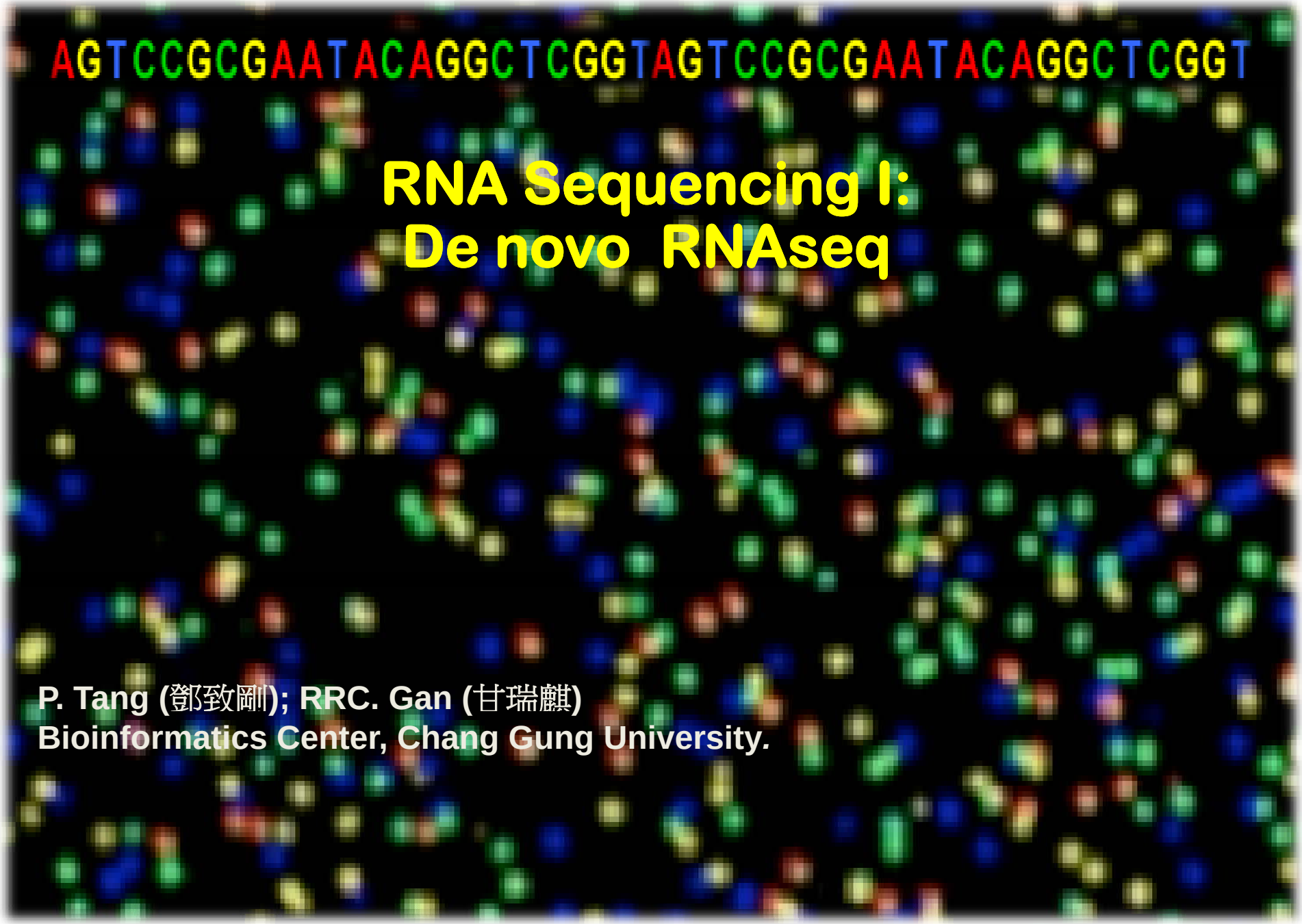
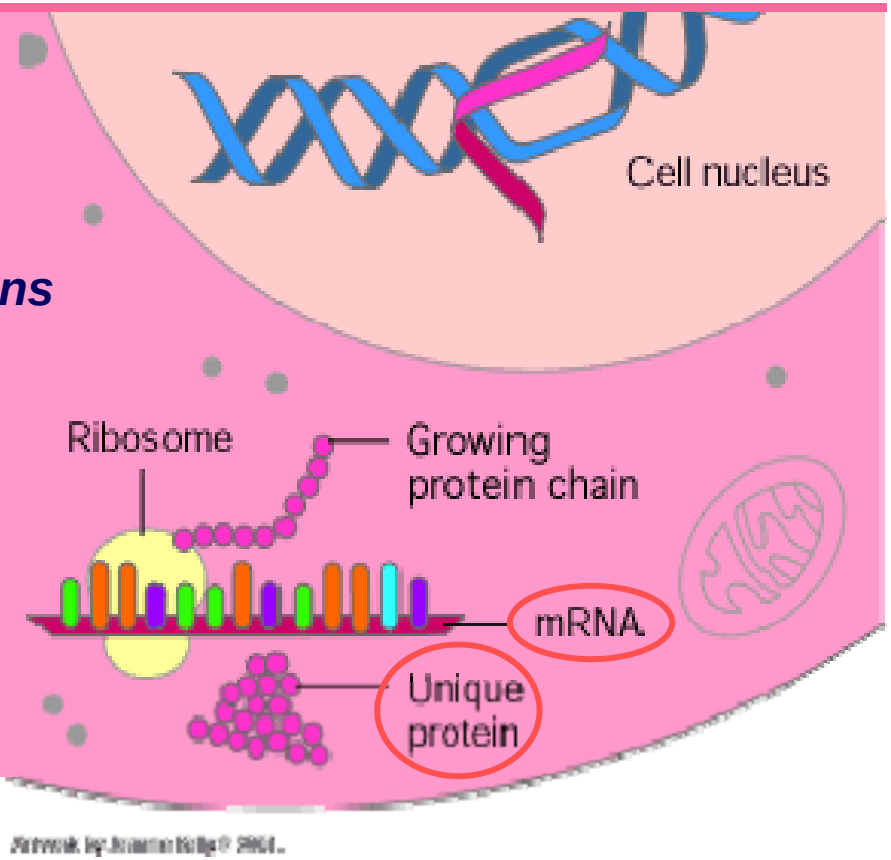


RNA Sequencing I: De novo RNAseq

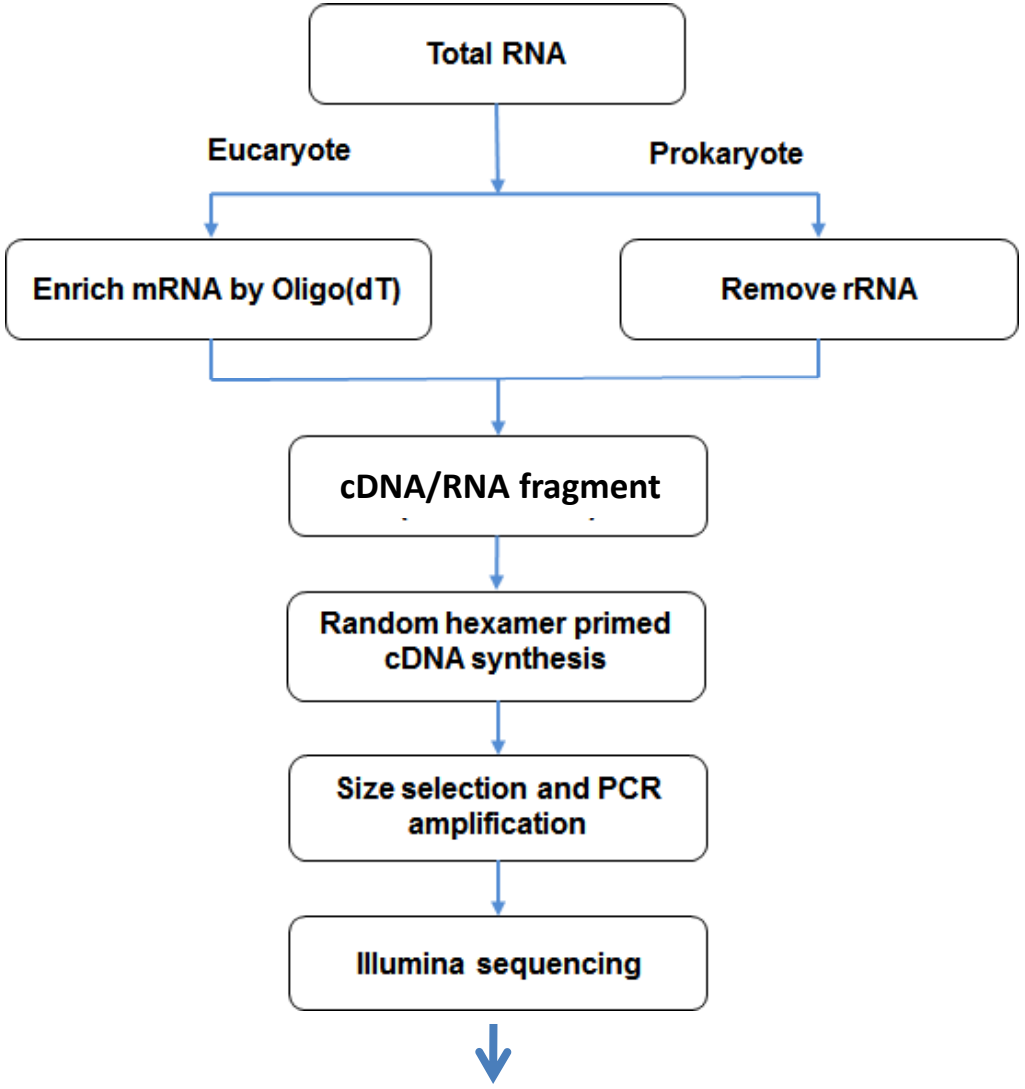
P. Tang (鄧致剛); RRC. Gan (甘瑞麒)
Bioinformatics Center, Chang Gung University.

Why Measure Gene Expression?

Unique set of genes
are expressed
at different growth conditions
and
at different stages.



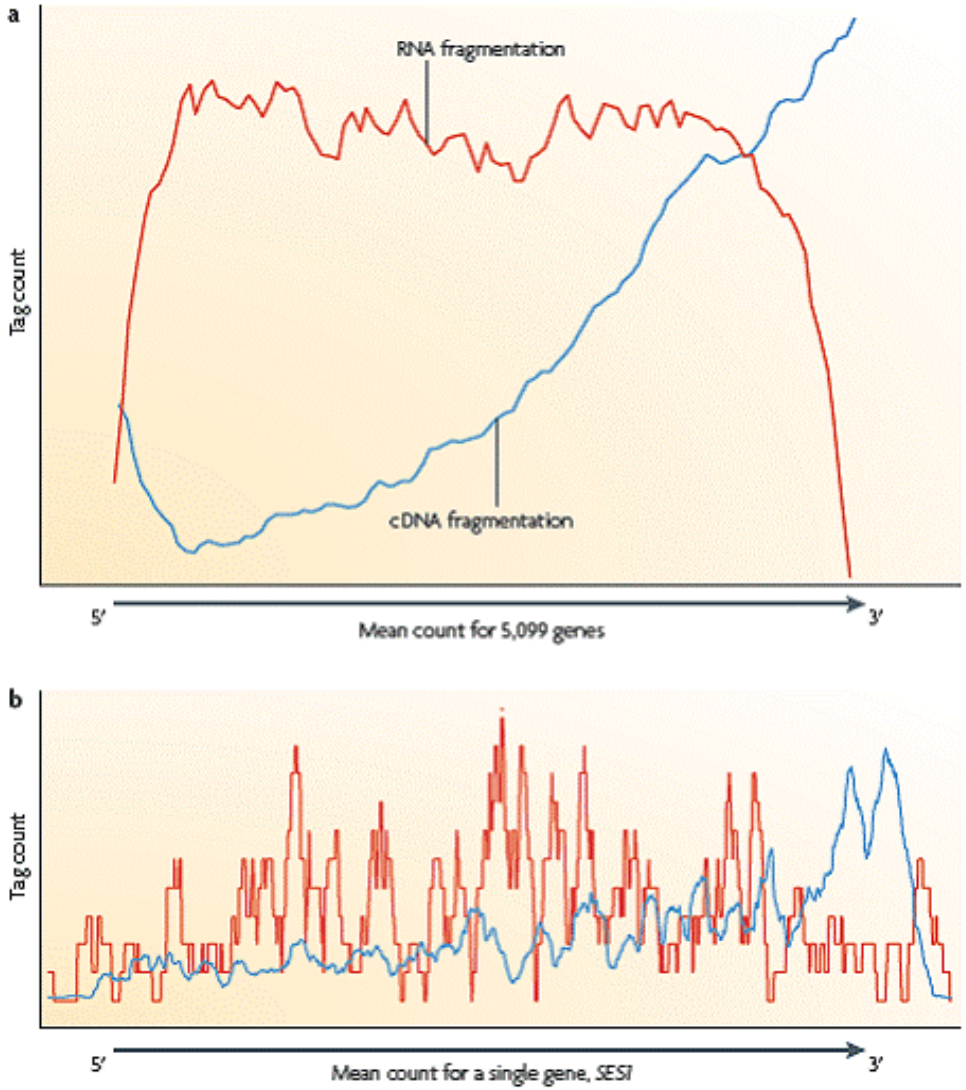
Experimental Workflow



De novo Transcriptome Analysis
Transcriptome Analysis with Regerence

Library Preparation *vs* Sequencing randomness

Fragmentation of mRNA/cDNA was performed through the physical or chemical methods during the experiment of transcriptome analysis. If the randomness of fragmentation is poor, reads would more frequently generated from specific regions of the original transcripts and the following analysis will be affected.



De novo Transcriptome Sequencing

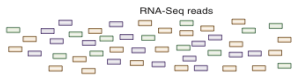
Assembly is the only option when working with a creature with no genome sequence, alignment of contigs may be to ESTs, cDNAs etc

RNAseq reads



Filer clean reads

Remove reads which containing adaptors
Remove reads in which unknown bases are more than 5%
Remove low quality reads (more than half of the bases' qualities are less than 5)



De novo assembly

Contigs



Functional Annotation

- **BLASTx NCBI nr**
- **BLASTx Uuiprot**
- **Protein domain/motif search**
- **Gene Ontology**
- **KEGG**
- **Specific databases**

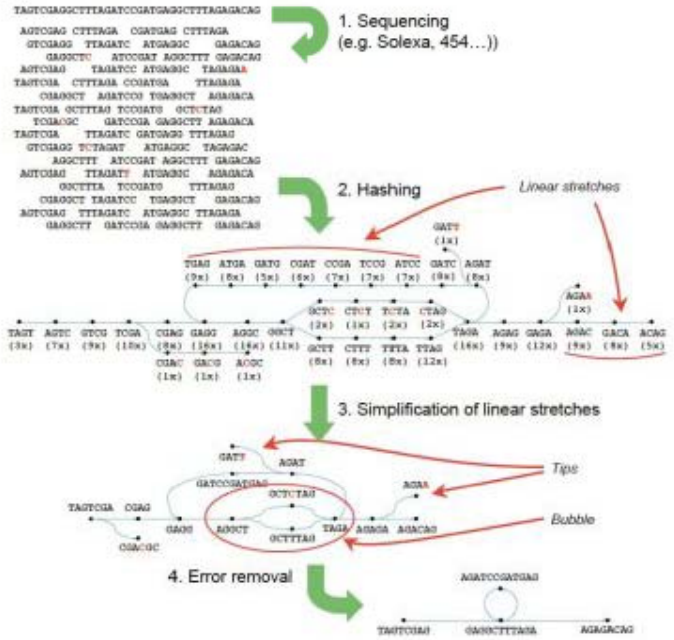
De novo Assembler

Velvet *http://www.ebi.ac.uk/~zerbino/velvet/*

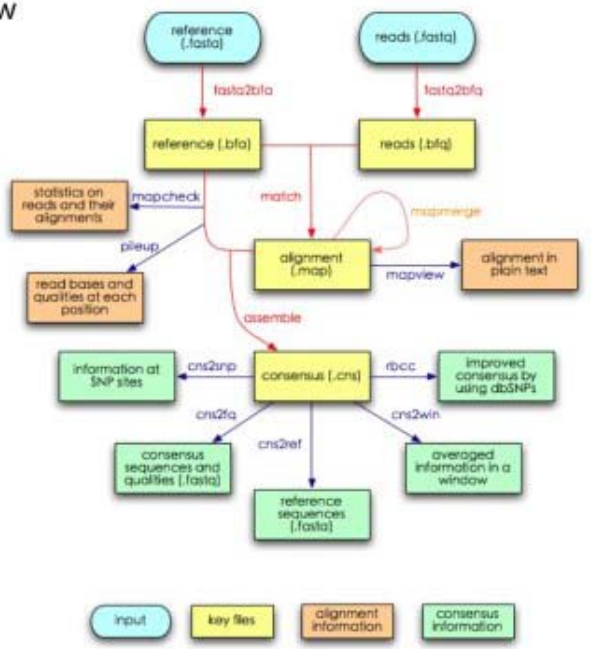
Maq *http://maq.sourceforge.net/*

SOAP de novo *http://soap.genomics.org.cn/*

Velvet assembly process



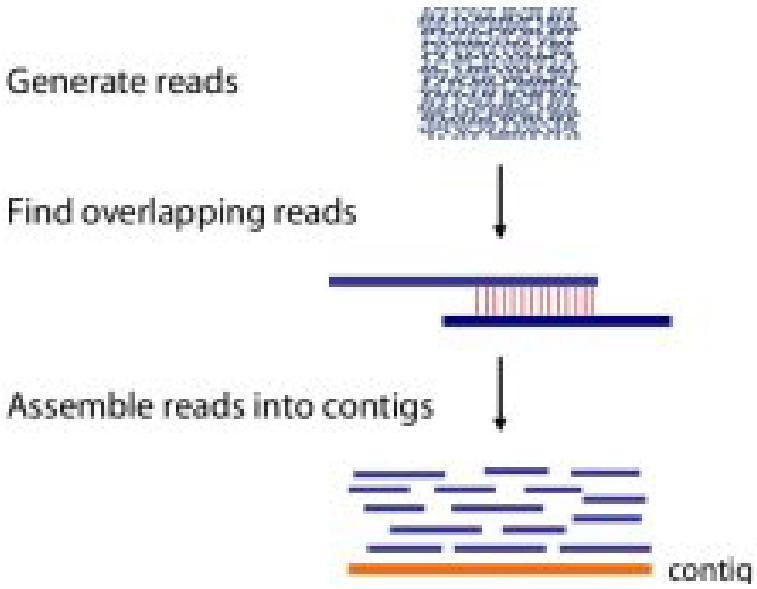
Maq dataflow



SOAP

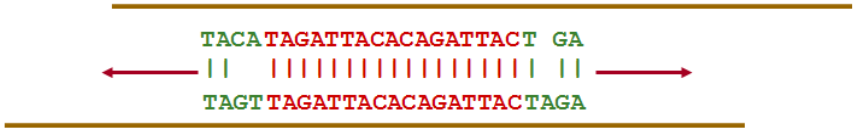
Short Oligonucleotide Analysis Package

Parameters for Assemble



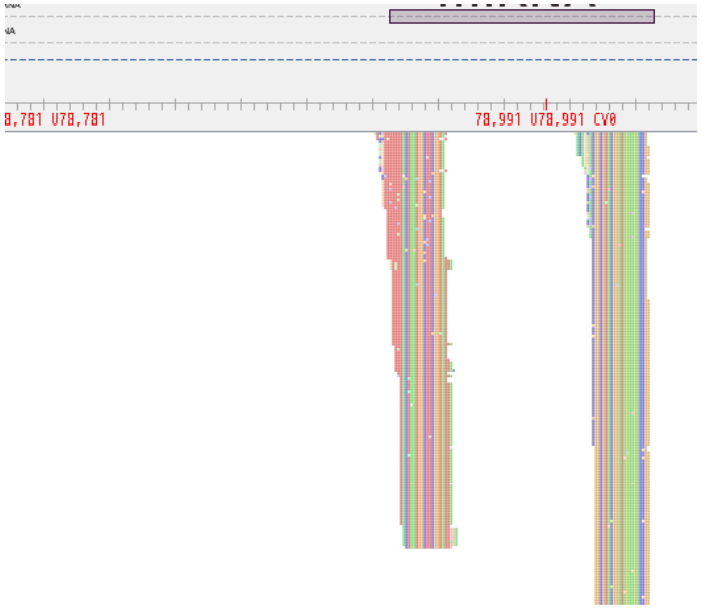
Important Parameters:

1. Percentage of Overlap
   - 100%, 80%, 50%, 20%?
2. Percentage of allowed mismatches
   - 10% or 20%?



Assembled/Aligned Reads

Contig/Gene



- Total reads in a contig/gene (mapped reads)
- Forward reads
- Reverse reads
- Non-specific reads
- Non-perfect reads
- Unique reads (Total reads – non specific reads)

Gene Expression Annotation

Gene coverage

Gene coverage is the percentage of a gene been covered by reads.
This value equals to ratio of the number of bases in a gene covered by **unique** mapping reads to number of total bases in that gene

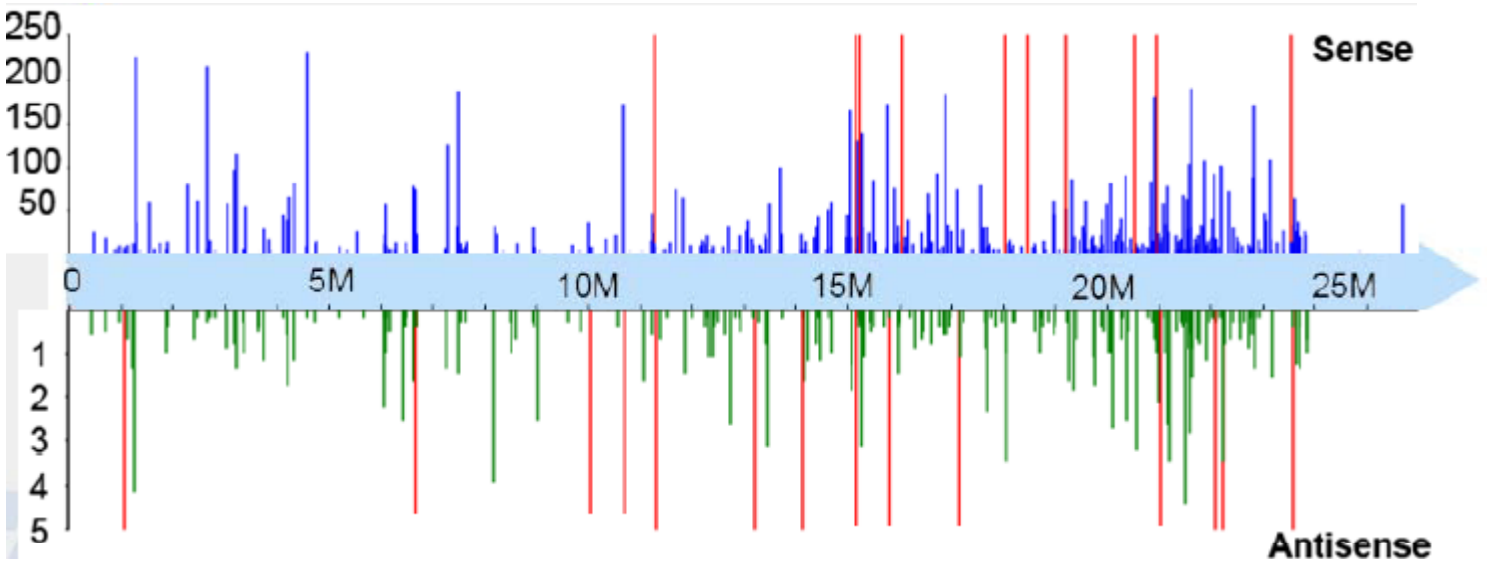
Gene expression levels

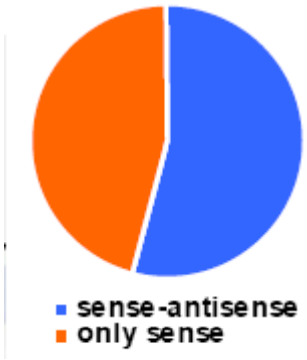
The calculation of Unigene expression uses RPKM method (Reads Per kb per Million reads)
The RPKM method is able to eliminate the influence of different gene length and sequencing discrepancy on the calculation of gene expression. Therefore, the calculated gene expression can be directly used for comparing the difference of gene expression among samples

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

C = number of reads that uniquely aligned to gene A,
N = total number of reads that uniquely aligned to all genes,
L = number of bases on gene A.

Sense *vs* Anti-sense Transcripts



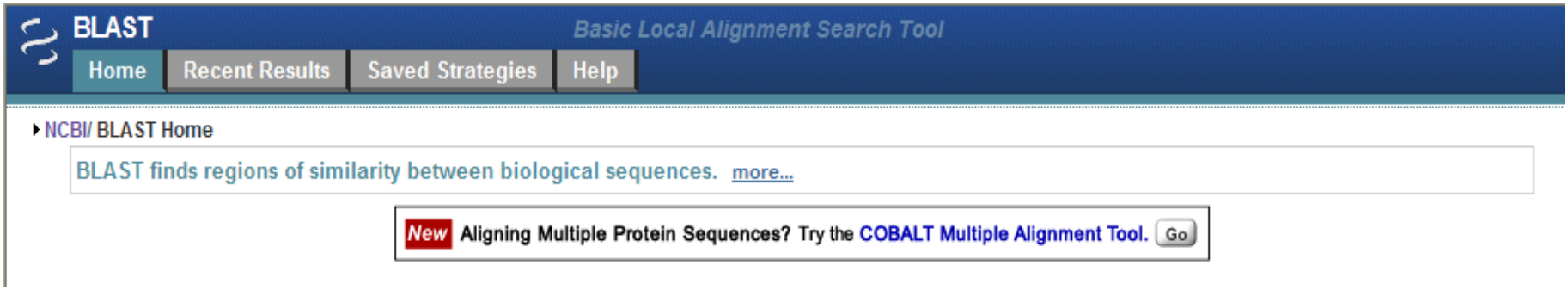


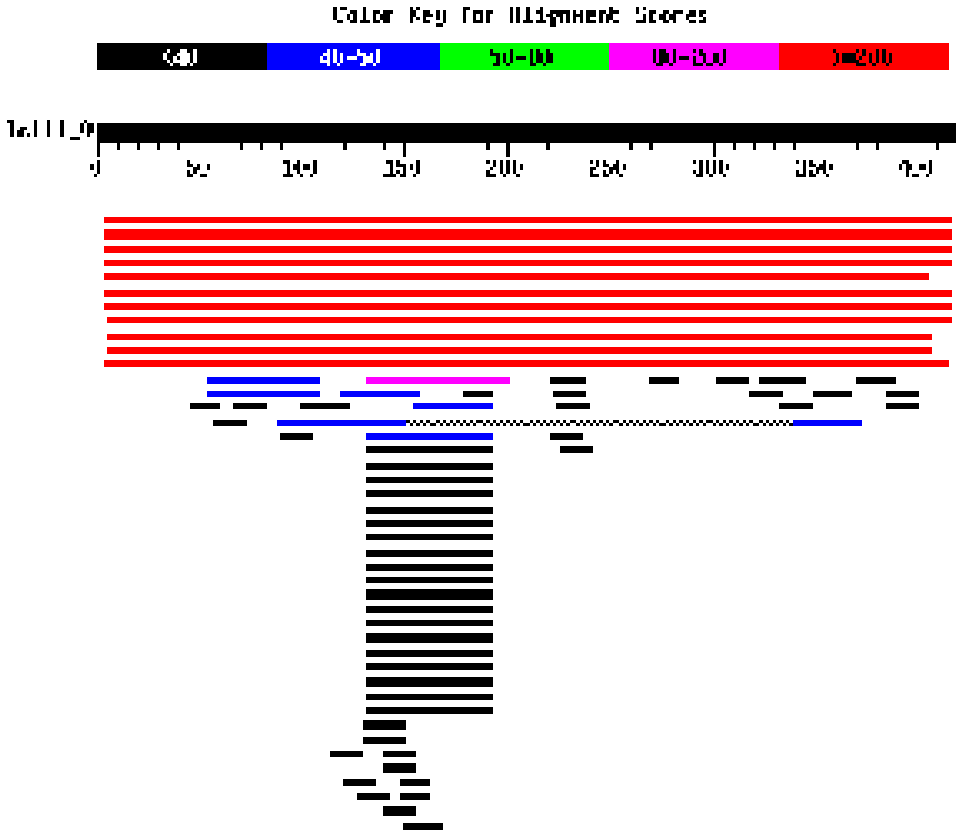
Human



Mouse

BLAST





E-vale
Score
% Identity
% Length

Stand-alone BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

Download BLAST Software and Databases

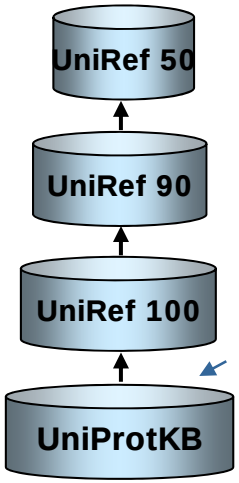
BLAST+ executables

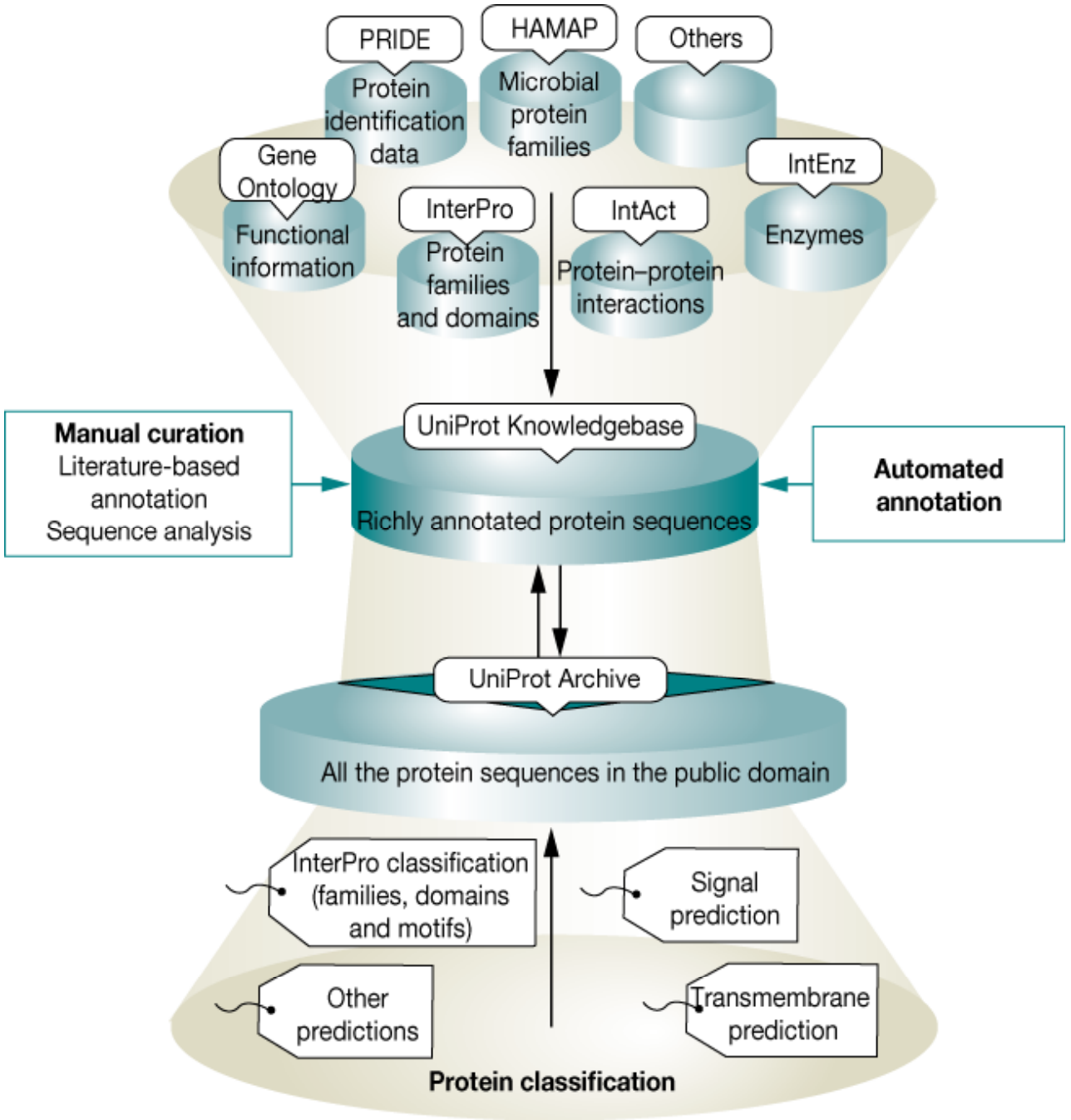
BLAST+ is a new suite of BLAST tools that utilizes the NCBI C++ Toolkit. The BLAST+ applications have a number of performance and feature improvements over the legacy BLAST applications. For details, please see the BLAST+ user manual and the article in BMC Bioinformatics (PubMed link).

platform	archive
Windows (32-bit x86, MSI installer)	download
Windows (64-bit x64, MSI installer)	download
Linux (32-bit x86, RPM)	download
Linux (64-bit x64, RPM)	download
Mac OS X (universal, DMG)	download
Solaris 10 (64-bit SPARC, .tar.gz)	download
Solaris 10 (64-bit x64, .tar.gz)	download

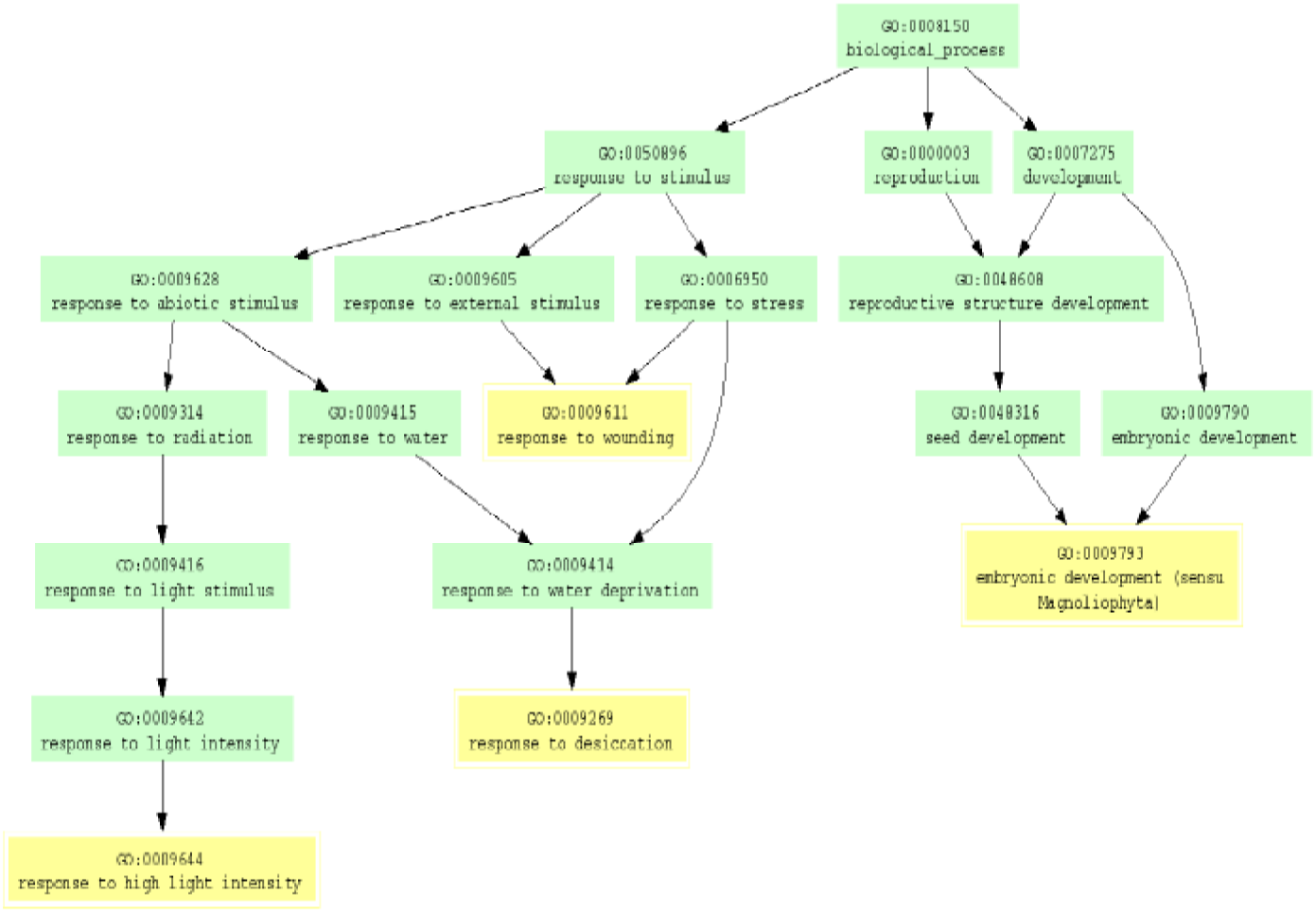
Source code and other archive formats are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ .

UniProt

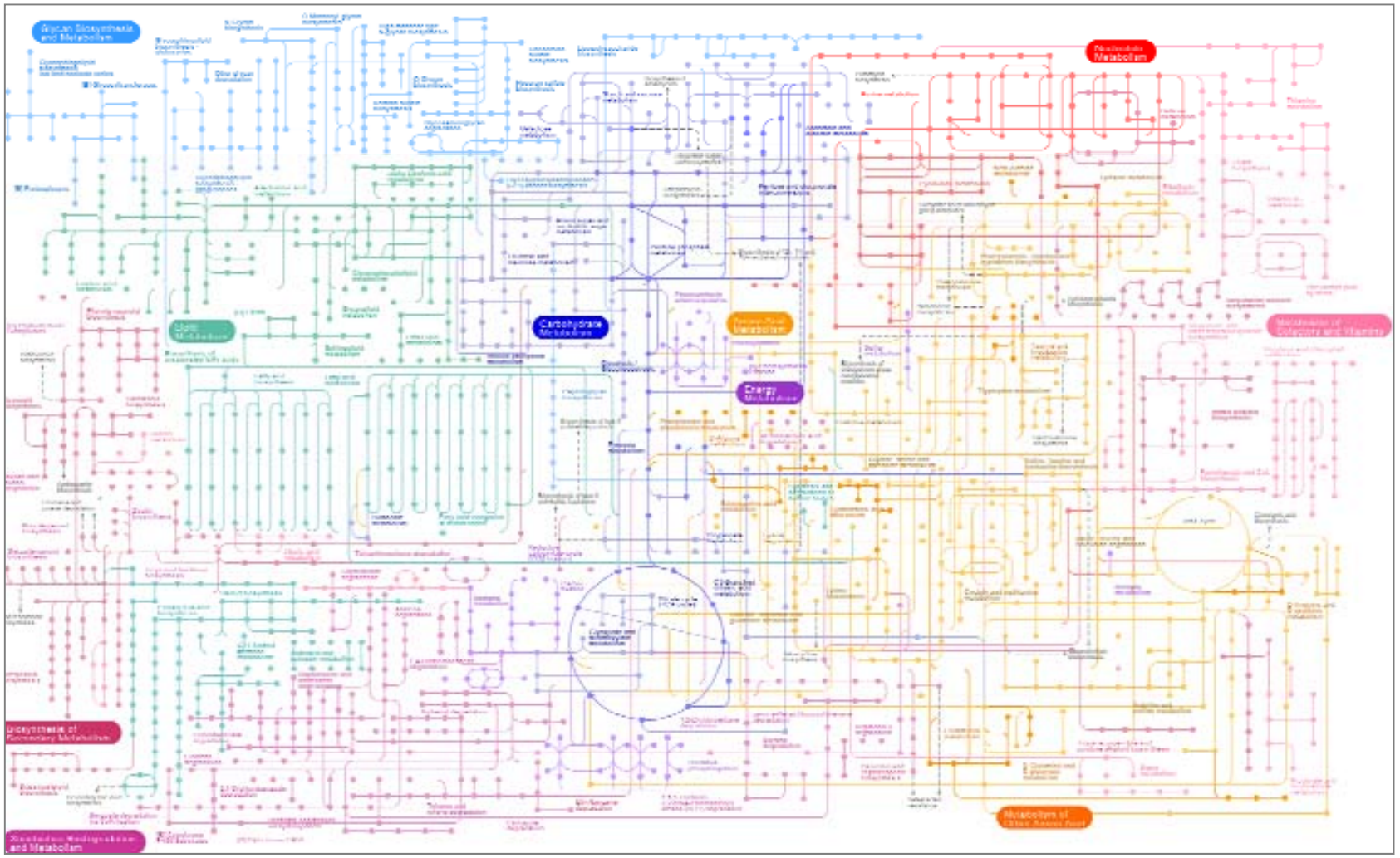

UniRef 50
UniRef 90
UniRef 100
UniProtKB


PRIDE
Protein identification data
HAMAP
Microbial protein families
Others
Gene Ontology
Functional information
InterPro
Protein families and domains
IntAct
Protein–protein interactions
IntEnz
Enzymes
Manual curation
Literature-based annotation
Sequence analysis
UniProt Knowledgebase
Richly annotated protein sequences
Automated annotation
UniProt Archive
All the protein sequences in the public domain
InterPro classification (families, domains and motifs)
Signal prediction
Other predictions
Transmembrane prediction
Protein classification

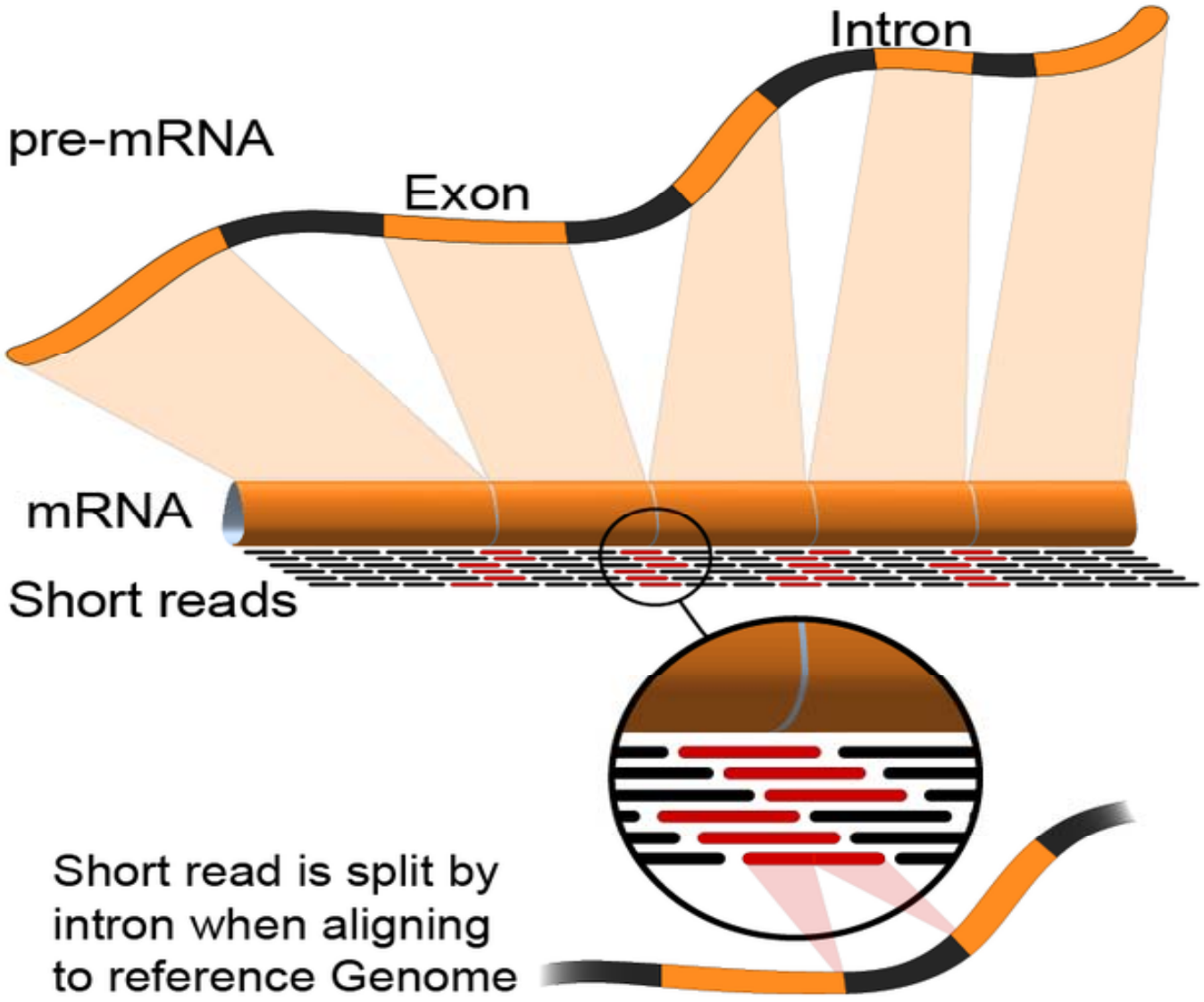
Gene Ontology



KEGG


Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites
Metabolism of Other Amino Acids
Xenobiotics Biodegradation and Metabolism

Transcriptome Sequencing with Reference



To be continue