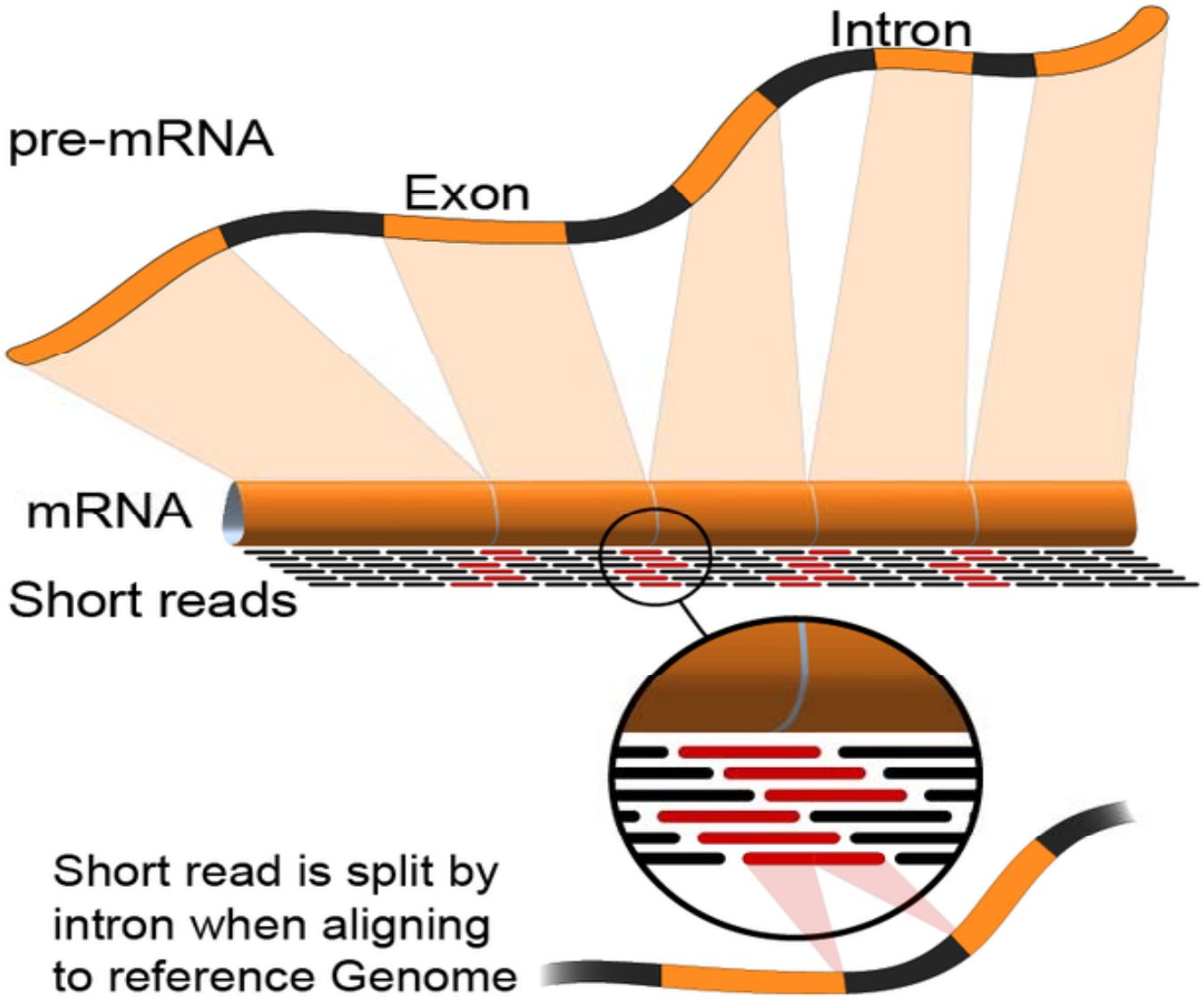
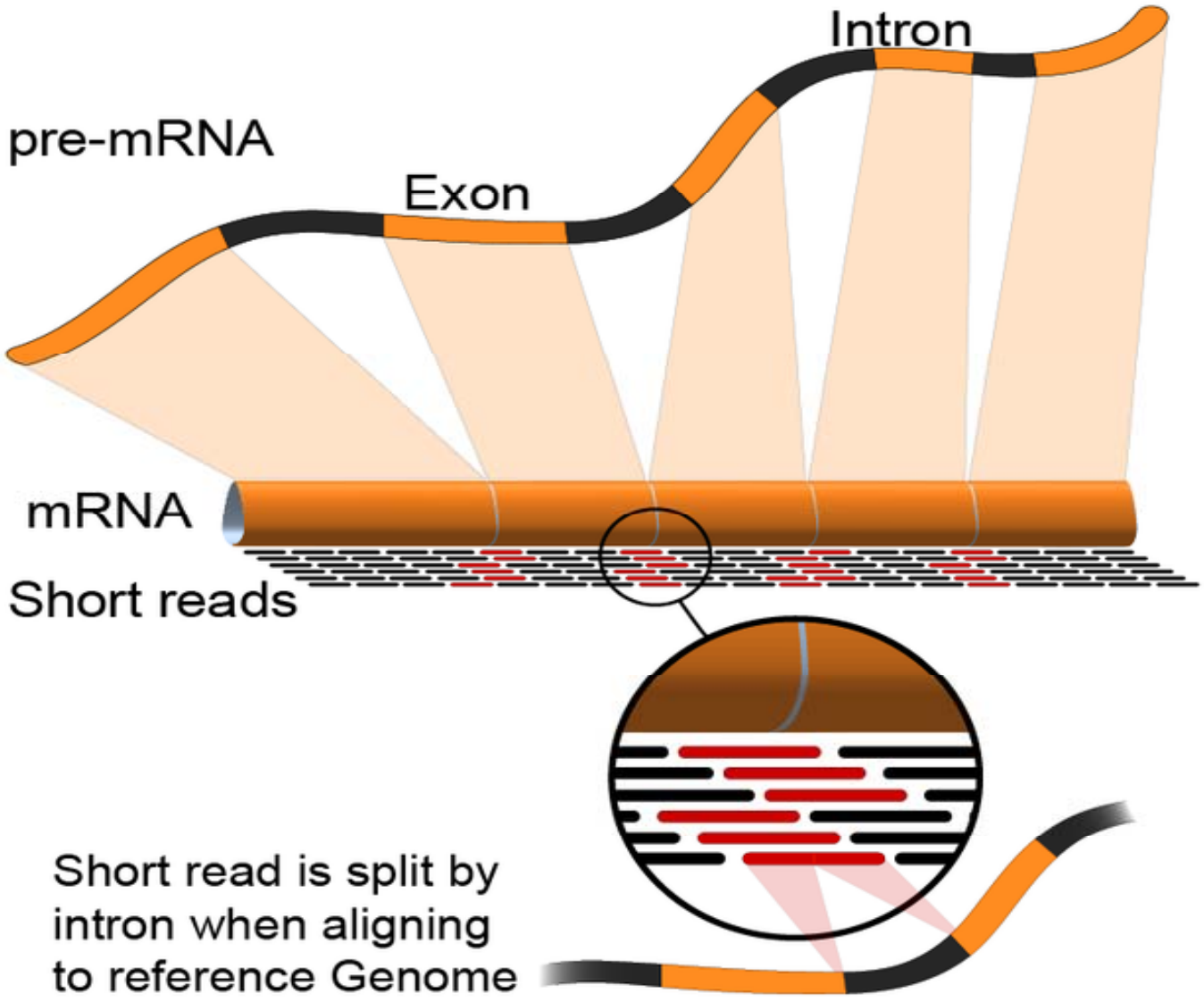
Transcriptome Sequencing with Reference

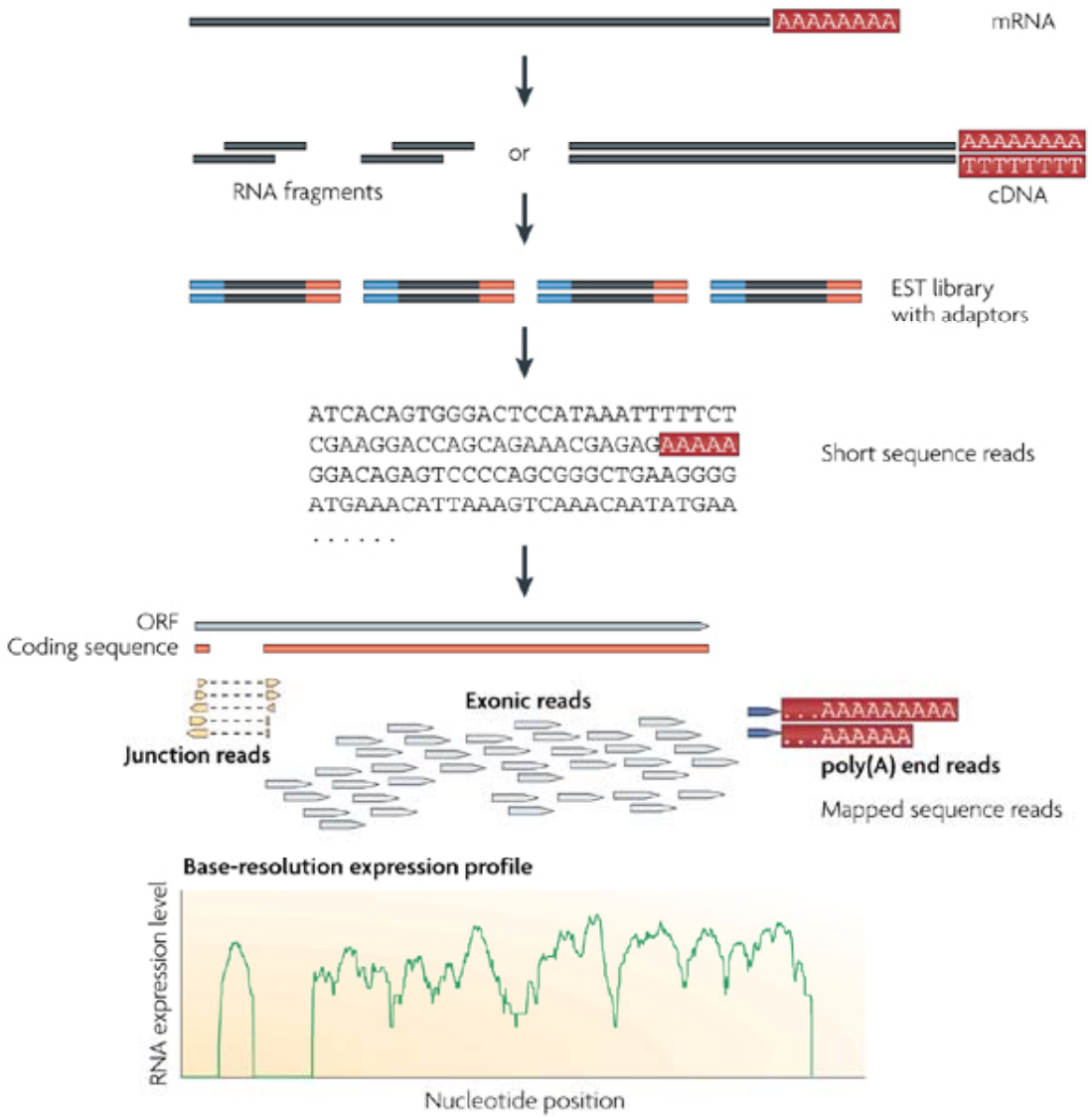


To be continue

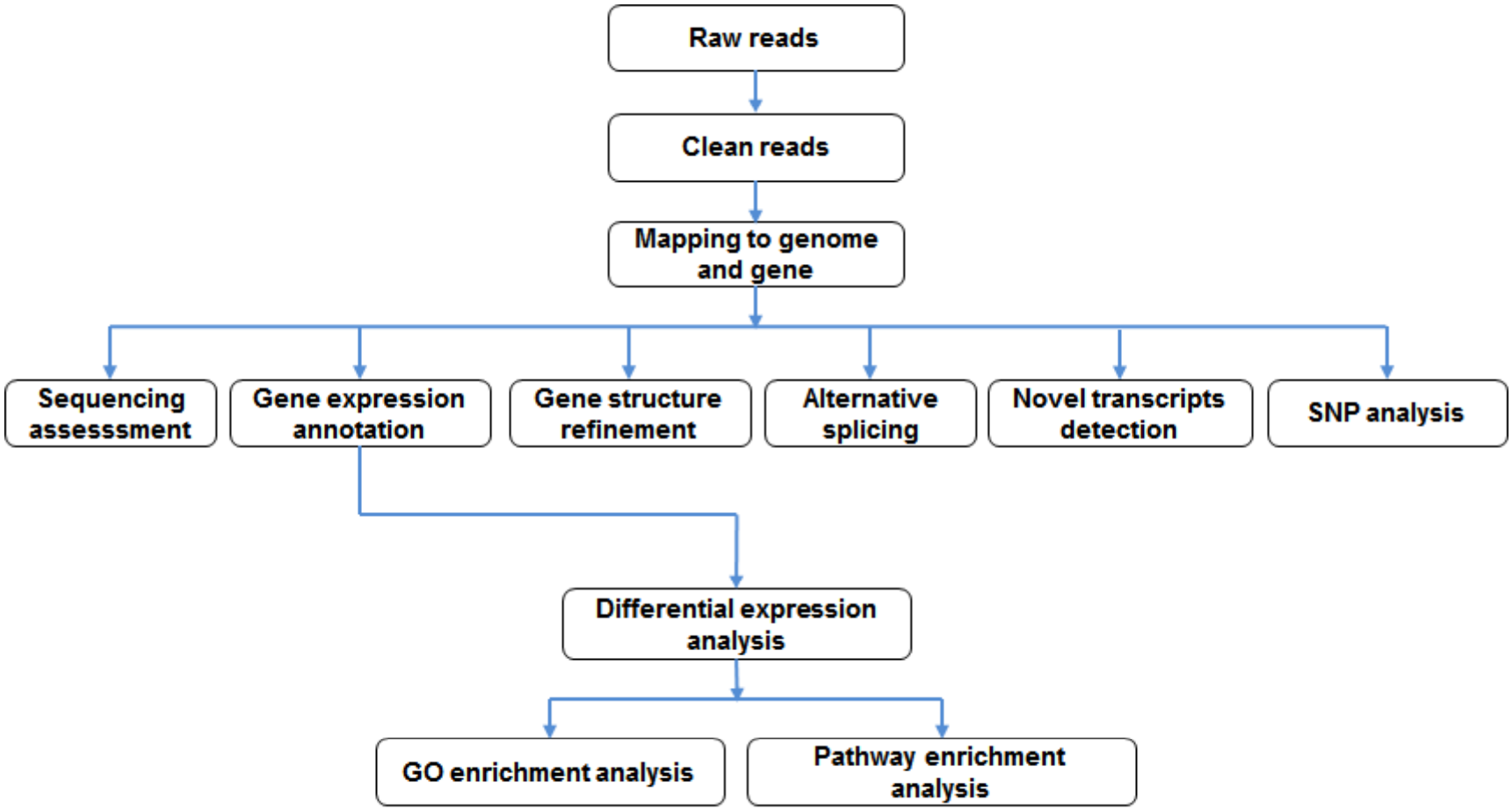
Transcriptome Sequencing with Reference



Overview of RNA-Seq

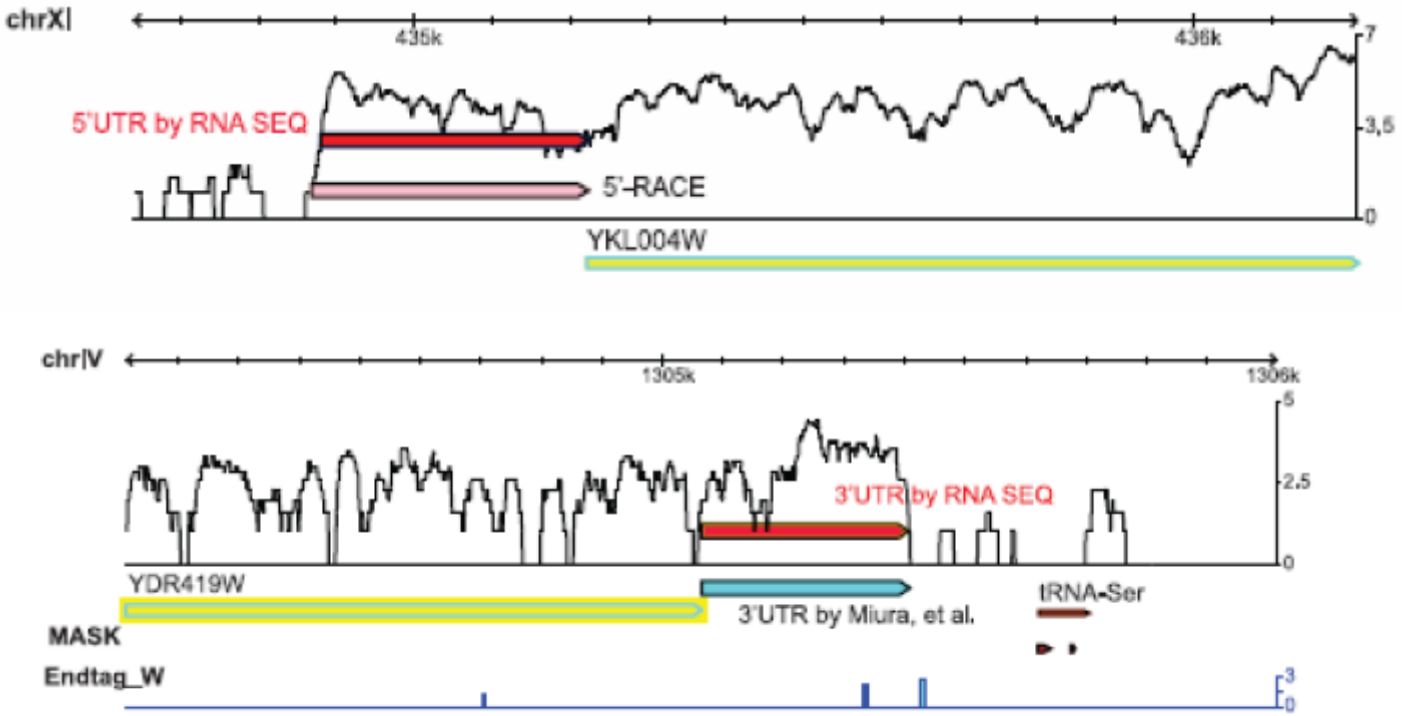


Bioinformatics Analysis Pipeline for Transcritpome Sequencing with Reference



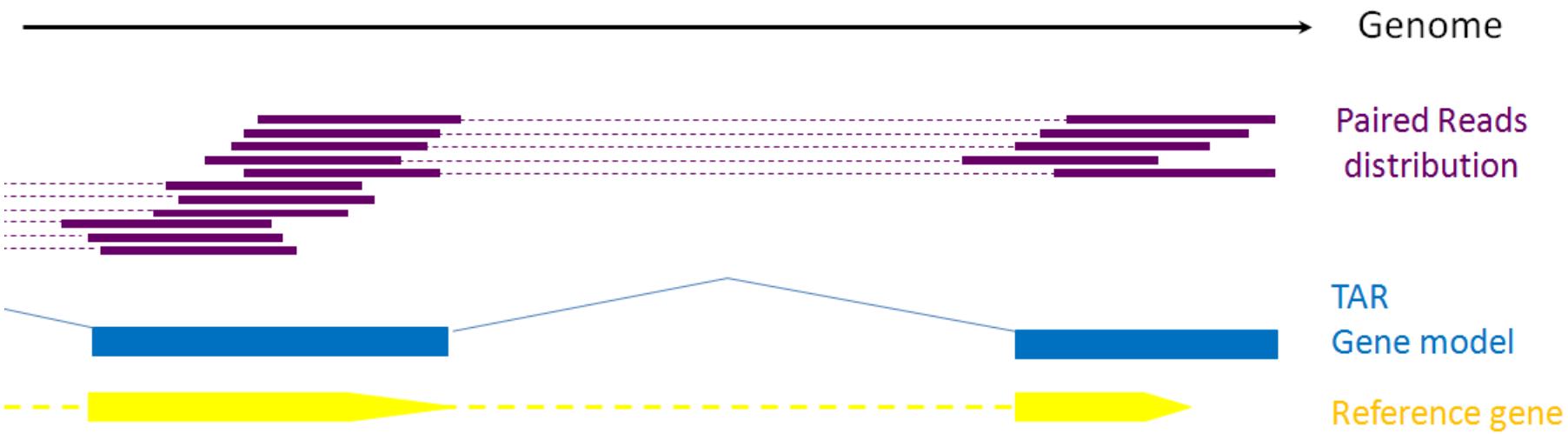
Gene structure refinement

UTRs by RNA-Seq in yeast



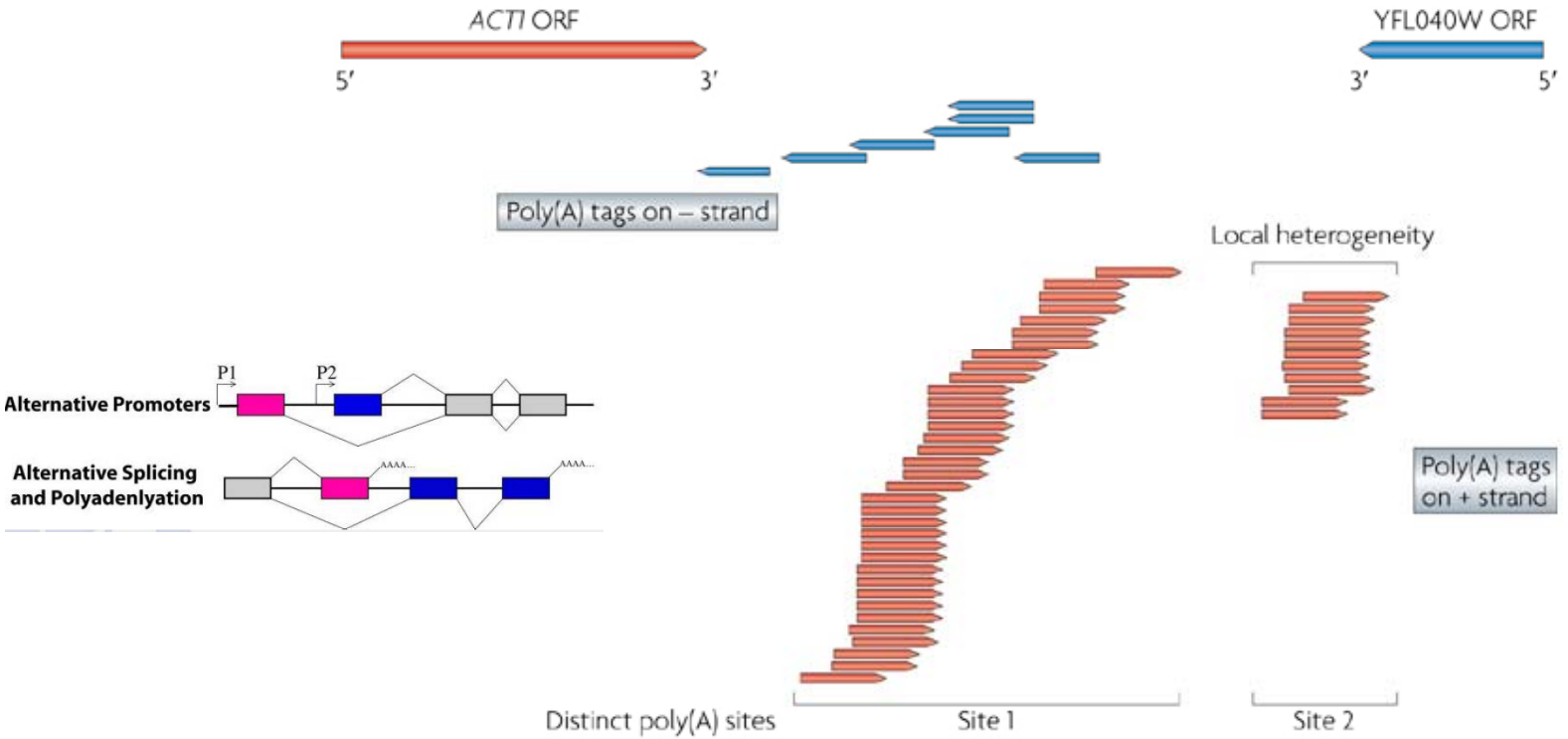
(Nagalakshmi, U. et al.,2008)

Gene structure refinement



The gene structure was optimized according to the distribution of the reads, information of paired-end and the annotation of reference gene. We can get the distribution of reads in the genome by aligning the continuous and overlap reads form a Transcription Active Region (TAR). According to paired-end data, we can connect the different TARs to form a potential gene model. We can compare the gene model with the existing gene annotated to extend the gene 5'and 3' end

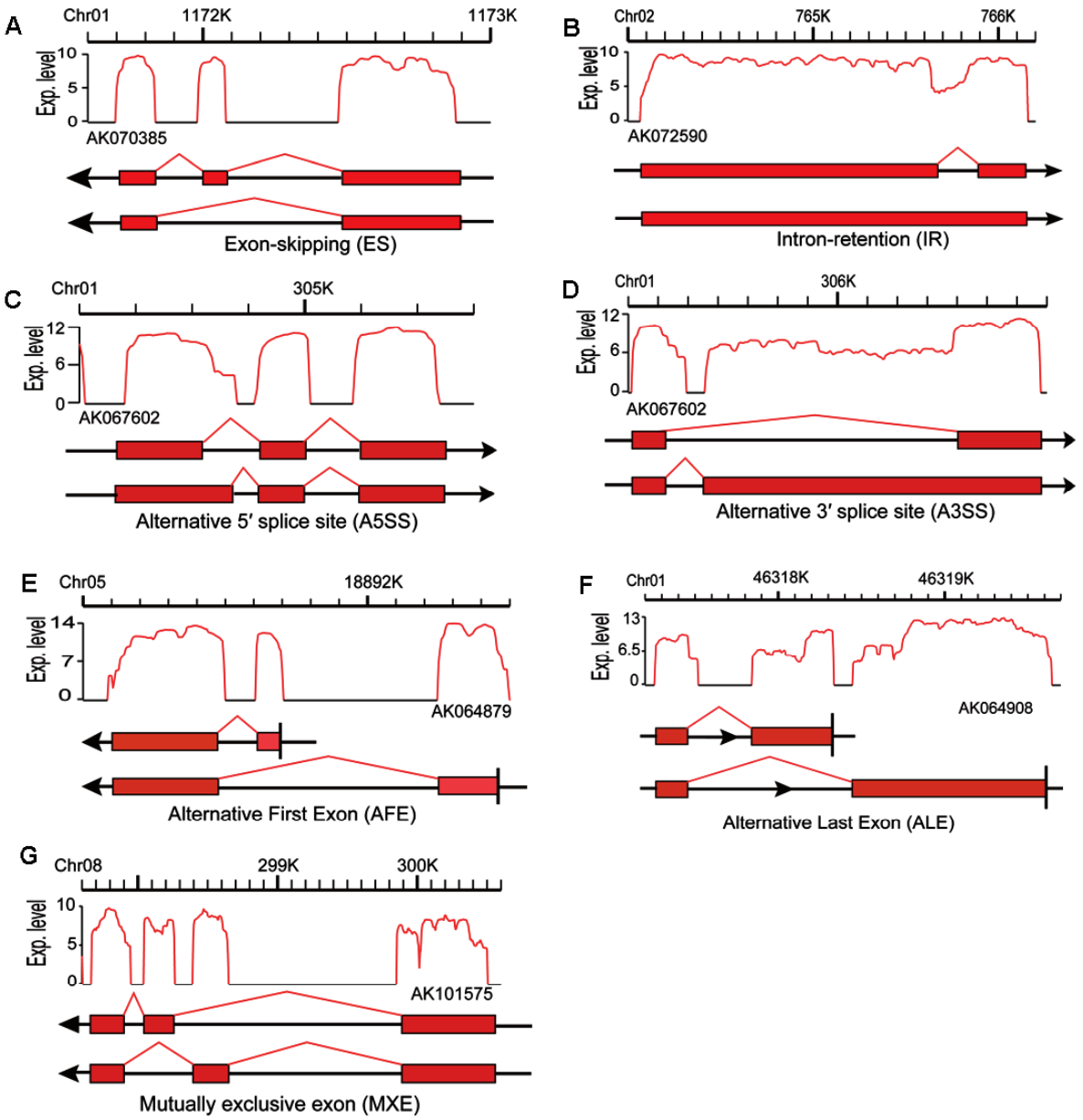
Poly(A) tags from RNA-Seq



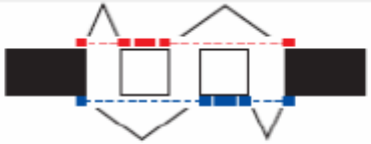
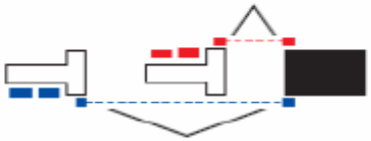
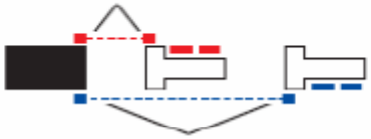
A region containing two overlapping transcripts (ACT1, from the actin gene, and YFL040W, an uncharacterized ORF) from the Saccharomyces cerevisiae genome is shown. Arrows point to transcription direction. The poly(A) tags from RNA-Seq experiments are shown below these transcripts, with arrows indicating transcription direction. The precise location of each locus identified by poly(A) tags reveals the heterogeneity in poly(A) sites, for example, ACT1 has two big clusters, both with a few bases of local heterogeneity. The transcription direction revealed by poly(A) tags also helps to resolve 3'-end overlapping transcribed regions

Nature Reviews Genetics ***10***, *57-63*

Alternative Splicing (AS)

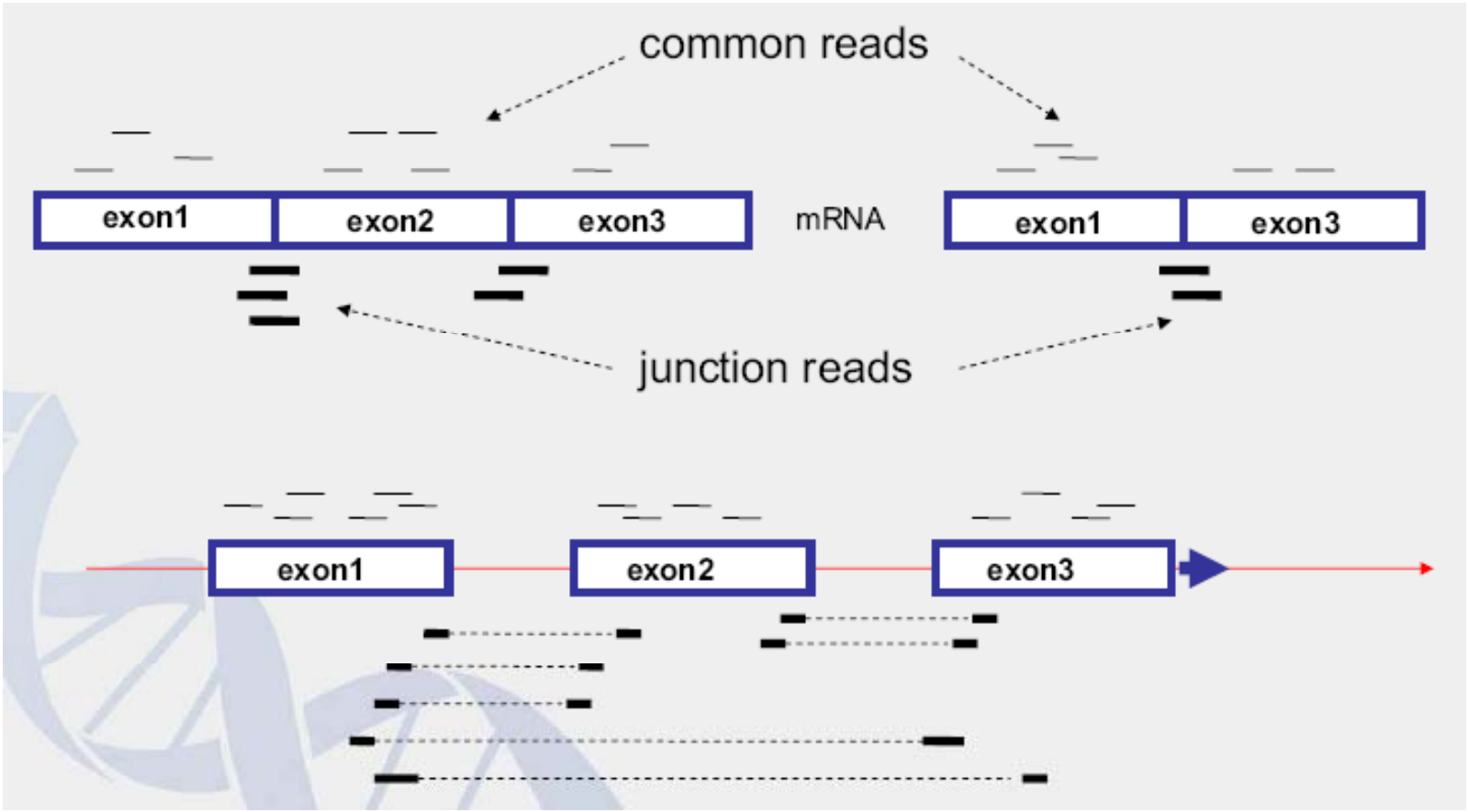


92%~94% of human multi-exon gene undergo Alternative Splicing (AS)

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)
Mutually exclusive exon (MXE)		4	4
Alternative first exon (AFE)		14	13
Alternative last exon (ALE)		9	8
Tandem 3′ UTRs		7	7



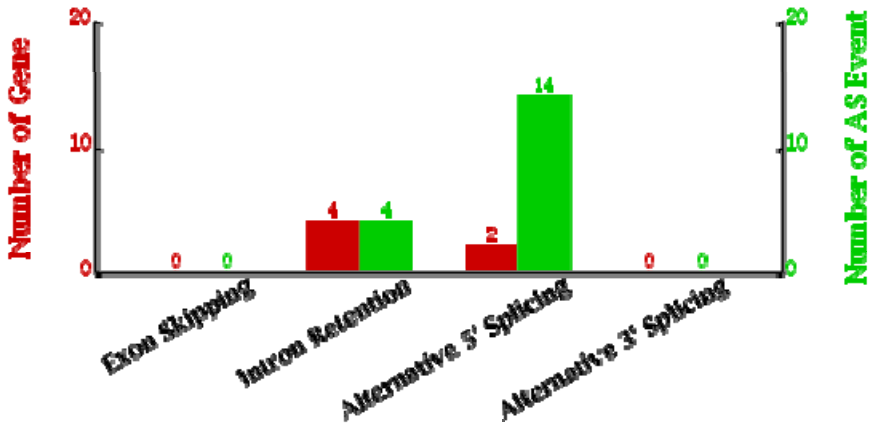
Exon junction reads



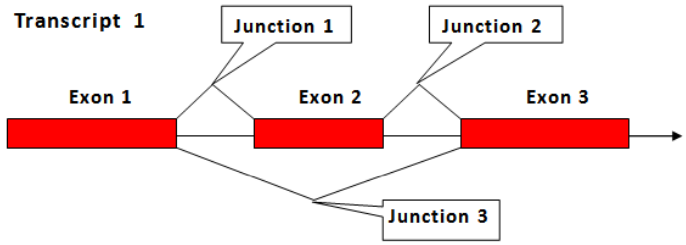
SOAP Short Oligonucleotide Analysis Package

Output of SOAPals

D	E	F	G	H
Name	strand	retained intron	Gene	Location
	+	34279-34388		
CAMK family protein kinase	-	4588-4721	3,745 to 5,333	complement(join(4722..5333,3745..4587))
	-	4588-4721		
	+	99395-99499		
	+	99395-99499		
CAMK family protein kinase	-	56818-56916	56,024 to 57,470	complement(join(57457..57470,56917..57457,56069..56817,56024..56069))
	-	56818-56916		
	-	56818-56916		
	-	56818-56916		
	-	23850-23917		
	-	72375-72446		
nuclear lim interactor-interacting factor	-	8624-8699	7,991 to 9,110	complement(join(8700..9110,8066..8623,7991..8066))
	-	8624-8699		
	-	8624-8699		
	-	8624-8699		
	-	8624-8699		
ubiquitin	+	26766-26993	26,229 to 27,193	join(26229..26765,26994..27146,27146..27193)
	+	26766-26993		
	-	9497-9589		
	-	78748-78814		
	-	78748-78814		



Splice Junction Database

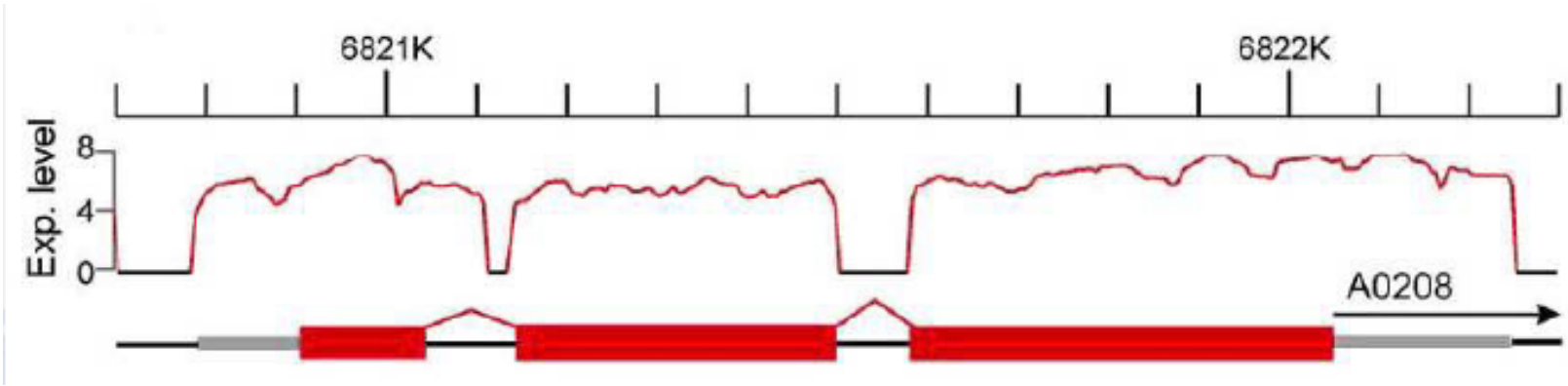


Species: Oryza Sativa
Chromosome: 1
All the Genes

```
SELECT * FROM oryzasativa_junction where GeneID=4326813;
```

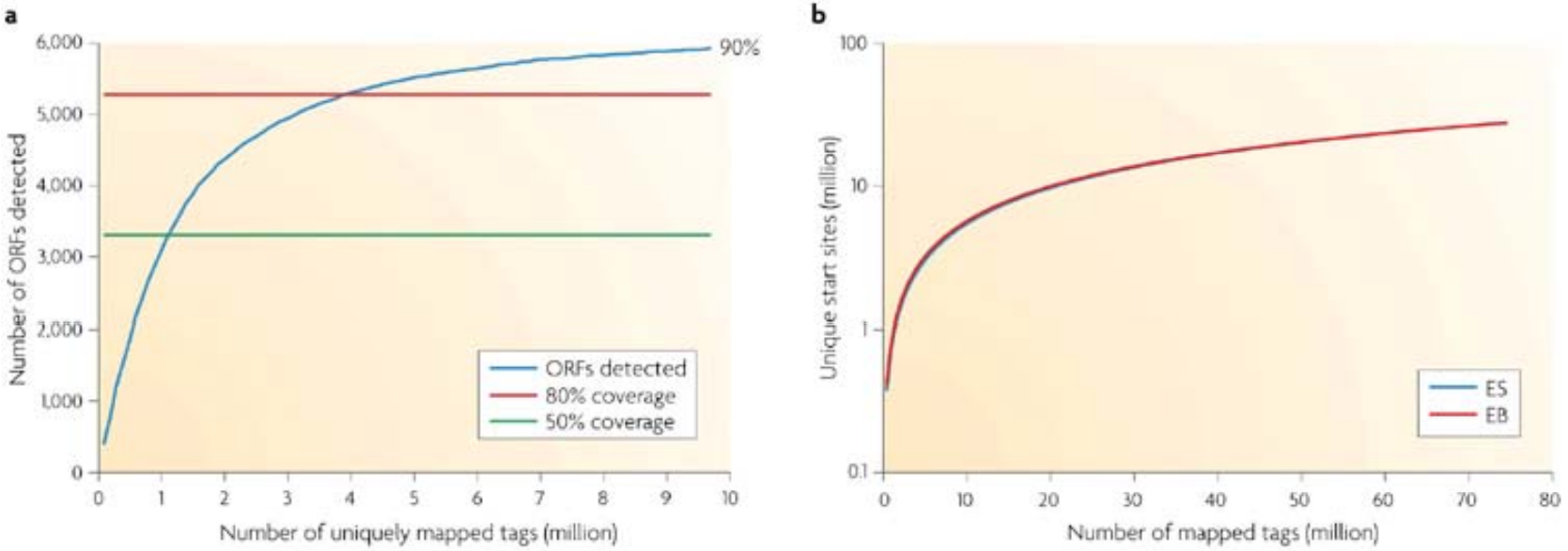
JunctionID	GeneID	exon3P	exon5P	splicejunc
490591	4326813	1	2	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGGAGCAAAATGCTCAAACCGGAAAAATGGCACAC
490592	4326813	2	3	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGGTGTGGATCCCACTATTCGAGCTGAAGTTTGGGA
490593	4326813	3	4	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490594	4326813	4	5	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACGATTACAATGACATGGGTGAACCAAGATATGACAC
490595	4326813	5	6	AGTCTCTGAATGGCTTTGGACACTGCACCGAATTGTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490596	4326813	6	7	CCTTTTGGTGCTTTGAGATGCTACTGAGAAGGATGCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT
490597	4326813	7	8	TTAGCATGTGGGAGGTGTGGCATTTGAAGACACATATGATGTGGGCTGCTGACTTTAATGAAGATGTGAT
490598	4326813	8	9	ATGATCTGCCTATATTCTGCGTAGCAGCAATCCTTATGTTCAATGATAATATGCTGAAGATCAATGTAAA
490599	4326813	9	10	TGGCTATCAAGCTGAGGAAGAAGTATATTTACAAGGGAATCTGTACATATGCACAGTATACTTGAAAACA
490600	4326813	1	3	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGTGTGGATCCCACTATTCGAGCTGAAGTTTGGGA
490601	4326813	2	4	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490602	4326813	3	5	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGGATTACAATGACATGGGTGAACCAAGATATGACAC
490603	4326813	4	6	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490604	4326813	5	7	AGTCTCTGAATGGCTTTGGACACTGCACCGAATTGCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT
490605	4326813	6	8	CCTTTTGGTGCTTTGAGATGCTACTGAGAAGGATGATGATGTGGGCTGCTGACTTTAATGAAGATGTGAT
490606	4326813	7	9	TTAGCATGTGGGAGGTGTGGCATTTGAAGACACATATGTTCAATGATAATATGCTGAAGATCAATGTAAA
490607	4326813	8	10	ATGATCTGCCTATATTCTGCGTAGCAGCAATCCTTGGAATCTGTACATATGCACAGTATACTTGAAAACA
490608	4326813	1	4	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490609	4326813	2	5	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGATTACAATGACATGGGTGAACCAAGATATGACAC
490610	4326813	3	6	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490611	4326813	4	7	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT

Novel Transcripts



Novel transcripts can be found by high throughput sequencing since present databases may be incomplete. Gene models found in intergenic regions (200 bp away from upstream or downstream genes) were thought to be candidate of novel transcripts

How deep should we go?



(a) 80% of yeast genes (**genome size: ~12MB**) were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3' end.

(b) The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body.

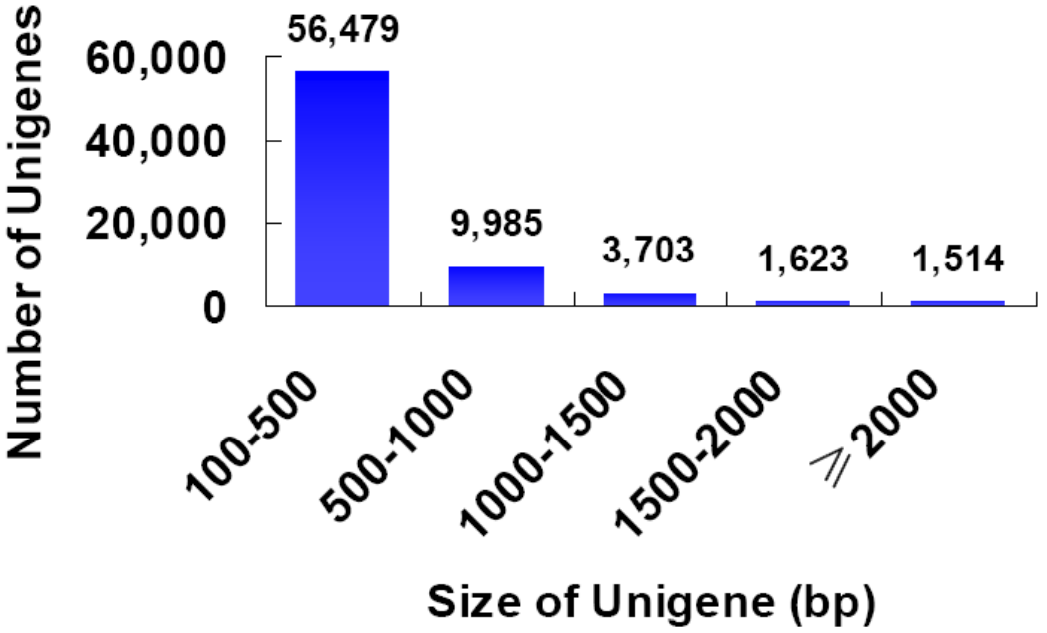
Nature Reviews Genetics ***10***, *57-63*

How deep should we go?

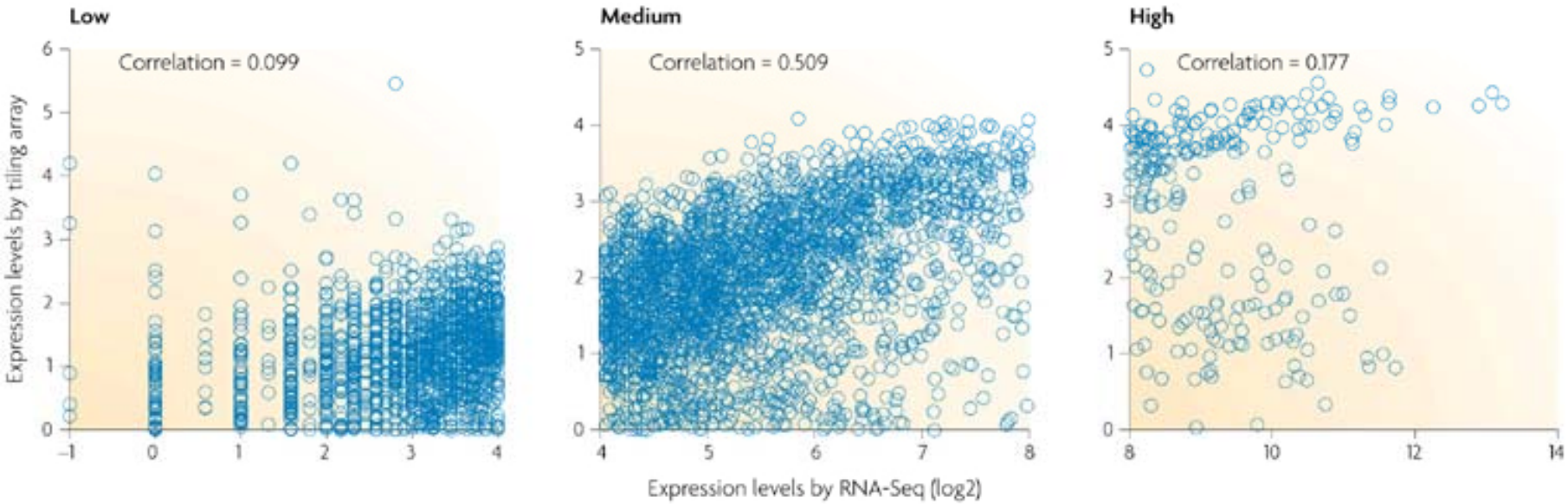
De novo assembled rice transcriptome

1.3 Gb RNA-Seq data (genome size: ~400MB)

85% of assembled unigenes were covered by genemodels



RNA-Seq and microarray compared

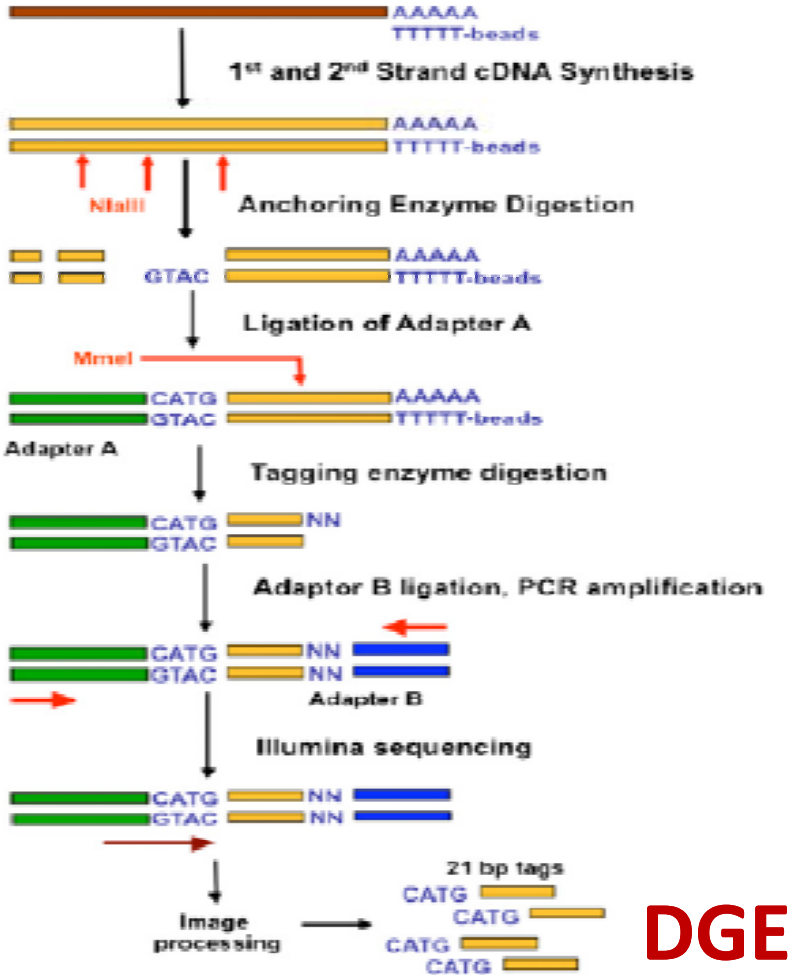


Expression levels are shown, as measured by RNA-Seq and tiling arrays, for Saccharomyces cerevisiae cells grown in nutrient-rich media. The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels.

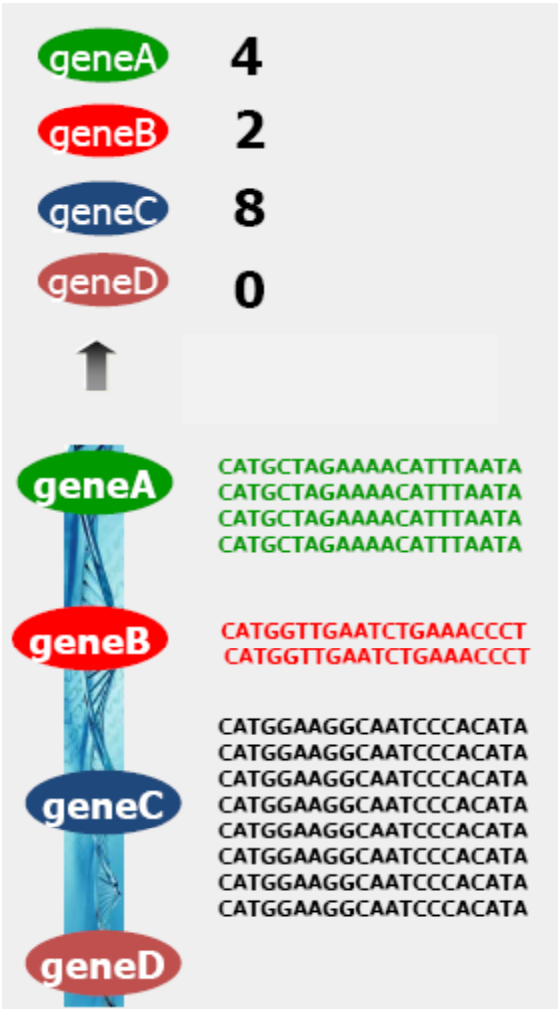
Nature Reviews Genetics ***10***, *57-63*

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Digital Gene Expression Profiling

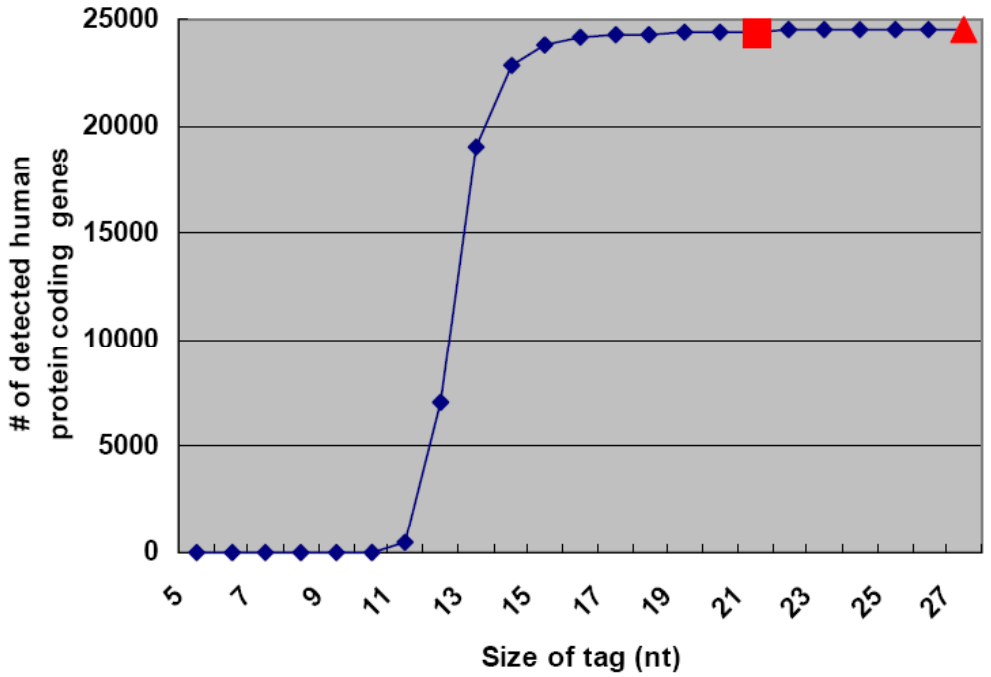


DGE



Tag length *vs* Complexity

Tag length (*n* base pairs)	Complexity[a]	Tag uniqueness probability[b]
14	1,048,576	0.00%
15	4,194,304	0.08%
16	16,777,216	16.73%
17	67,108,864	63.95%
18	268,435,456	89.43%
19	1,073,741,824	97.24%
20	4,294,967,296	99.30%
21	17,179,869,184	99.83%



Exercise AGAIN……