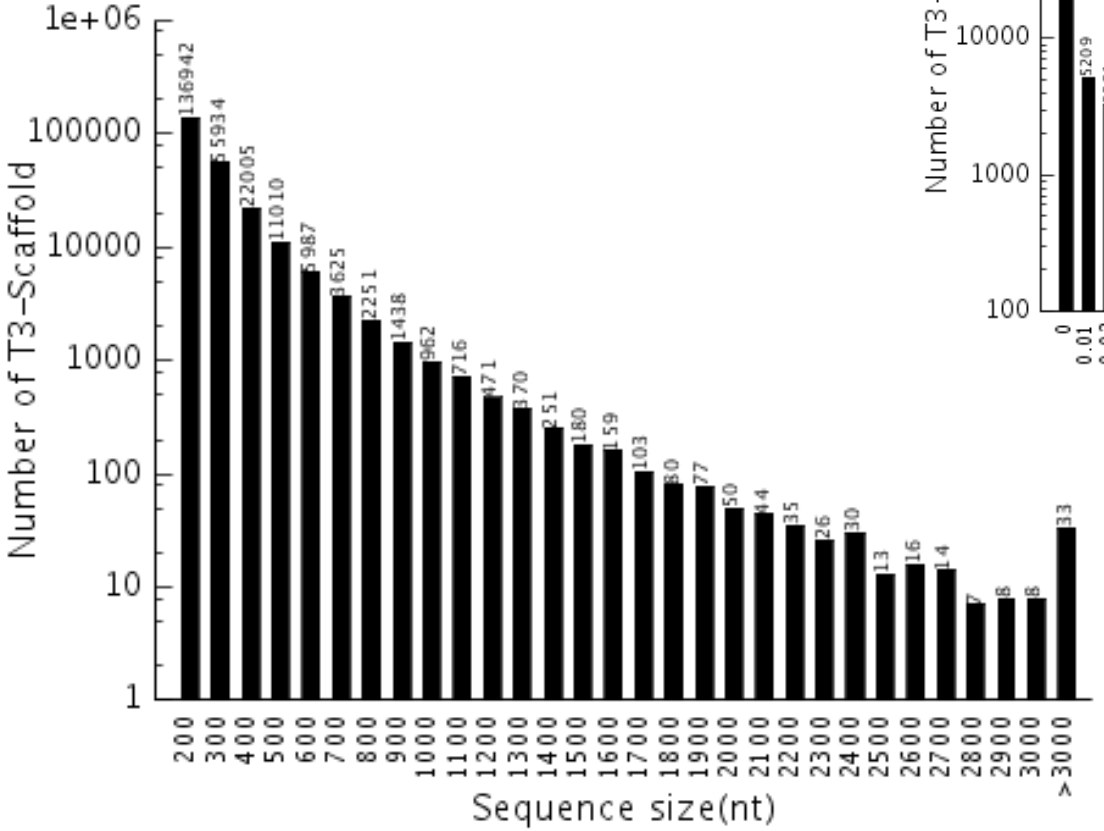
Challenges of transcriptome assembly

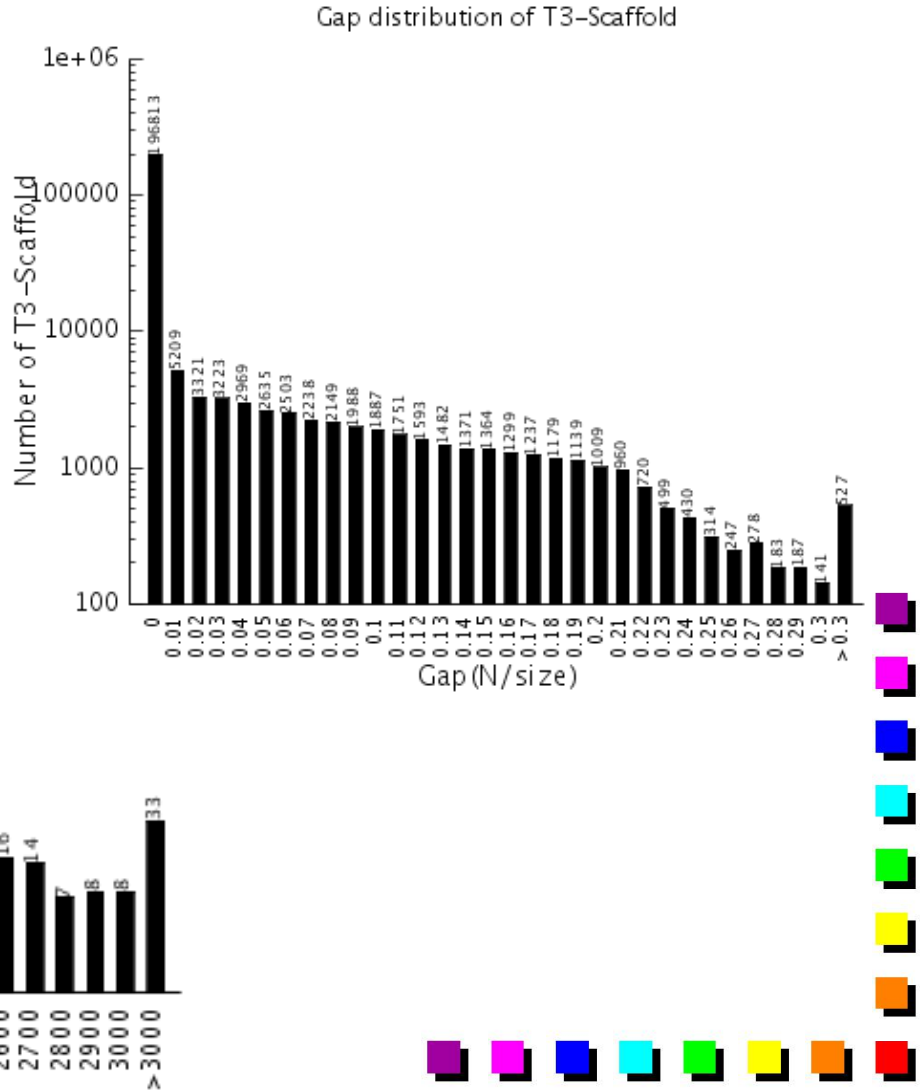
- *de-novo* assembler: ABySS, Velvet, SOAPdenovo, Oases, Trinity, ABySS
- This can be somewhat overcome by having larger sequences obtained from the same sample using other techniques as Sanger sequencing, and using larger reads as a "skeleton" or a "template" to help assemble reads in difficult regions (e.g. regions with repetitive sequences).

Assembly results: Contig

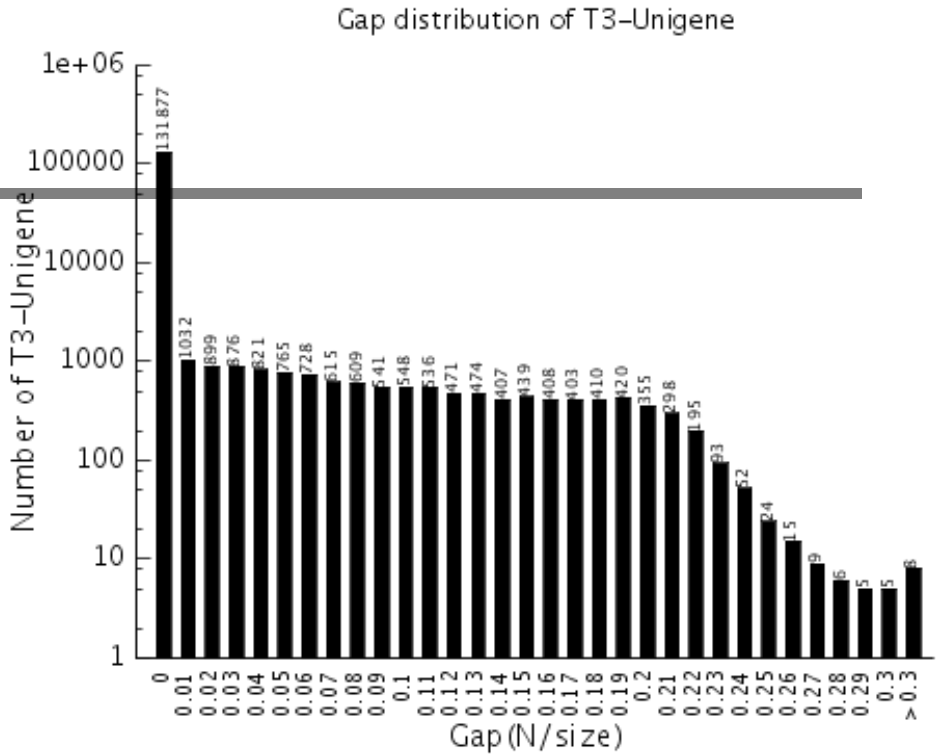


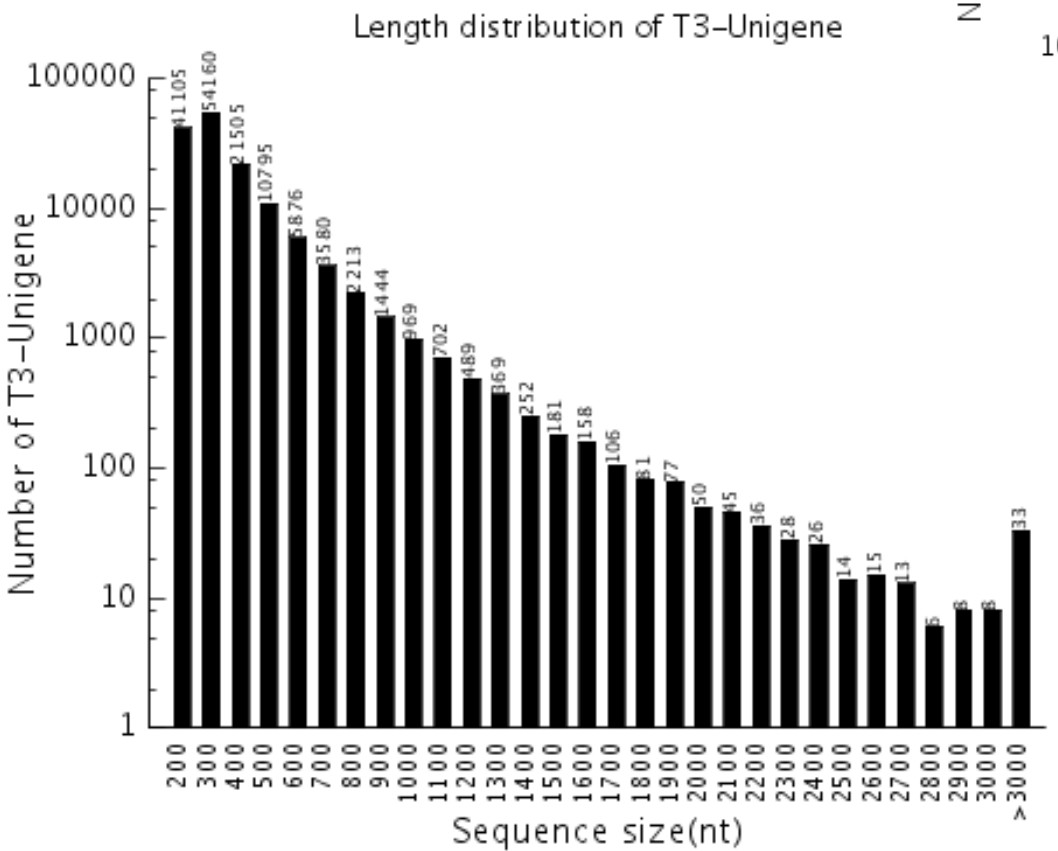
Scaffold

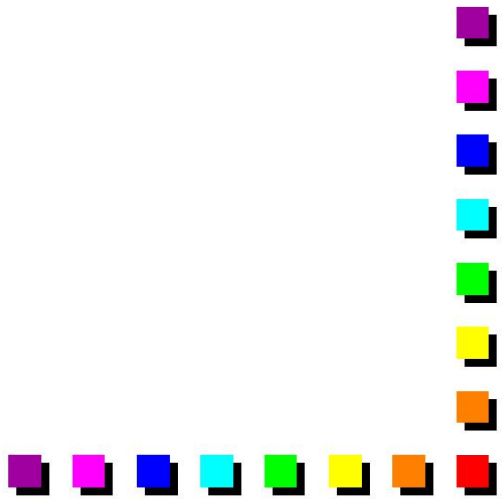

Length distribution of T3-Scaffold
Number of T3-Scaffold
1e+06
100000
10000
1000
100
10
1
136942
55934
22005
11010
5987
3625
2251
1438
962
716
471
370
251
180
159
103
80
77
50
44
35
26
30
13
16
14
7
8
8
33
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
>3000
Sequence size(nt)


Gap distribution of T3-Scaffold
Number of T3-Scaffold
1e+06
100000
10000
1000
100
196813
5209
3321
3223
2969
2635
2503
2238
2149
1988
1887
1751
1593
1482
1371
1364
1299
1237
1179
1139
1009
960
720
499
430
314
247
278
183
187
141
627
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
0.3
>0.3
Gap (N/size)

Unigene

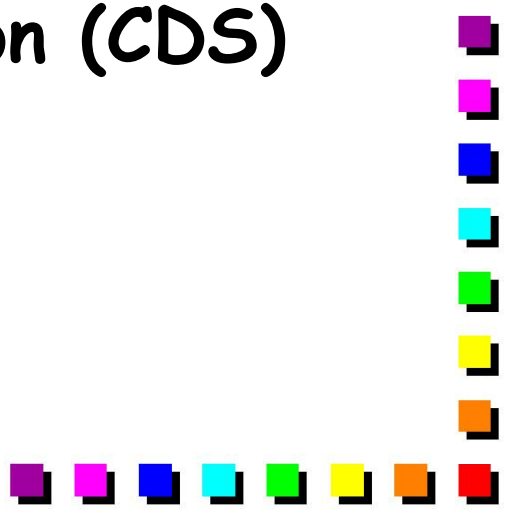

Gap distribution of T3-Unigene
Number of T3-Unigene
1e+06
100000
10000
1000
100
10
1
131877
1032
899
876
821
765
728
615
609
541
548
536
471
474
407
439
408
403
410
420
355
298
195
93
62
24
15
9
6
5
5
8
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
0.3
>0.3
Gap (N/size)


Length distribution of T3-Unigene
Number of T3-Unigene
100000
10000
1000
100
10
1
41105
54160
21505
10795
6876
3580
2213
1444
969
702
489
369
252
181
158
106
81
77
50
45
36
28
26
14
15
13
6
8
8
33
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
>3000
Sequence size(nt)

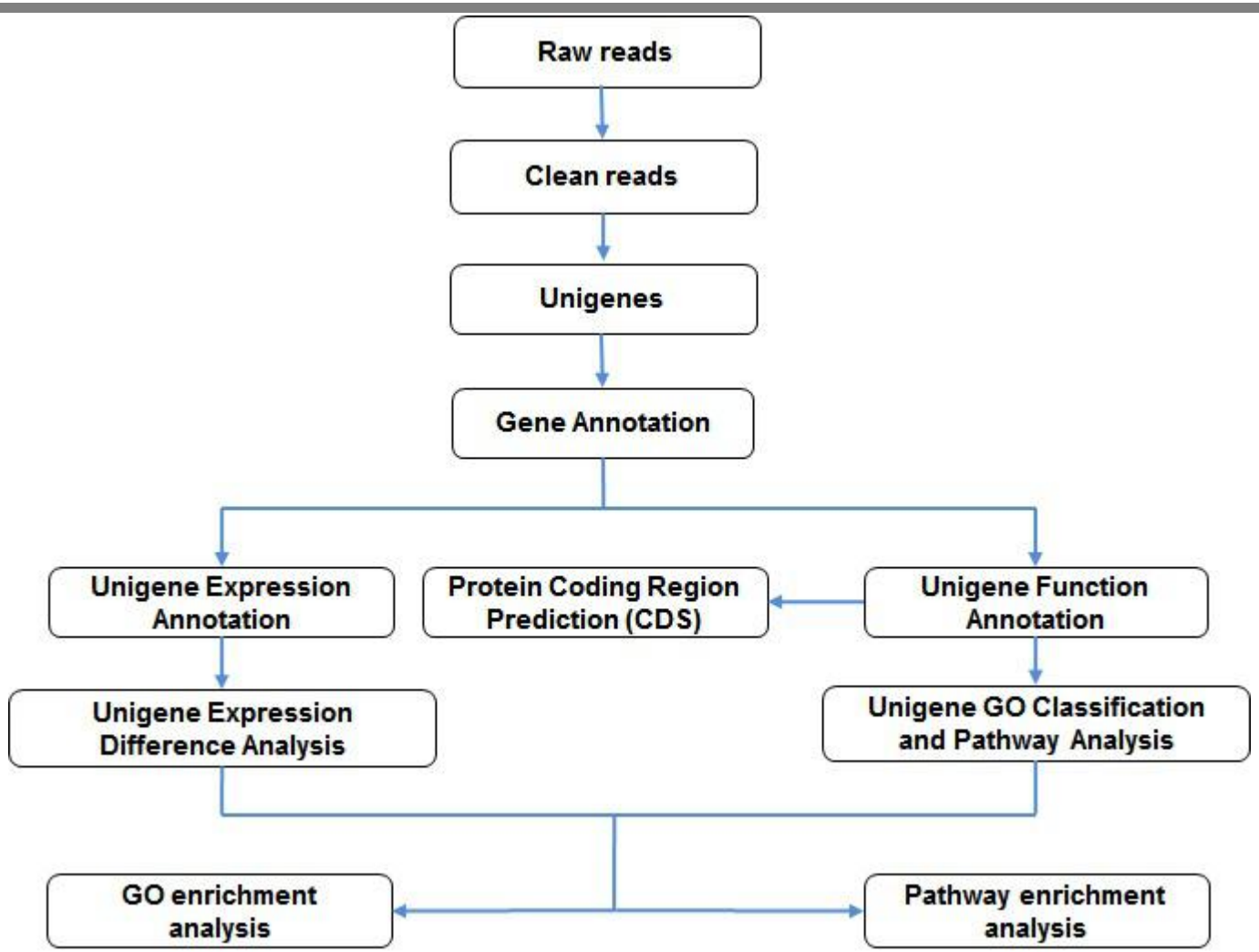


Annotation

- **Assembly Results**
  - **Statistics of assembly quality**
- **Unigene Function Annotation**
  - protein functional annotaiton, Pathway annotation, COG functional annotation and Gene Ontology (GO) functional annotation.
- **Protein Coding Region Prediction (CDS)**
  - **CDS nucleotide sequence**

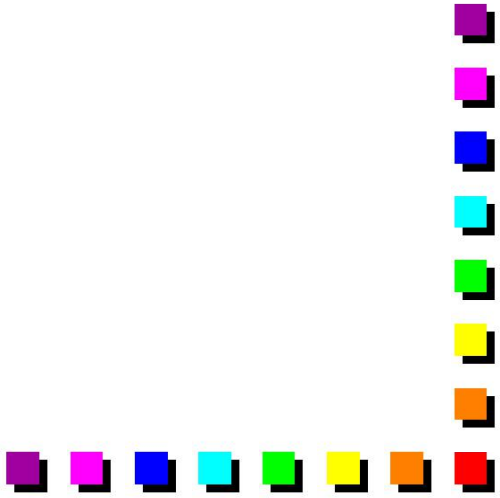


Annotation pipeline



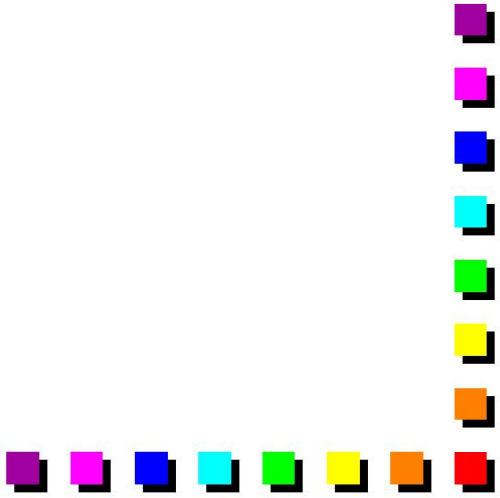
Unigene Function Annotation

- protein functional annotation
- COG functional annotation
- Gene Ontology (GO) functional annotation
- KEGG pathway annotation



Protein functional annotation

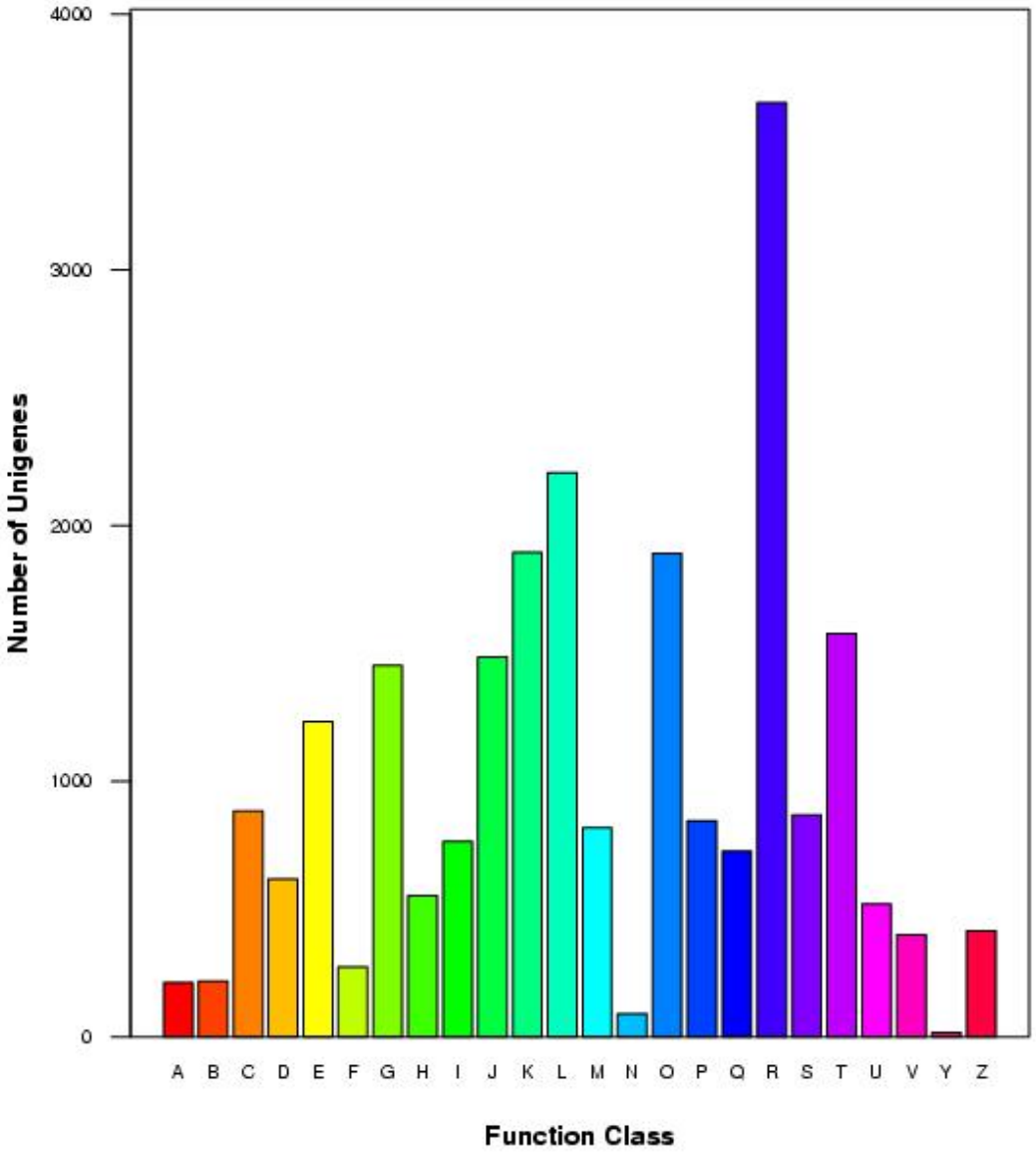
- Unigene sequences are firstly aligned by blastx to protein databases like nr, Swiss-Prot, KEGG and COG (e-value<0.00001), retrieving proteins with the highest sequence similarity with the given Unigenes along with their protein functional annotations.



COG classification

- COG is a database where orthologous gene products were classified. Every protein in COG is assumed to be evolved from an ancestor protein, and the whole database is built on coding proteins with complete genome as well as system evolution relationships of bacteria, algae and eukaryotic creatures. Unigenes are aligned to COG database to predict and classify possible functions of Unigenes.

COG Function Classification of T3-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

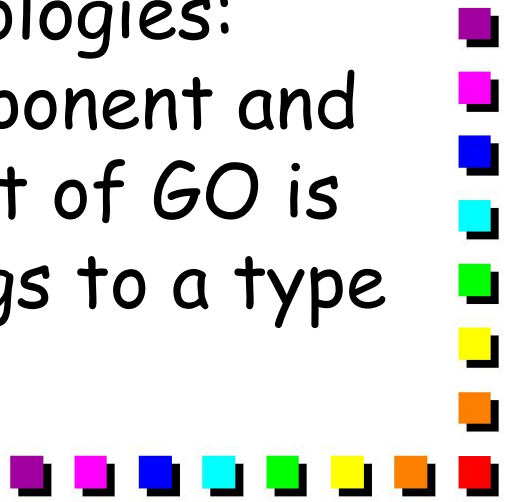
V: Defense mechanisms

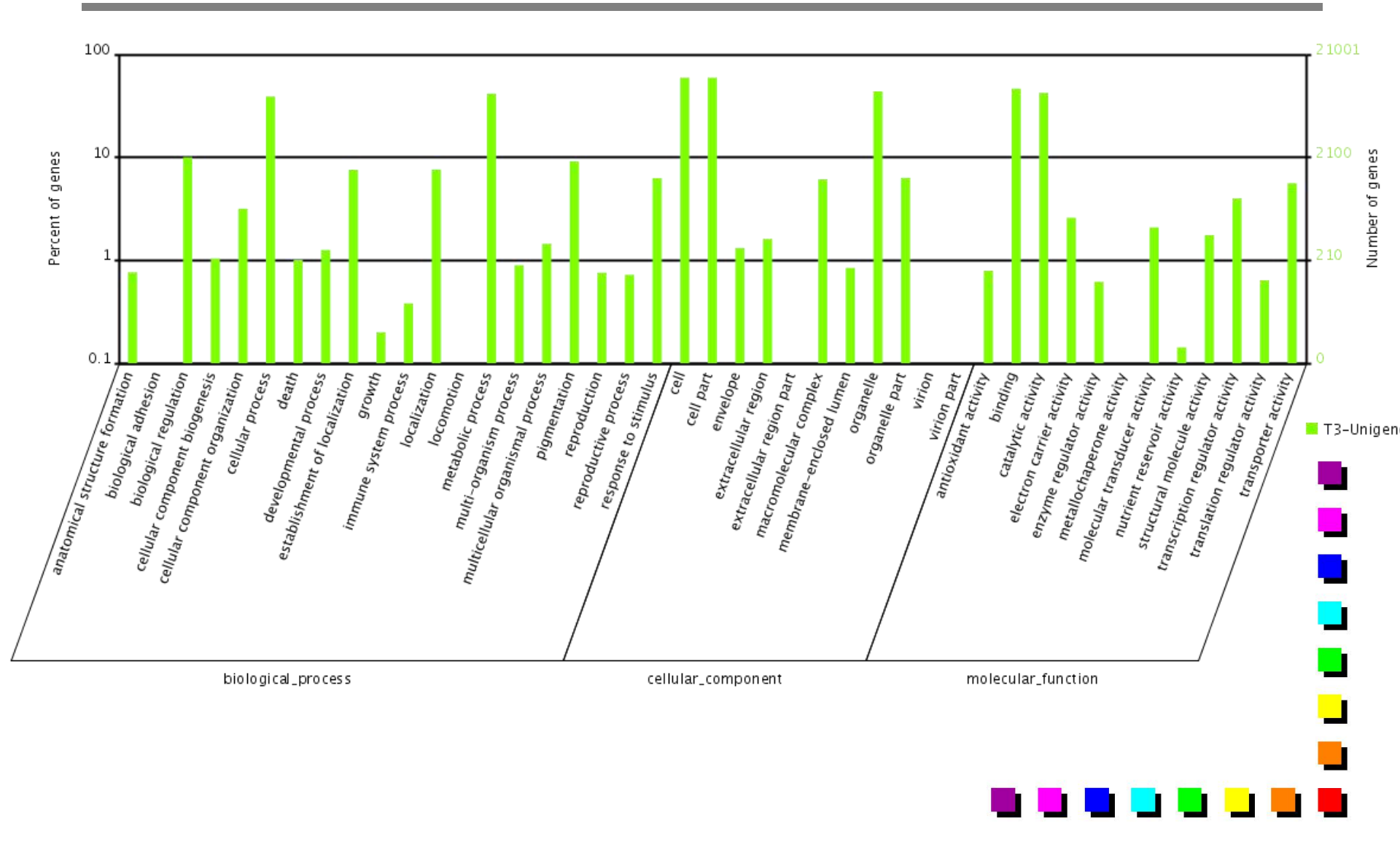
Y: Nuclear structure

Z: Cytoskeleton

GO classification

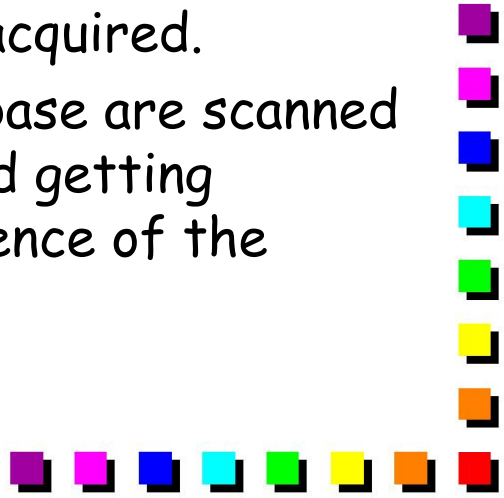
- Gene Ontology (GO) is an international standardized gene functional classification system which offers a dynamic-updated controlled vocabulary and a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component and biological process. The basic unit of GO is GO-term. Every GO-term belongs to a type of ontology.



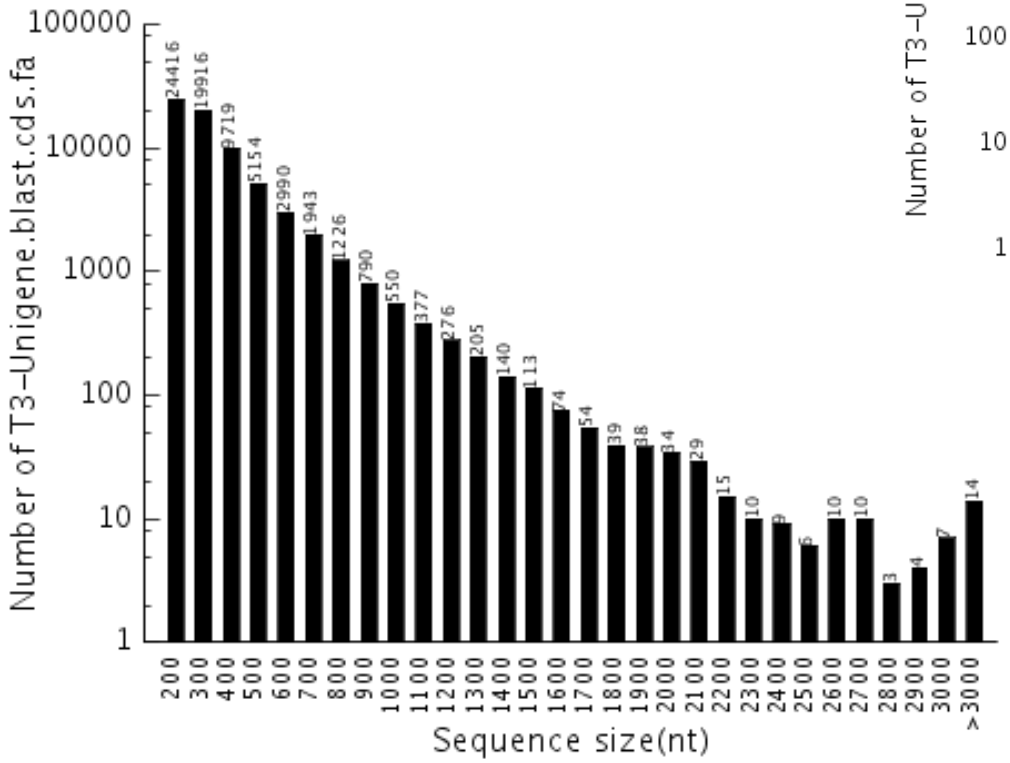

100
10
1
0.1
Percent of genes
21001
2100
210
0
Number of genes
anatomical structure formation
biological adhesion
biological regulation
cellular component biogenesis
cellular component organization
cellular process
death
developmental process
establishment of localization
growth
immune system process
localization
locomotion
metabolic process
multi-organism process
multicellular organismal process
pigmentation
reproduction
reproductive process
response to stimulus
cell
cell part
envelope
extracellular region
extracellular region part
macromolecular complex
membrane-enclosed lumen
organelle
organelle part
virion
virion part
antioxidant activity
binding
catalytic activity
electron carrier activity
enzyme regulator activity
metallochaperone activity
molecular transducer activity
nutrient reservoir activity
structural molecule activity
transcription regulator activity
translation regulator activity
transporter activity
biological_process
cellular_component
molecular_function
T3-Unigen

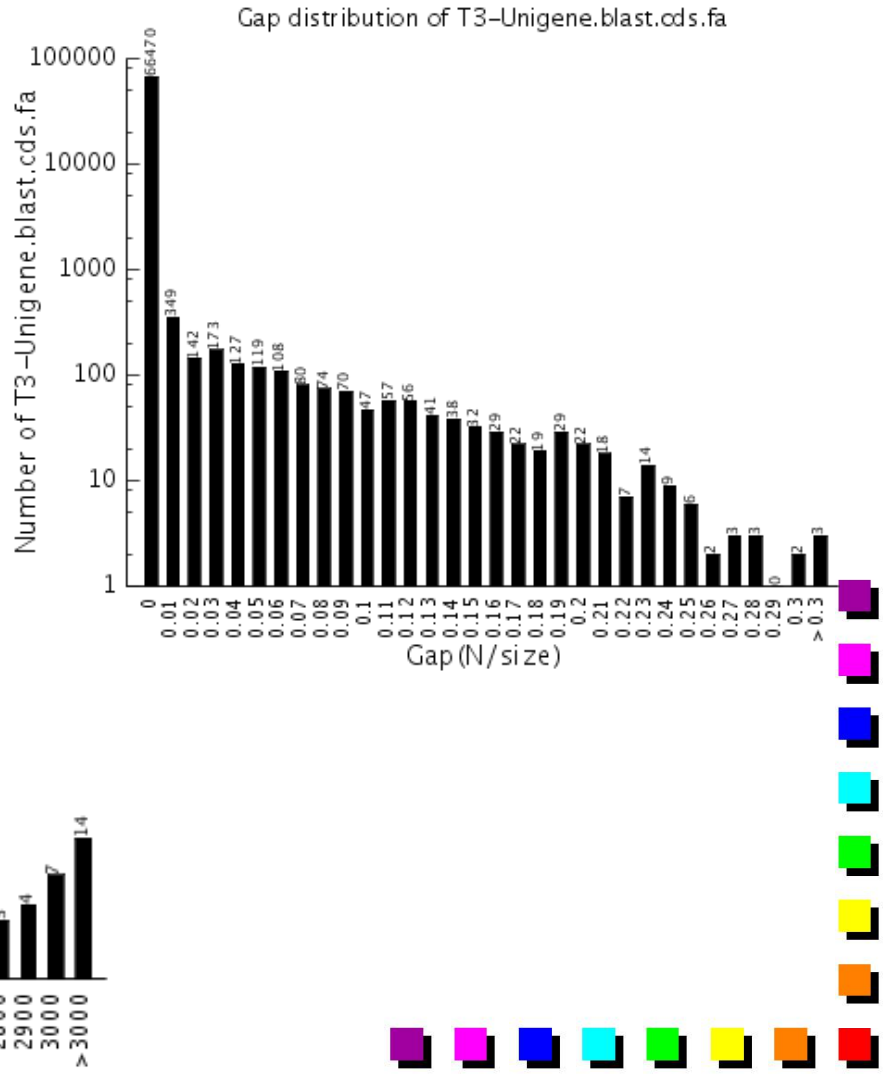
CDS Prediction: Pipeline

- Unigenes are firstly aligned by blastx (evalue<0.00001) to protein databases in the priority order of nr, Swiss-Prot, KEGG and COG. Unigenes aligned to databases with higher priority will not enter the next circle. The alignments end when all circles are finished. Proteins with highest ranks in blast results are taken to decide the coding region sequences of Unigenes, then the coding region sequences are translated into amino sequences with the standard codon table. So both the nucleotide sequences (5' - 3') and amino sequences of the Unigene coding region are acquired.
- Unigenes that cannot be aligned to any database are scanned by ESTScan (Iseli, Jongeneel *et al*. 1999) and getting nucleotide sequence (5' - 3') and amino sequence of the coding regions

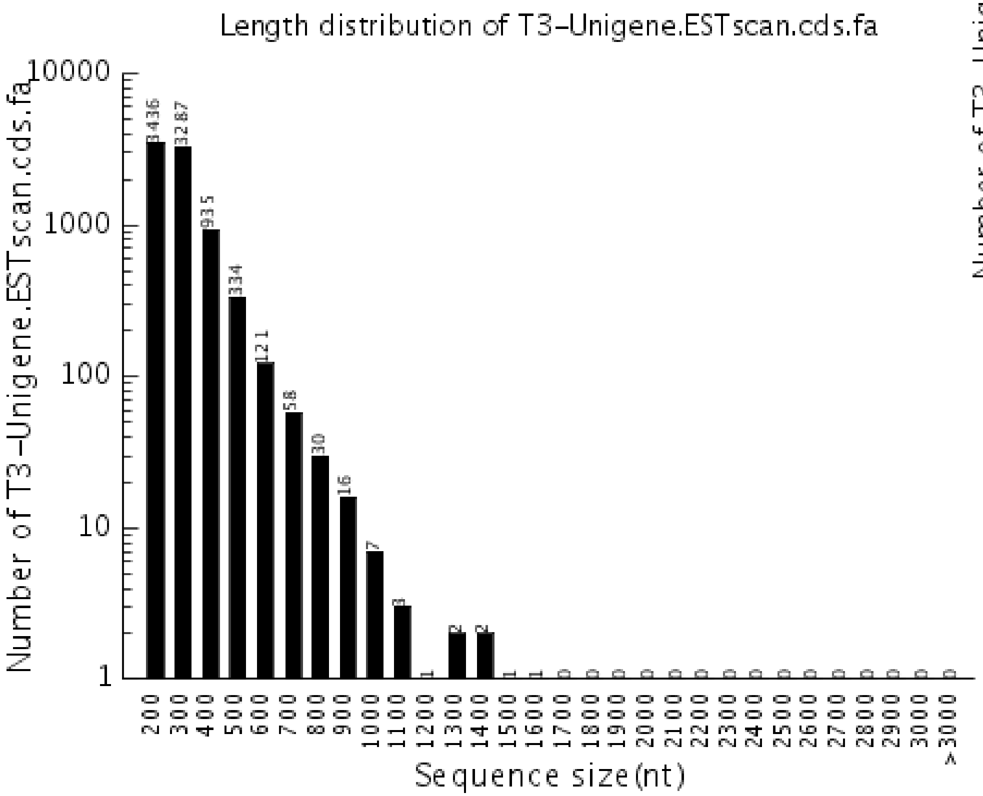


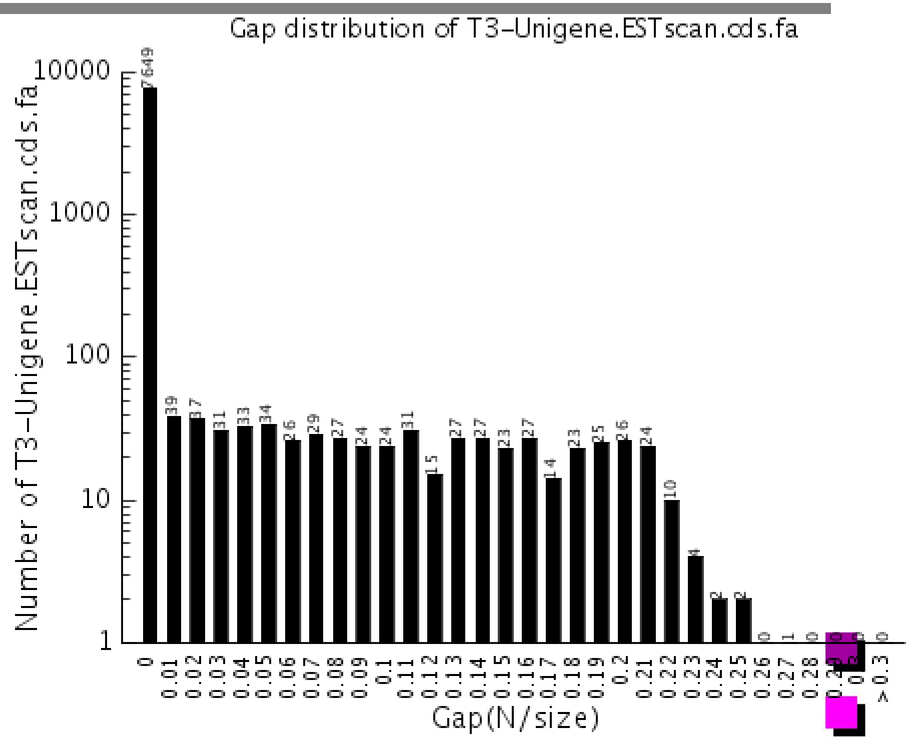
CDS results


Length distribution of T3-Unigene.blast.cds.fa
Number of T3-Unigene.blast.cds.fa
100000
10000
1000
100
10
1
24416
19916
9719
5154
2990
1943
1226
790
550
377
276
205
140
113
74
54
39
38
34
29
15
10
9
6
10
10
3
4
7
14
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
>3000
Sequence size(nt)


Gap distribution of T3-Unigene.blast.cds.fa
Number of T3-Unigene.blast.cds.fa
100000
10000
1000
100
10
1
66470
349
142
173
127
119
108
80
74
70
47
57
56
41
38
32
29
22
19
29
22
18
7
14
9
6
2
3
3
0
2
3
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
0.3
>0.3
Gap(N/size)

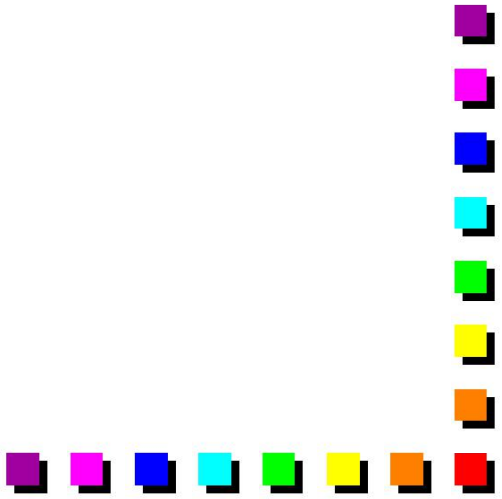
ESTScan results





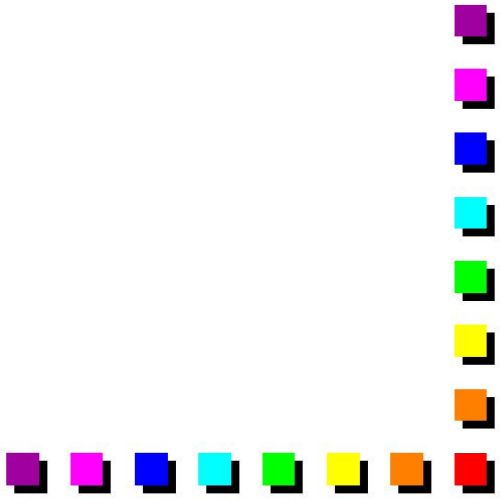
2. Applying RNA-Seq

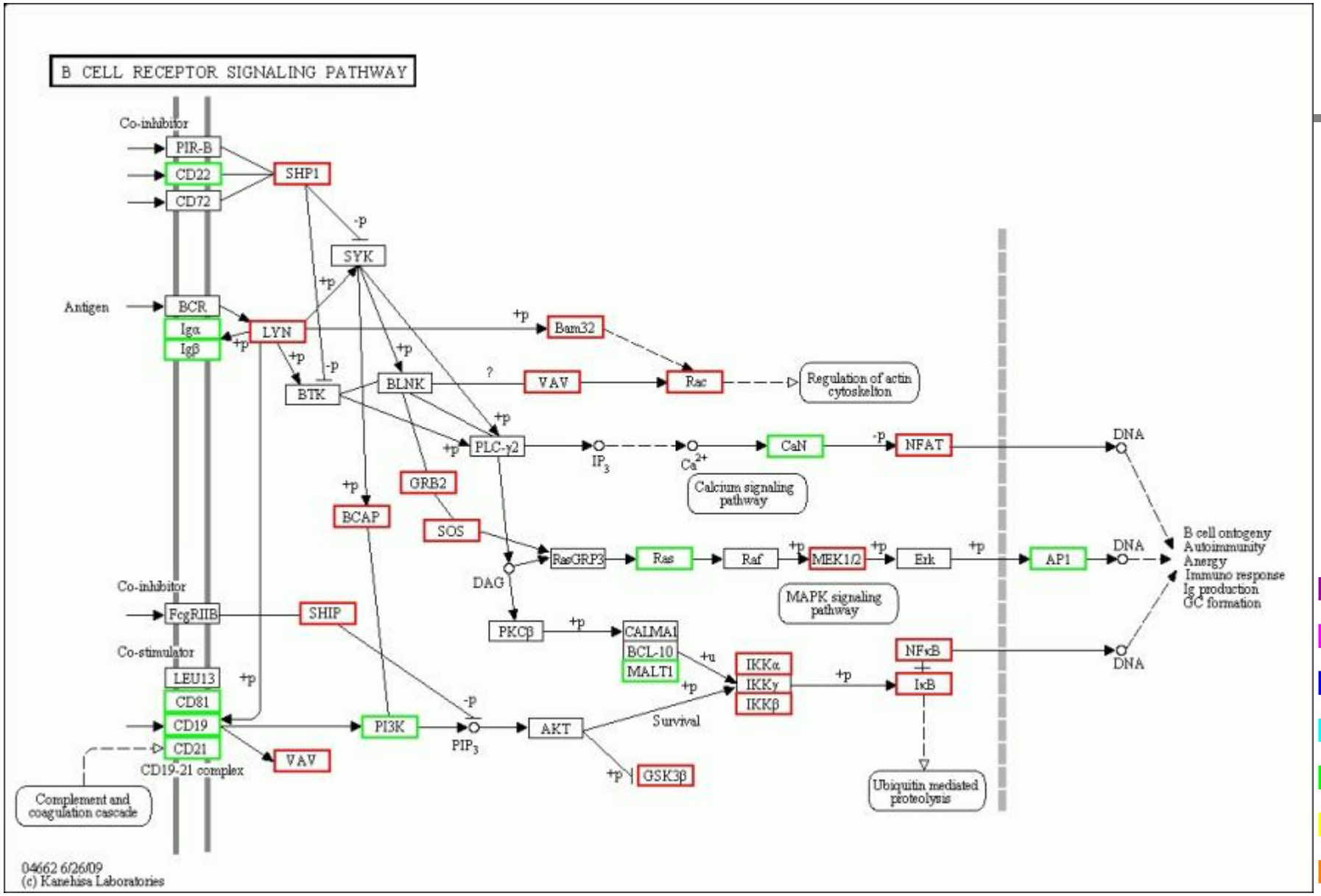
- Gene finding
- Gene expression; Single nucleotide variation discovery; Post-transcriptional SNVs; Fusion gene detection; alternative splice etc.



Metabolic Pathway Analysis: KEGG

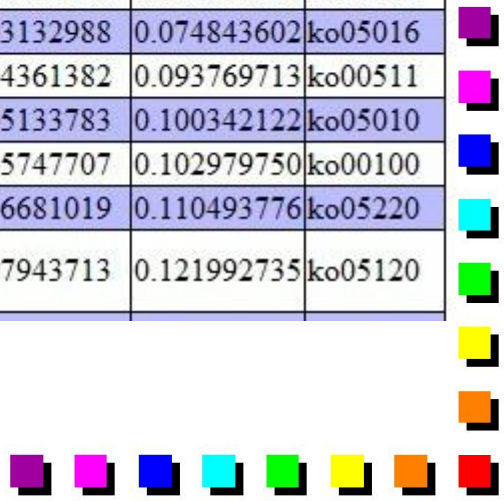
- KEGG database contains systematic analysis of inner-cell metabolic pathways and functions of gene products. It helps studying complicated biological behaviors of genes. With KEGG annotation we can get Pathway annotation of Unigenes.




B CELL RECEPTOR SIGNALING PATHWAY
Co-inhibitor
PIR-B
CD22
CD72
SHP1
-p
SYK
+p
Antigen
BCR
Igα
Igβ
LYN
+p
Bam32
+p
BTK
-p
BLNK
?
VAV
Rac
Regulation of actin cytoskelton
+p
PLC-γ2
IP3
Ca2+
CaN
-p
NFAT
DNA
Calcium signaling pathway
+p
GRB2
BCAP
SOS
RasGRP3
Ras
Raf
+p
MEK1/2
+p
Erk
+p
AP1
DNA
B cell ontogeny
Autoimmunity
Anergy
Immuno response
Ig production
GC formation
DAG
MAPK signaling pathway
Co-inhibitor
FcgRIIB
SHIP
PKCβ
+p
CALMA1
BCL-10
MALT1
+u
IKKα
IKKγ
IKKβ
+p
+p
NFκB
IκB
DNA
Co-stimulator
LEU13
CD81
CD19
CD21
+p
-p
PI3K
PIP3
AKT
Survival
+p
GSK3β
VAV
CD19-21 complex
Complement and coagulation cascade
Ubiquitin mediated proteolysis
04662 6/26/09
(c) Kanehisa Laboratories

Pathway Functional Enrichment Analysis

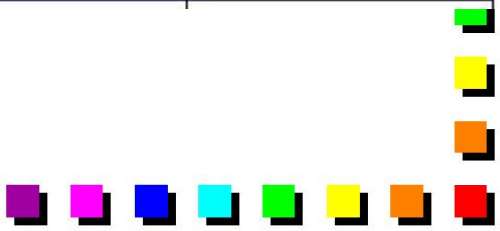
#	Pathway	DEGs with pathway annotation (2085)	All genes with pathway annotation (8986)	Pvalue	Qvalue	Pathway ID
1	Metabolic pathways	307 (14.72%)	1081 (12.03%)	1.354119e-05	0.002911356	ko01100
2	Proteasome	23 (1.1%)	48 (0.53%)	0.0001482570	0.015937627	ko03050
3	B cell receptor signaling pathway	29 (1.39%)	70 (0.78%)	0.0005085341	0.036444944	ko04662
4	Apoptosis	34 (1.63%)	89 (0.99%)	0.001018471	0.045737882	ko04210
5	Hematopoietic cell lineage	31 (1.49%)	80 (0.89%)	0.001271905	0.045737882	ko04640
6	Primary immunodeficiency	16 (0.77%)	33 (0.37%)	0.001276406	0.045737882	ko05340
7	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	13 (0.62%)	25 (0.28%)	0.001618825	0.049721054	ko00563
8	N-Glycan biosynthesis	18 (0.86%)	40 (0.45%)	0.001901140	0.051093137	ko00510
9	Huntington's disease	60 (2.88%)	187 (2.08%)	0.003132988	0.074843602	ko05016
10	Other glycan degradation	9 (0.43%)	16 (0.18%)	0.004361382	0.093769713	ko00511
11	Alzheimer's disease	56 (2.69%)	176 (1.96%)	0.005133783	0.100342122	ko05010
12	Biosynthesis of steroids	13 (0.62%)	28 (0.31%)	0.005747707	0.102979750	ko00100
13	Chronic myeloid leukemia	27 (1.29%)	74 (0.82%)	0.006681019	0.110493776	ko05220
14	Epithelial cell signaling in Helicobacter pylori infection	25 (1.2%)	68 (0.76%)	0.007943713	0.121992735	ko05120



Gene Ontology Functional Enrichment Analysis

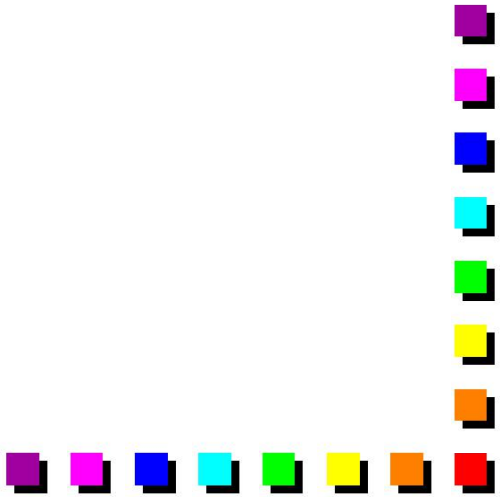
Terms from the Process Ontology with p-value as good or better than 0.05

Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value	Expression Profile
immune response view genes	82 out of 807 genes, 10.2%	663 out of 13525 genes, 4.9%	2.74e-07	View Result
immune system process view genes	100 out of 807 genes, 12.4%	921 out of 13525 genes, 6.8%	3.77e-06	View Result
response to virus view genes	21 out of 807 genes, 2.6%	105 out of 13525 genes, 0.8%	0.00138	View Result
regulation of apoptosis view genes	63 out of 807 genes, 7.8%	583 out of 13525 genes, 4.3%	0.00508	View Result
regulation of programmed cell death view genes	63 out of 807 genes, 7.8%	592 out of 13525 genes, 4.4%	0.00840	View Result
&regulation of cell death view genes	63 out of 807 genes, 7.8%	593 out of 13525 genes, 4.4%	0.00888	View Result
regulation of cell death view genes	63 out of 807 genes, 7.8%	593 out of 13525 genes, 4.4%	0.00888	View Result



Profile expression levels

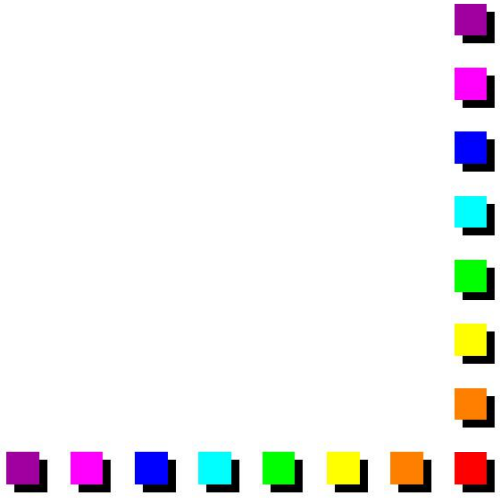
- Statistical issues



Differentially Expressed Genes (DEG)

- "The significance of digital gene expression profiles" (Audic et al, 1997, Genome Research)
- Denote the number of unambiguous clean tag from gene A as x, as every genes expression occupies only a small part of the library, the p(x) is in the Poisson distribution.

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$



- The total clean tag number of the sample 1 is N1, and total clean tag number of sample 2 is N2; gene A holds x tags in sample1 and y tags in sample2. The probability of gene A expressed equally between two samples can be calculated with the following formula:

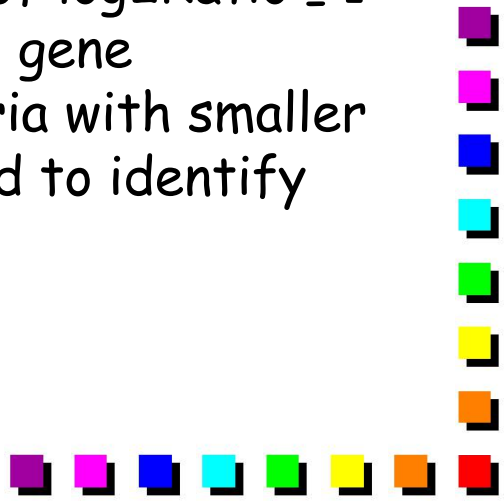
$$2\sum_{i=0}^{i=y} p(i \mid x)$$

$$\text{Or } 2\times(1-\sum_{i=0}^{i=y} p(i \mid x)) \quad (\text{if} \sum_{i=0}^{i=y} p(i \mid x) > 0.5)$$

$$p(y \mid x) = (\frac{N_2}{N_1})^y \frac{(x+y)!}{x!y!(1+\frac{N_2}{N_1})^{(x+y+1)}}$$

False Discovery Rate (FDR)

- FDR is a method to determine the threshold of P Value in multiple test and analysis through manipulating the FDR value. Assume that we have picked out R differentially expressed genes in which S genes are really show differential expression and the other V genes are false positive. If we decide that the error ratio "Q = V/R" must stay below a cutoff (e.g. 5%), we should preset the FDR to a number no larger than 0.05.
- We use "FDR ≤ 0.001 and the absolute value of log2Ratio ≥ 1" as the threshold to judge the significance of gene expression difference. More stringent criteria with smaller FDR and bigger fold-change value can be used to identify DEGs.



RPKM

- The calculation of unigene expression uses RPKM method (Reads Per kb per Million reads) (Mortazavi et al, 2008)

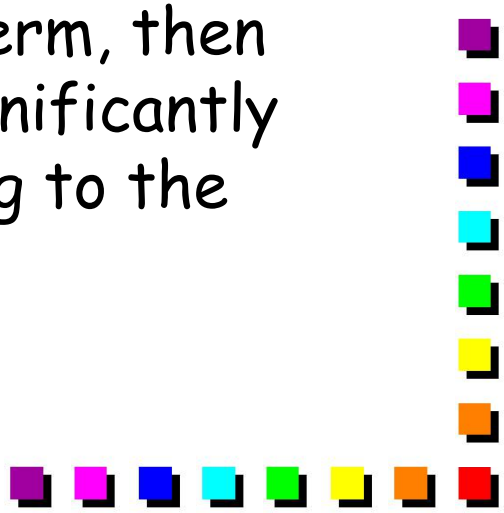
$$RPKM = \frac{10^6 C}{NL/10^3}$$

  - C to be number of reads that uniquely aligned to Unigene A, N to be total number of reads that uniquely aligned to all Unigenes, and L to be number of bases on Unigene A.

- The RPKM method is able to eliminate the influence of different gene length and sequencing level on the calculation of gene expression. Therefore the calculated gene expression can be directly used for comparing the difference of gene expression between samples.

Gene Ontology Functional Enrichment Analysis for DEGs

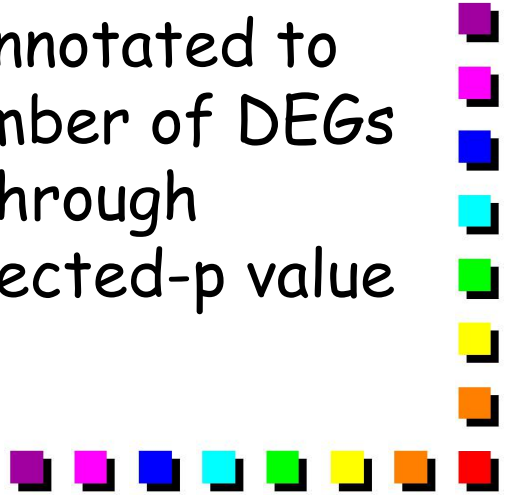
- GO functional enrichment analysis provides GO terms which significantly enrich in DEGs comparing to the genome background, showing which DEGs are connected to wanted biological functions.
- The analysis firstly maps all DEGs to GO terms in the database (http://www.geneontology.org/), calculating gene numbers for every term, then using ultra-geometric test to find significantly enriched GO terms in DEGs comparing to the genome background.



- The calculating formula is

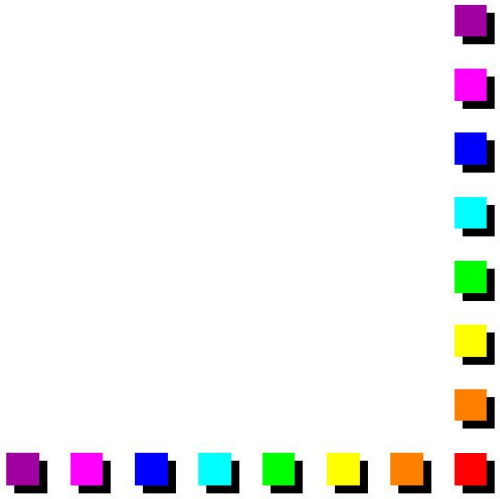
$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

  - where N is the number of all genes with GO annotation; n is the number of DEGs in N; M is the number of all genes that are annotated to the certain GO terms; m is the number of DEGs in M. The calculated p value goes through Bonferroni Correction, taking corrected-p value ≤ 0.05 as a threshold.



Qvalue

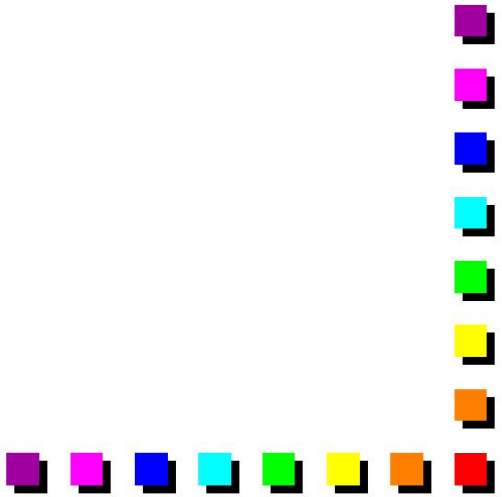
- The q-value is defined to be the FDR analogue of the p-value. The q-value of an individual hypothesis test is the minimum FDR at which the test may be called significant.

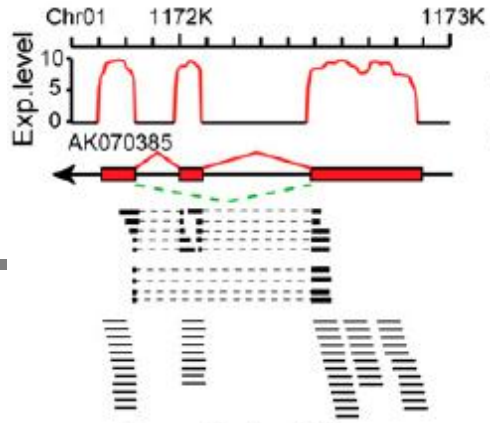


Applying RNA-Seq: others

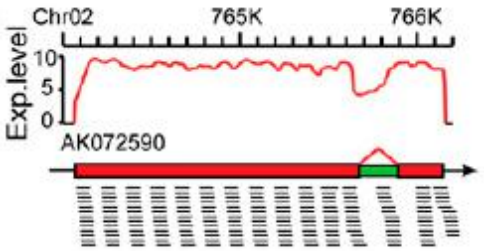
- Wikipedia
  - Gene expression; Microarray approach; Coverage as measure of expression; Single nucleotide variation discovery; Coverage/depth; Germline vs expressed alleles; Post-transcriptional SNVs; Fusion gene detection

- A case of rice: two GR papers
  - Identify alternative splices
  - Identify novel transcripts

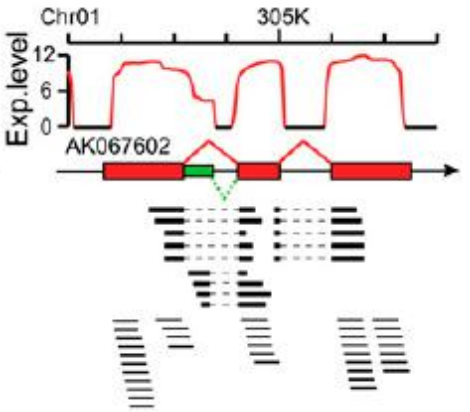




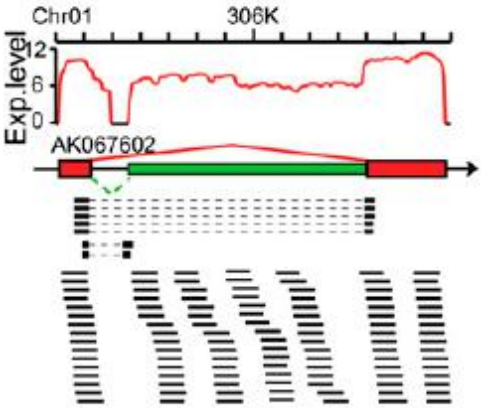
Exon-skipping (ES)



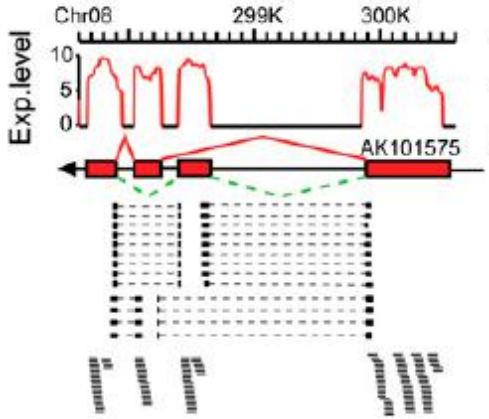
Intron-retention (IR)



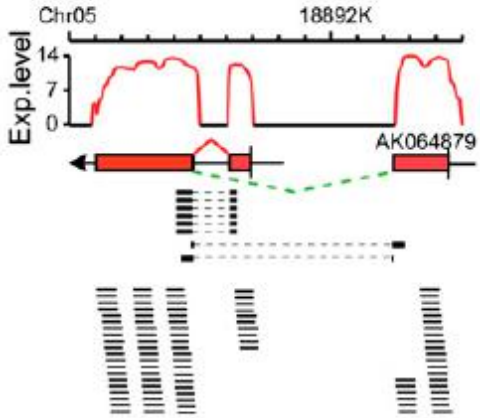
Alternative 5′ splice site (A5SS)



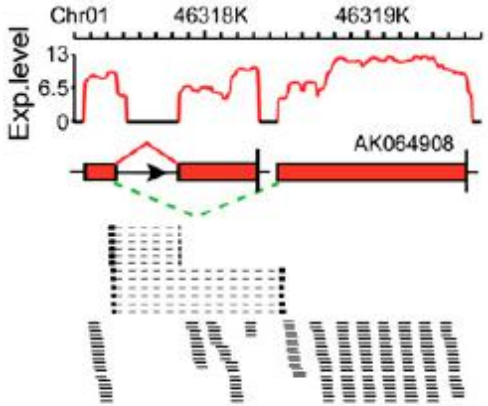
Alternative 3′ splice site (A3SS)



Mutually exclusive exon (MXE)



Alternative First Exon (AFE)



Alternative Last Exon (ALE)

Alternative splicing events in the rice transcriptome

Zhang *et al*. 2010

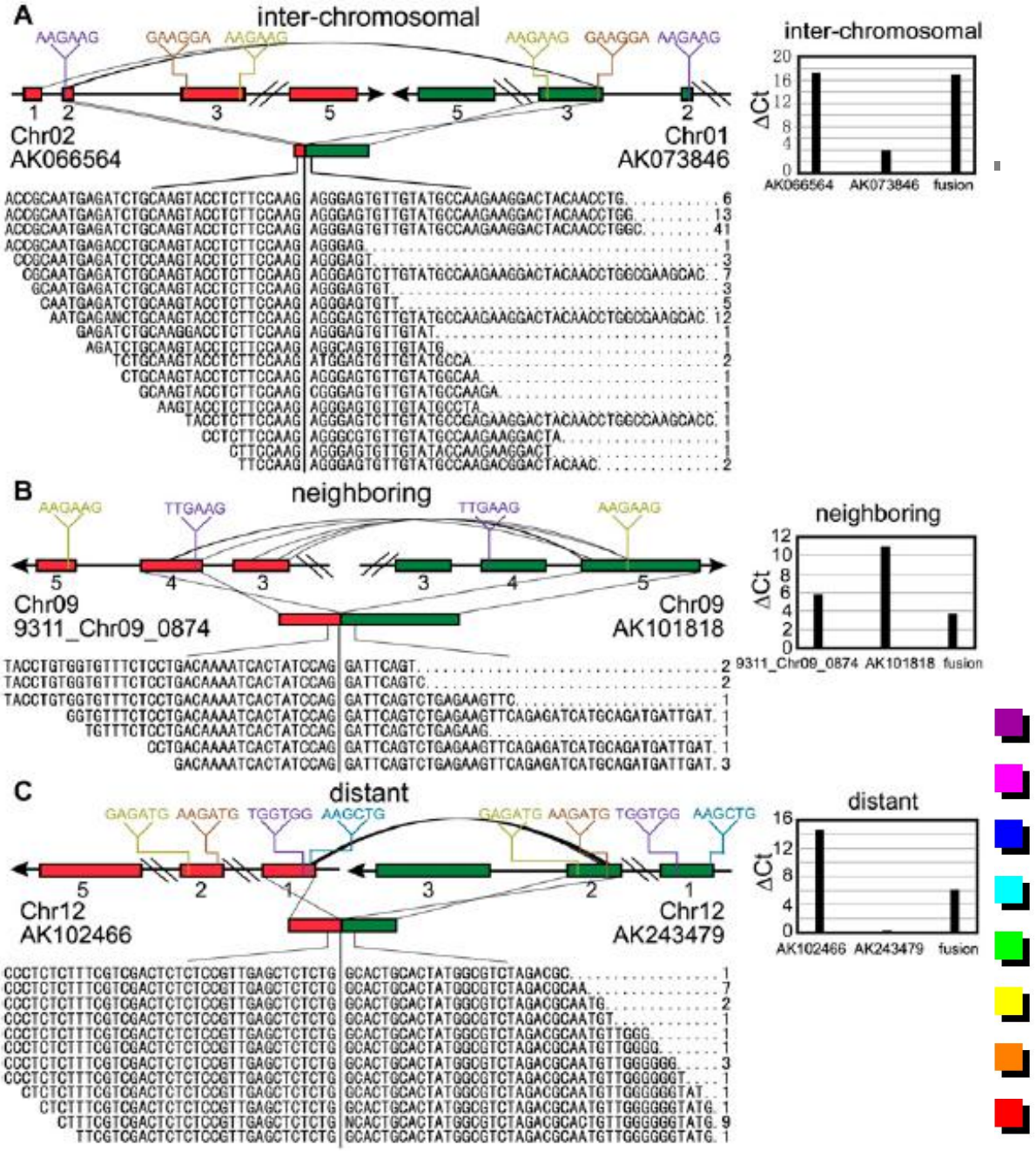
Table 1. The number of each type of AS event and the number of organ-specific AS events

AS events	ES	IR	A5SS	A3SS	MXE	AFE	ALE
Total organ	6048(25.4%)	11,297(47.5%)	1871(7.9%)	3483(14.6%)	567(2.4%)	258(1.1%)	276(1.2%)
Organ specific	3296(26.2%)	5800(46.0%)	1057(8.4%)	1637(13.0%)	306(2.4%)	240(1.9%)	262(2.1%)

ES: Exon-skipping; IR: intron retention; A5SS: alternative 5′ spliced site; A3SS: alternative 3′ spliced site; MXE: mutually exclusion exon; AFE: alternative first exon; ALE: alternative last exon.

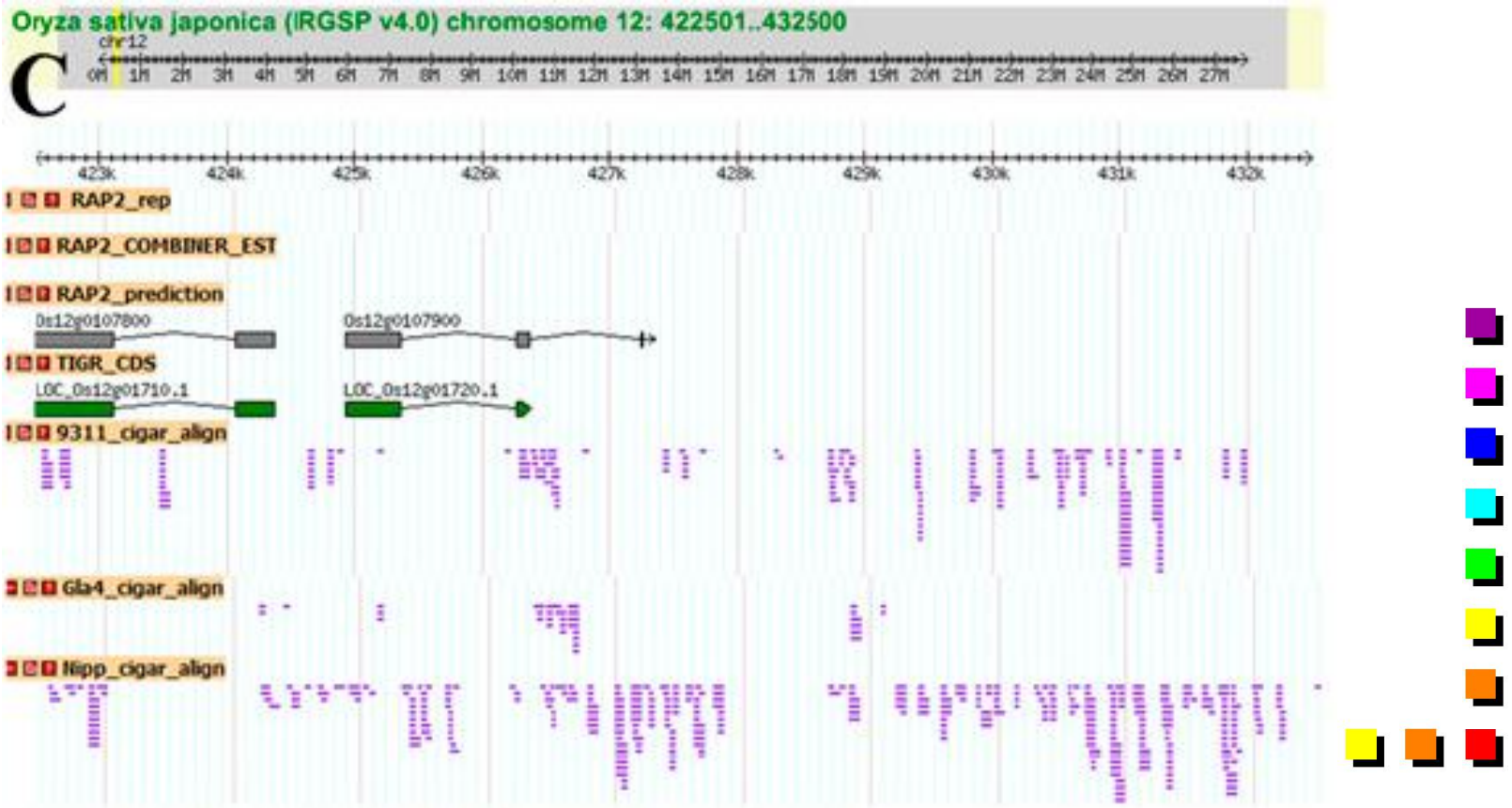
- An analysis of alternative splicing in the rice transcriptome revealed that alternative cis-splicing occurred in 33% of all rice genes. (Zhang *et al.* 2010)
- We found that ~48% of rice genes show alternative splicing patterns. (Lu *et al.* 2010)

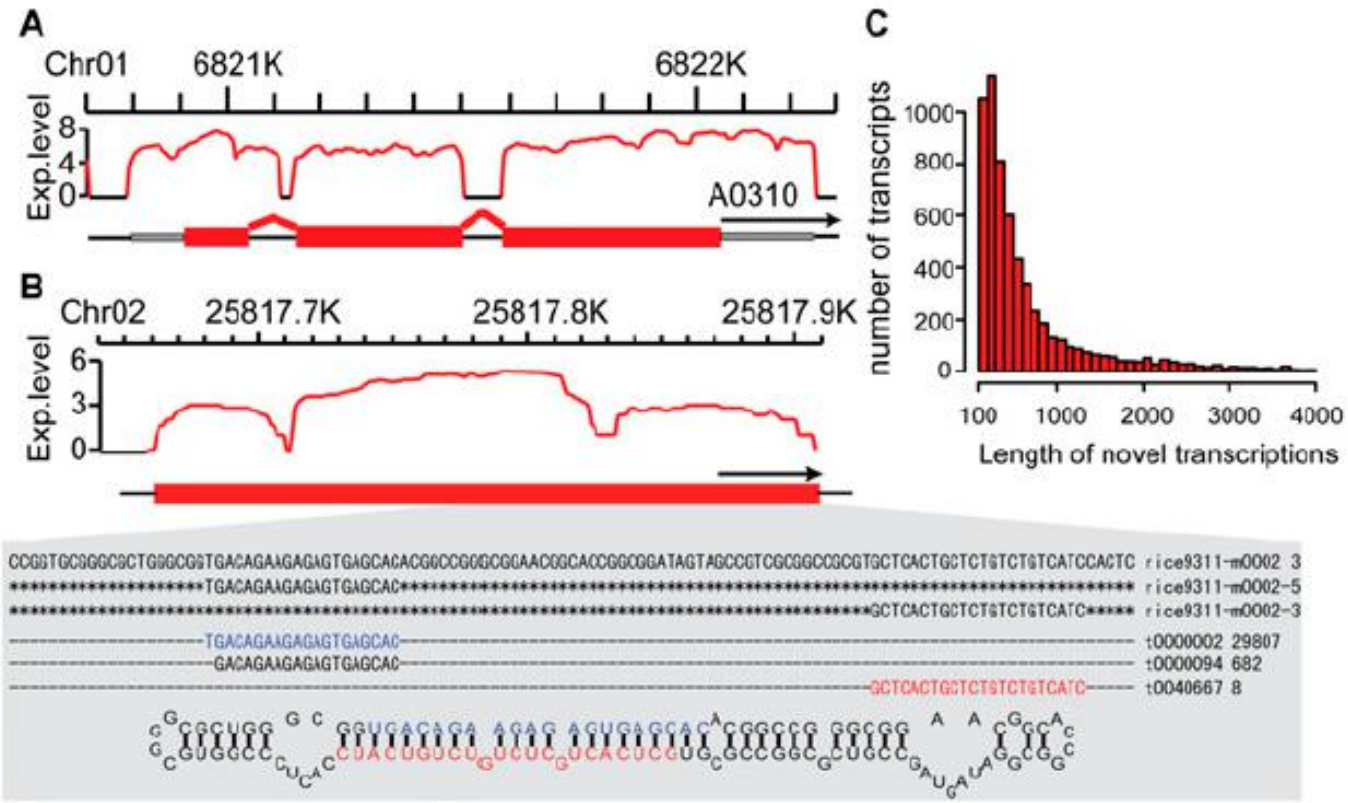
Gene fusion



(Zhang *et al.* 2010)

- We identified 15,708 novel transcriptional active regions (nTARs), of which 51.7% have no homolog to public protein data and >63% are putative single-exon transcripts, which are highly different from protein-coding genes (<20%). (Lu et al. 2010)





(Zhang *et al.* 2010)

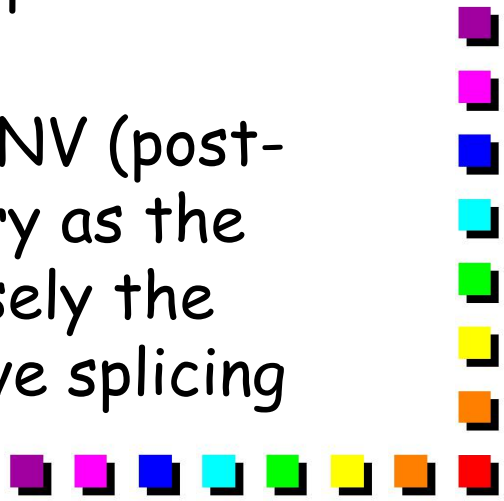
SNP CALLING

- RNA-Seq-based whole genome EST-SNP calling: genetic experimental populations (genetic map; QTL); cultivar populations (population structure; GWAS)
  - Bancroft et al. 2011. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nature Biotech
  - Harper et al. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nature Biotech

(Rapeseed genome is not yet available)

Some considerations

- Same limitations as other RNA expression analysis pipelines.
  - a) Tissue specific
  - b) Time dependent
- Care must be taken when drawing conclusions from the sequencing experiment, as some of the information gathered might not be representative of the individual itself.
- An example of this would be during SNV (post-transcriptional modification) discovery as the mutations discovered are more precisely the mutations being expressed; alternative splicing



How many Gs (data) are enough for transcript assembly?

- 2G
- 4G
- 8G in future?

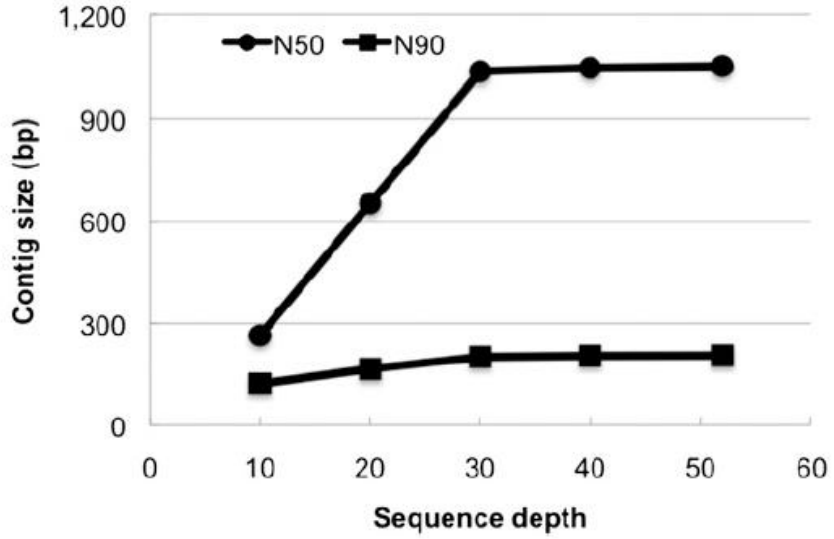


Figure 4. N50 and N90 size of assembled contigs by different sequence depths. We sampled subsets of randomly selected reads from the Asian genome data for de novo assembly of contigs. The same *K*-mer ($K = 25$) size was used for all the assemblies.

(Li et al. 2010, Genome Res)

三、专注功能位点与功能元件

1、高通量技术在功能研究领域大有可为

CAGE: Capture of the methylated cap at 5’ end of RNA, following by high-throughput sequencing of a small adjacent to the 5’ methylated caps. Identification of TSS by CAGE.

RNA-PET: simultaneous capture of RNAs with both 5’ methyl cap and a poly(A) tail, which is indicative of a full-length RNA.

DNase-seq: Chromatin accessibility characterized by Dnase 1 hypersensitive is the hallmark of regulatory DNA regions.

FAIRE: Formaldehyde assisted isolation of regulatory elements.

ChiA-PET: chromatin interaction analysis with paired-end tag sequencing for identification of chromosome-interaction regions. Physical interaction between distinct chromosome regions that can be separated by hundreds of kilobases is thought to be important in the regulation of gene expression.

(The ENCODE Project Consortium, Nature, 2012)