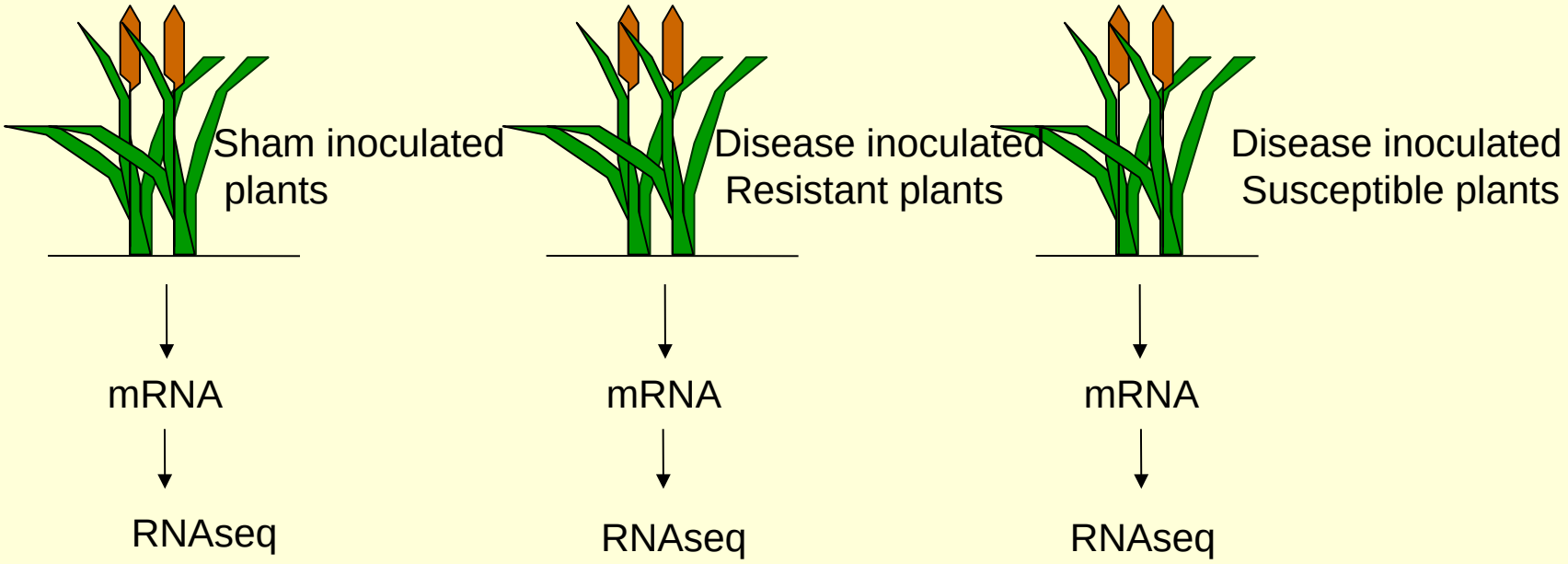
PLNT2530 Plant Biotechnology
2024
Unit 7

# Finding Genes with Genomics Technologies

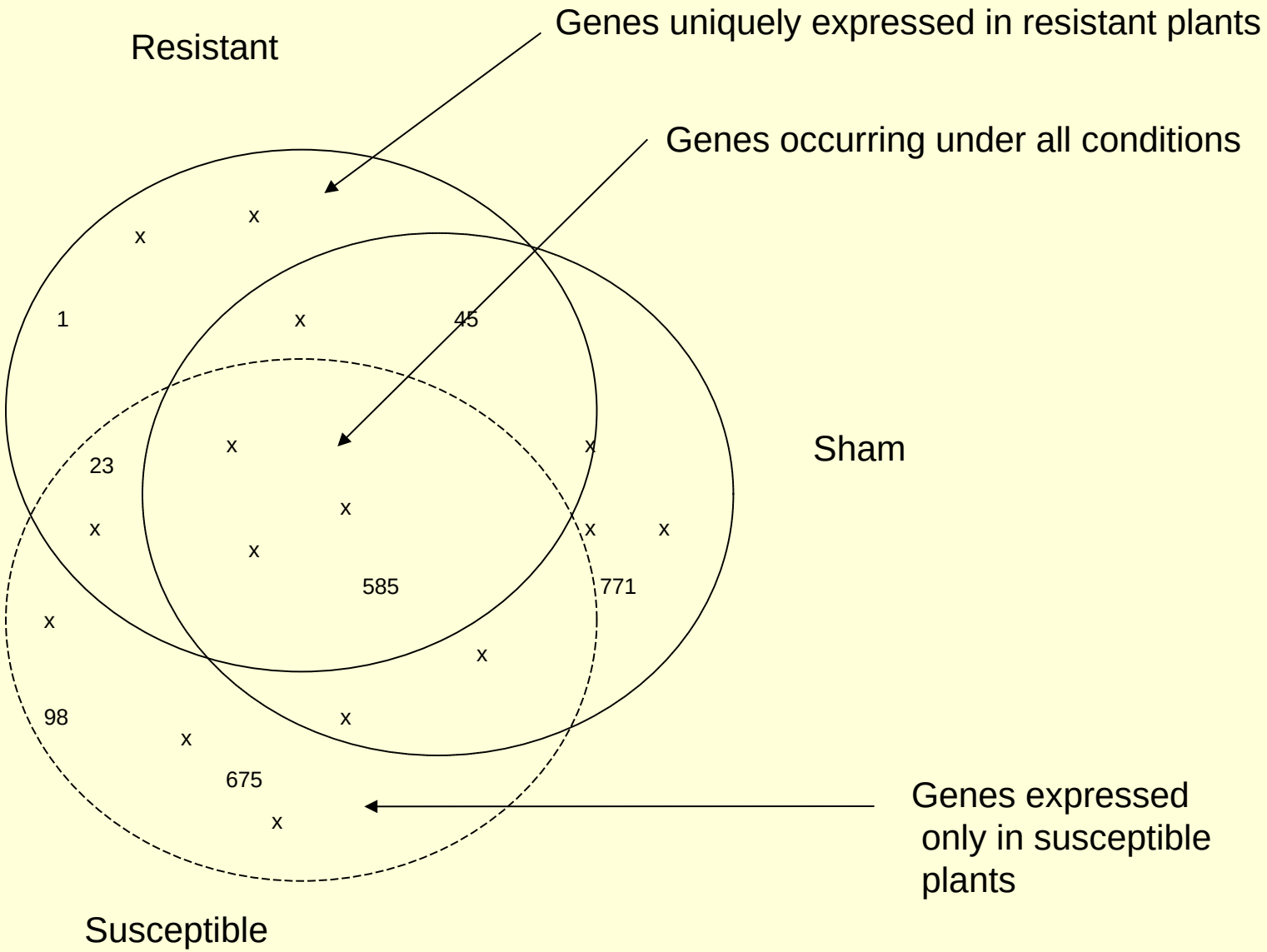
# “Omics” technologies

| Technology | Description | What you learn |
|---|---|---|
| Genomics | High throughput DNA sequencing of genomic DNA | which genes are present in an organism, and which alleles |
| Transcriptomics | High throughput sequencing of RNA populations | which genes are transcribed in a tissue, cell type, or in response to environmental stimuli |
| Proteomics | Identification of proteins in a protein population using mass spectrometry of oligopeptides | which proteins are expressed |
| Metabolomics | Identification of metabolites in cells | which biochemical products are produced, allowing inference of biochemical pathways |

Example: What genes are involved in a disease-resistance reaction?




Identify genes showing differential expression between treatments, either from RNAseq data, and search GenBank for the homologous genes of known function. Hope to see different transcripts are present in resistant plant, vs susceptible vs sham inoculated


Resistant
Genes uniquely expressed in resistant plants
Genes occurring under all conditions
1
45
23
585
771
98
675
Sham
Genes expressed only in susceptible plants
Susceptible

# RNA-seq (RNA sequencing)

Another approach to measuring gene expression:
**Just sequence the entire mRNA population!**

NGS now lets us do millions of reads at a reasonable price.

Why not just sequence a few million cDNAs?
In principle, **the number of transcripts found for each gene should be a good measure of the relative levels of mRNA for each gene**.

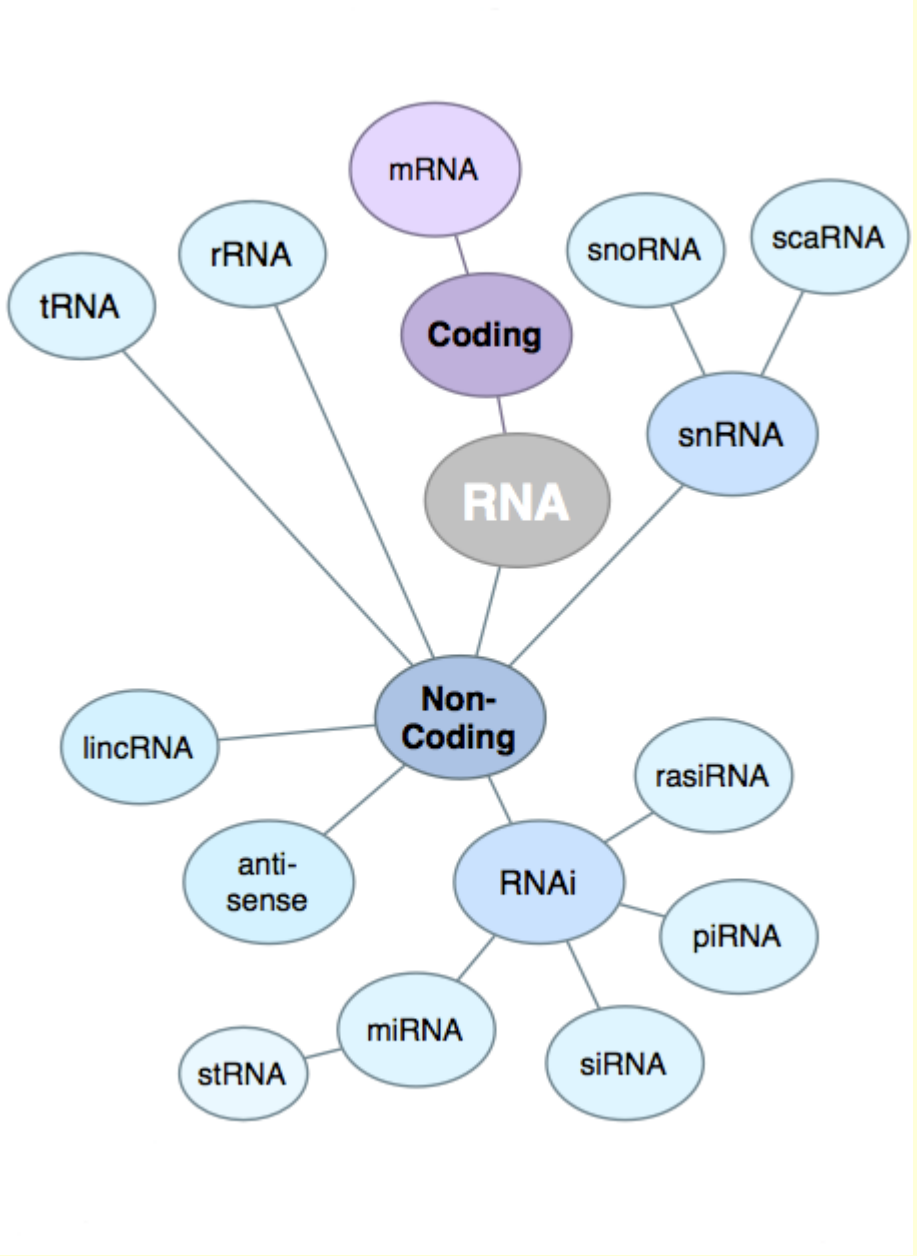
Shouldn't it?

http://finchtalk.geospiza.com/2009/05/small-rnas-get-smaller.html

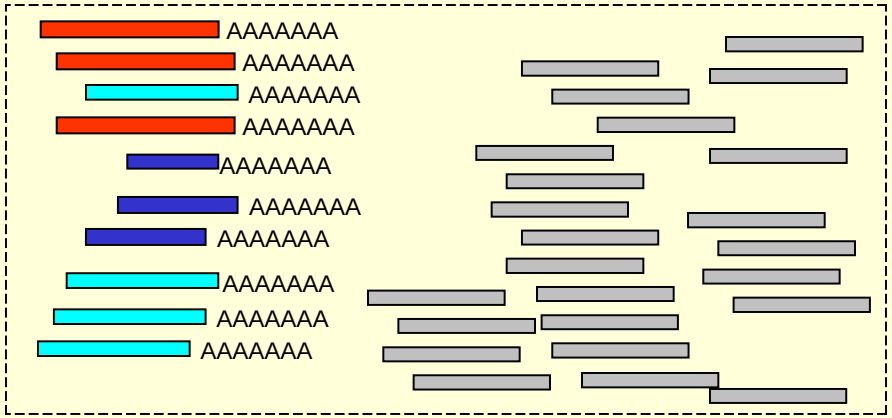
# RNA-seq (RNA sequencing)

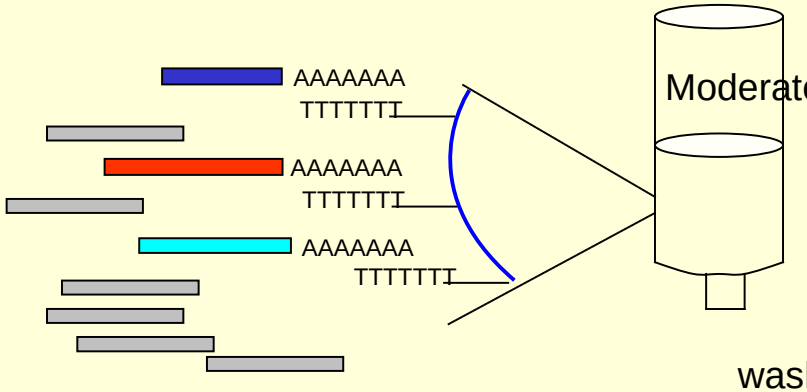
While in general the quality of the RNA is important to the success of RNA-seq experiments, the parameter that has the most effect is the **degree to which the sample has been enriched for mRNA**, by eliminating other RNAs. Especially in Eukaryotic RNA populations, mRNA usually makes up only a few percent of the total, which is predominantly rRNA. If no enrichment procedure was done, the depth of coverage of protein coding genes would be greatly compromised, because the vast majority of reads would be rRNA.

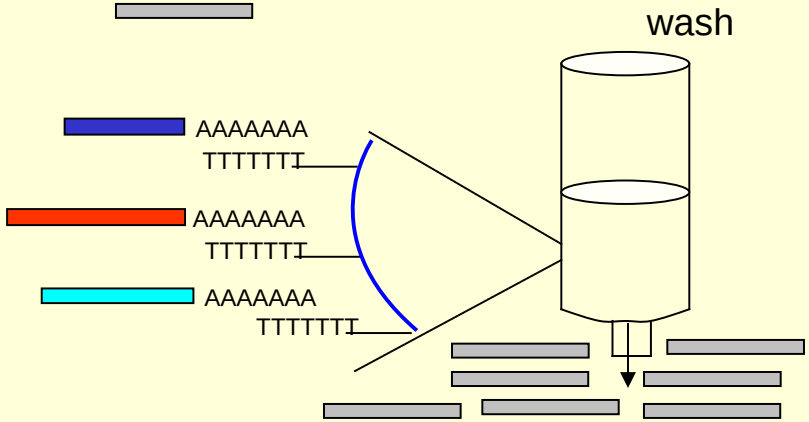
Although most RNA-seq library preparation protocols have a step for enriching for mRNAs, there will always be contamination from other RNAs. For this reason, it is important that there be a step in the data pipeline to eliminate reads that can be identified as other forms of RNA, such as rRNA or tRNA.

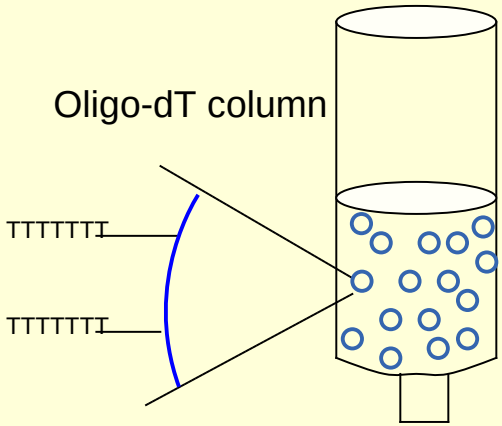


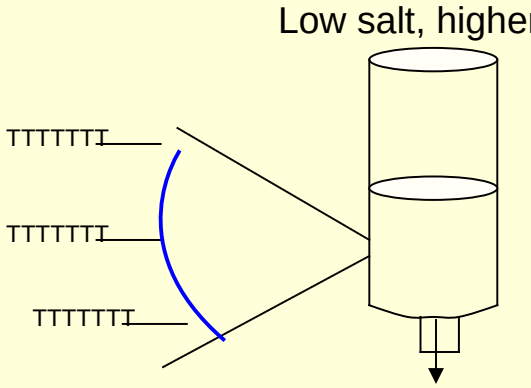

http://finchtalk.geospiza.com/2009/05/small-rnas-get-smaller.html

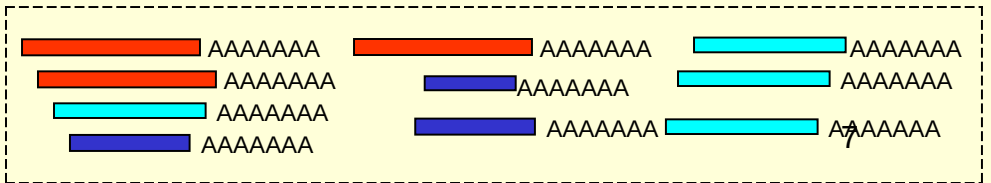

AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA


AAAAAAA
TTTTTTT
AAAAAAA
TTTTTTT
AAAAAAA
TTTTTTT
Moderate salt


wash
AAAAAAA
TTTTTTT
AAAAAAA
TTTTTTT
AAAAAAA
TTTTTTT


Oligo-dT column
TTTTTTT
TTTTTTT


Low salt, higher T
TTTTTTT
TTTTTTT
TTTTTTT

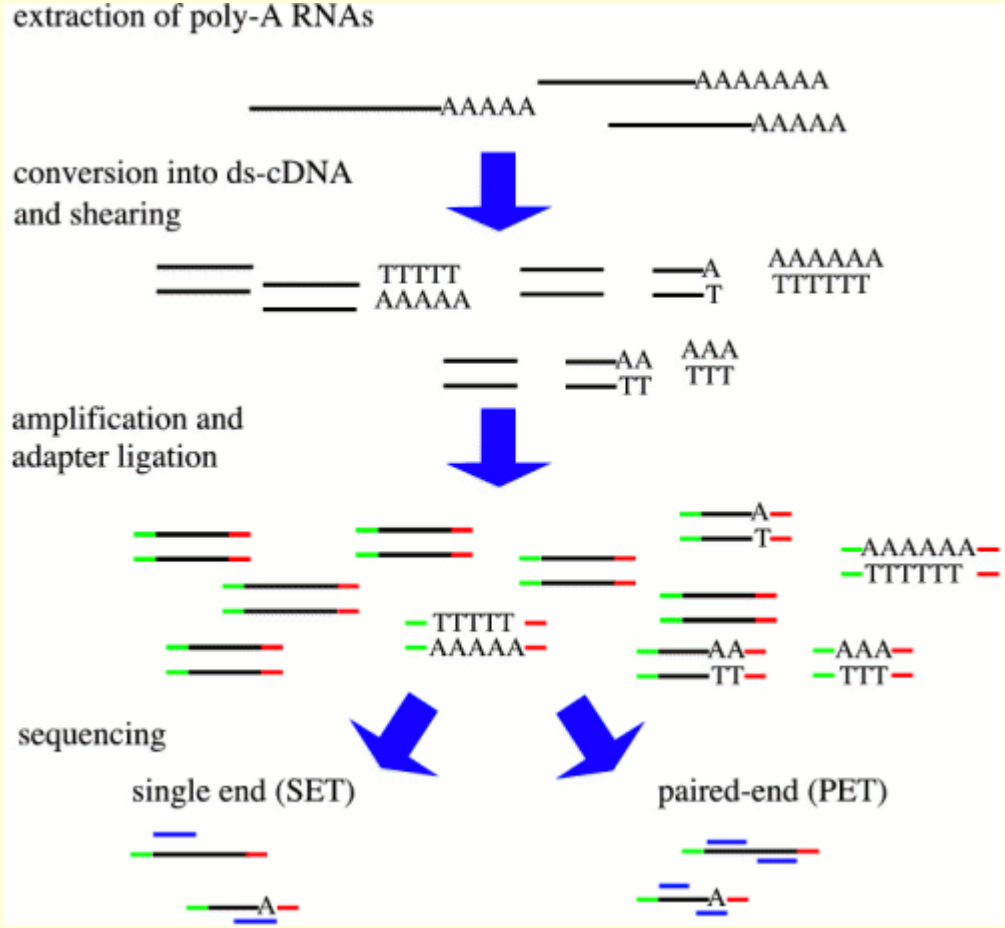

Eluted PolyA RNA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA
AAAAAAA

# RNA-seq - General strategy

There are many protocols for RNA sequencing, including Illumina GA/HiSeq, SOLiD, and Roche 454. Although these differ, the RNA-seq can be described generally as shown at right.

In some protocols, RNA is sheared, followed by random hexamer priming. In other protocols, the entire mRNA transcript is used as a template for cDNA synthesis, and the cDNA is fragmented.

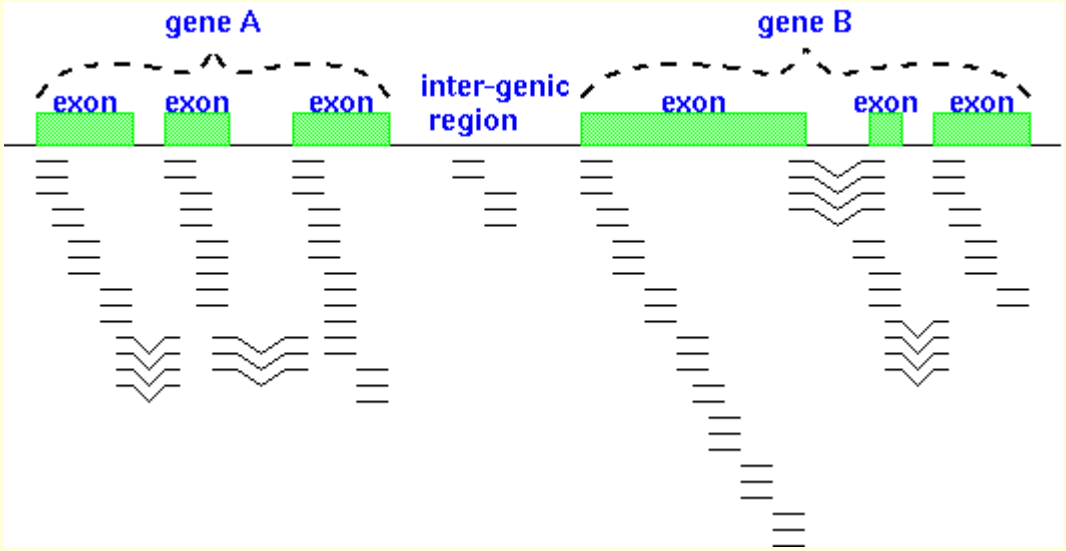
Adapters for PCR are ligated onto ds-cDNA, followed by PCR amplification. Sequencing reactions are either done from a single end, or for both ends (paired-end).




http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/RNA-seq.html

# RNA-seq - introns complicate the assembly process

The illustration at right shows RNA-seq reads aligned to two eukaryotic genes A and B. Reads that span part of an exon are shown as **single lines**, whereas reads that include parts of two adjacent exons are indicated by **V-shaped lines**. The presence of introns being spliced out of pre-mRNA transcripts means that alignment programs have to check to see whether a read contains part of the 3' end of one intron and part of a 5' end of another intron. We have to already have the genomic sequence to do this.




It is also noteable that most transcriptomics experiments contain reads that map in presumably **non-coding intergenic regions**. These can either indicate that there are previously-unannotated transcribed regions in the genome, or the presence of untranscribed pseudogenes.

Transcriptomics is also revealing that **alternative spllicing** occurs more frequently in eukaryotic gene expression than was previously appreciated.

# RNA-seq - Normalization

As shown in the illustration above, **more reads will be found for larger genes than for smaller genes**. In other words, we want to find out the number of reads or fragments that were mapped to each gene in the genome or transcriptome. Consequently, it is necessary to correct gene expression levels for

- The **size of each gene**.
- The total **number of reads in the dataset**. This makes results comparable across experiments

Depending on whether you are doing single reads or paired-end reads, there are two almost identical formulae.

# RNA-seq - Normalization

Depending on whether you are doing single reads or paired-end reads, there are two almost identical formulae.

RPKM - Reads Per Kilobase of transcript per Million mapped reads

**RPKM = C/LN**

where
C :number of mappable reads on a feature = #reads for single-end reads
L : Length of feature (in kb)
N: Total number of mappable features (in millions)

FPKM - Fragments Per Kilobase of transcript per Million mapped reads

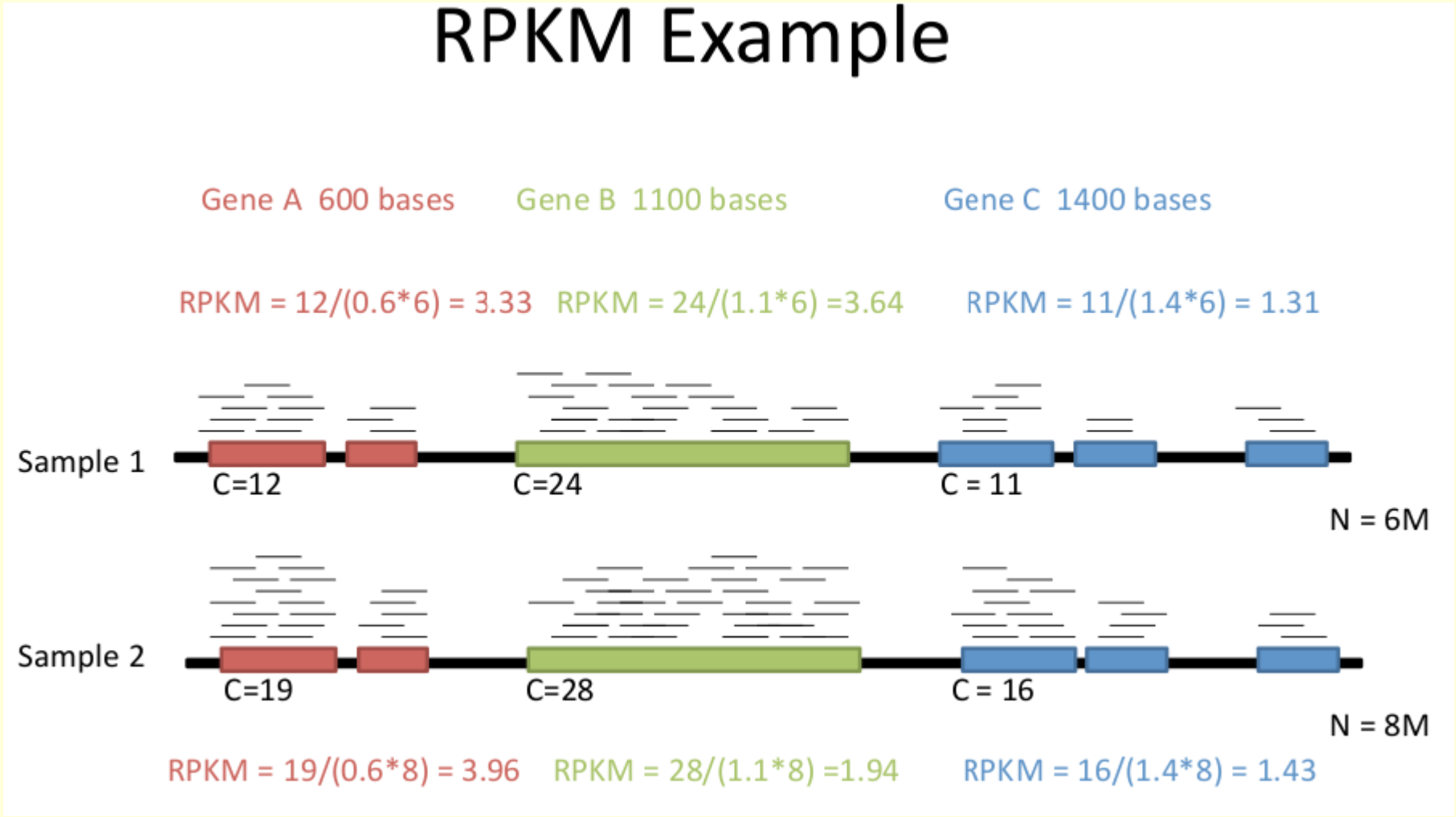
**FPKM = F/LN**

where
F :number of mappable reads on a feature* = #reads/2 for paried end reads
L : Length of feature (in kb)

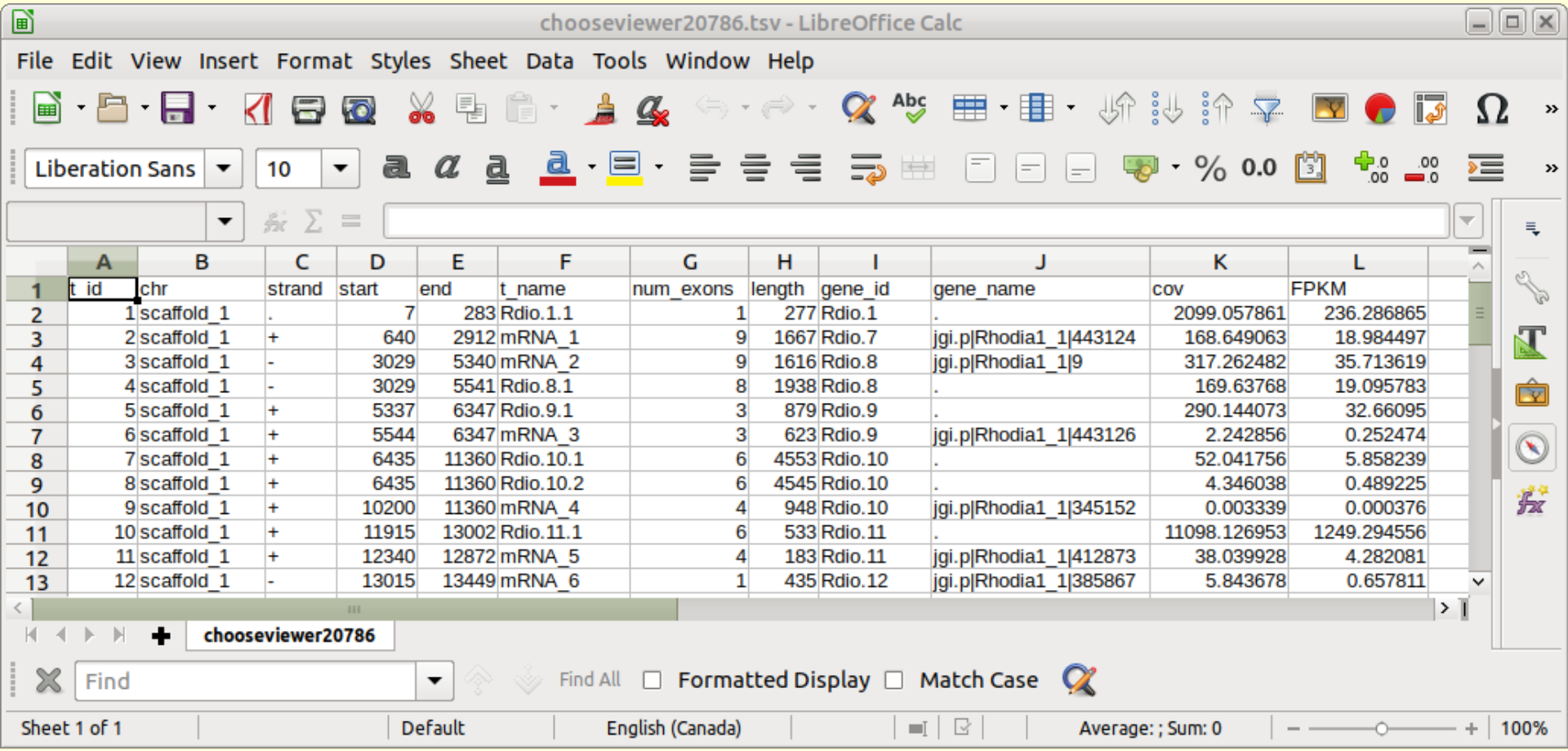
N: Total number of mappable features (in millions)

*feature - a small contig representing a particular mRNA

# RNA-seq (RNA sequencing)



Results are in the form of a spreadsheet, indicating the level of signal seen for each gene.

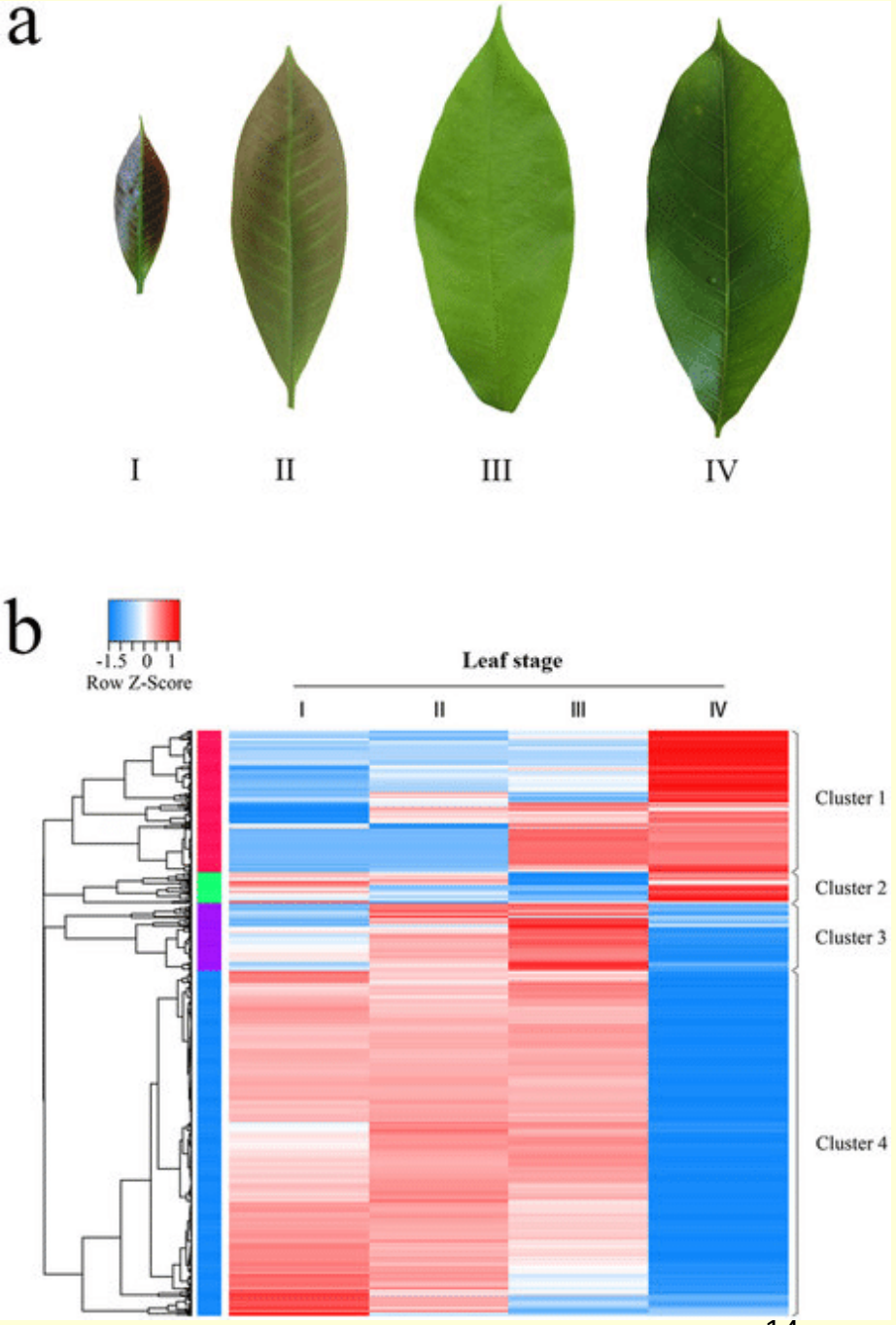



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | t_id | chr | strand | start | end | t_name | num_exons | length | gene_id | gene_name | cov | FPKM |
| 2 | 1 | scaffold_1 | . | 7 | 283 | Rdio.1.1 | 1 | 277 | Rdio.1 | . | 2099.057861 | 236.286865 |
| 3 | 2 | scaffold_1 | + | 640 | 2912 | mRNA_1 | 9 | 1667 | Rdio.7 | jgi.p\|Rhodia1_1\|443124 | 168.649063 | 18.984497 |
| 4 | 3 | scaffold_1 | - | 3029 | 5340 | mRNA_2 | 9 | 1616 | Rdio.8 | jgi.p\|Rhodia1_1\|9 | 317.262482 | 35.713619 |
| 5 | 4 | scaffold_1 | - | 3029 | 5541 | Rdio.8.1 | 8 | 1938 | Rdio.8 | . | 169.63768 | 19.095783 |
| 6 | 5 | scaffold_1 | + | 5337 | 6347 | Rdio.9.1 | 3 | 879 | Rdio.9 | . | 290.144073 | 32.66095 |
| 7 | 6 | scaffold_1 | + | 5544 | 6347 | mRNA_3 | 3 | 623 | Rdio.9 | jgi.p\|Rhodia1_1\|443126 | 2.242856 | 0.252474 |
| 8 | 7 | scaffold_1 | + | 6435 | 11360 | Rdio.10.1 | 6 | 4553 | Rdio.10 | . | 52.041756 | 5.858239 |
| 9 | 8 | scaffold_1 | + | 6435 | 11360 | Rdio.10.2 | 6 | 4545 | Rdio.10 | . | 4.346038 | 0.489225 |
| 10 | 9 | scaffold_1 | + | 10200 | 11360 | mRNA_4 | 4 | 948 | Rdio.10 | jgi.p\|Rhodia1_1\|345152 | 0.003339 | 0.000376 |
| 11 | 10 | scaffold_1 | + | 11915 | 13002 | Rdio.11.1 | 6 | 533 | Rdio.11 | . | 11098.126953 | 1249.294556 |
| 12 | 11 | scaffold_1 | + | 12340 | 12872 | mRNA_5 | 4 | 183 | Rdio.11 | jgi.p\|Rhodia1_1\|412873 | 38.039928 | 4.282081 |
| 13 | 12 | scaffold_1 | - | 13015 | 13449 | mRNA_6 | 1 | 435 | Rdio.12 | jgi.p\|Rhodia1_1\|385867 | 5.843678 | 0.657811 |

Example: Changes in gene expression during leaf development in rubber tree

RNA was extracted from leaves at four different stages of development. The RNA was used to measure gene expression.

Each line in the “heat map” at right represents a different gene. Only a small number of the genes are shown. The level of expression for each gene in the leaf, is represented using the color scale shown in b.




Fang, Yongjun & Mei, Hailiang & Zhou, Binhui & Xiao, Xiaohu & Yang, Meng & Huang, Yacheng & Long, Xiang-Yu & Hu, Songnian & Tang, Chaorong. (2016). De novo Transcriptome Analysis Reveals Distinct Defense Mechanisms by Young and Mature Leaves of Hevea brasiliensis (Para Rubber Tree). Scientific Reports. 6. 33151. 10.1038/srep33151.

# Proteomics

A family of methods for identifying and quantifying proteins which are expressed under specific conditions.

- Older methods involved isolation of proteins by 2D electrophoresis, but more recent approaches employ more high-throughput methods
- There are many different strategies. We will only talk about one method of the global strategy for studying the entire population of proteins in cells or tissues.

# Proteomics

General procedure:

1 Isolate total protein population from cells (thousands of different proteins)

2 Use proteases to digest proteins into smaller peptides

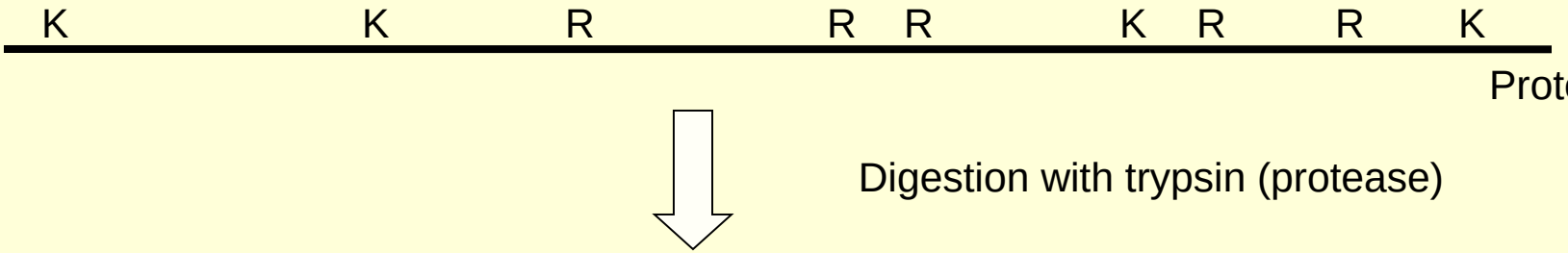
3 Separate the peptides into size classes by liquid chromatography

4 As fractionated peptides come off the LC column, they are fed into a mass spectrometer to precisely measure their molecular weights, and to quantify how much of each oligopeptide is present.

5 Oligopeptides are compared with a database of proteins to identify which proteins were present in the cell population.

# Proteomics

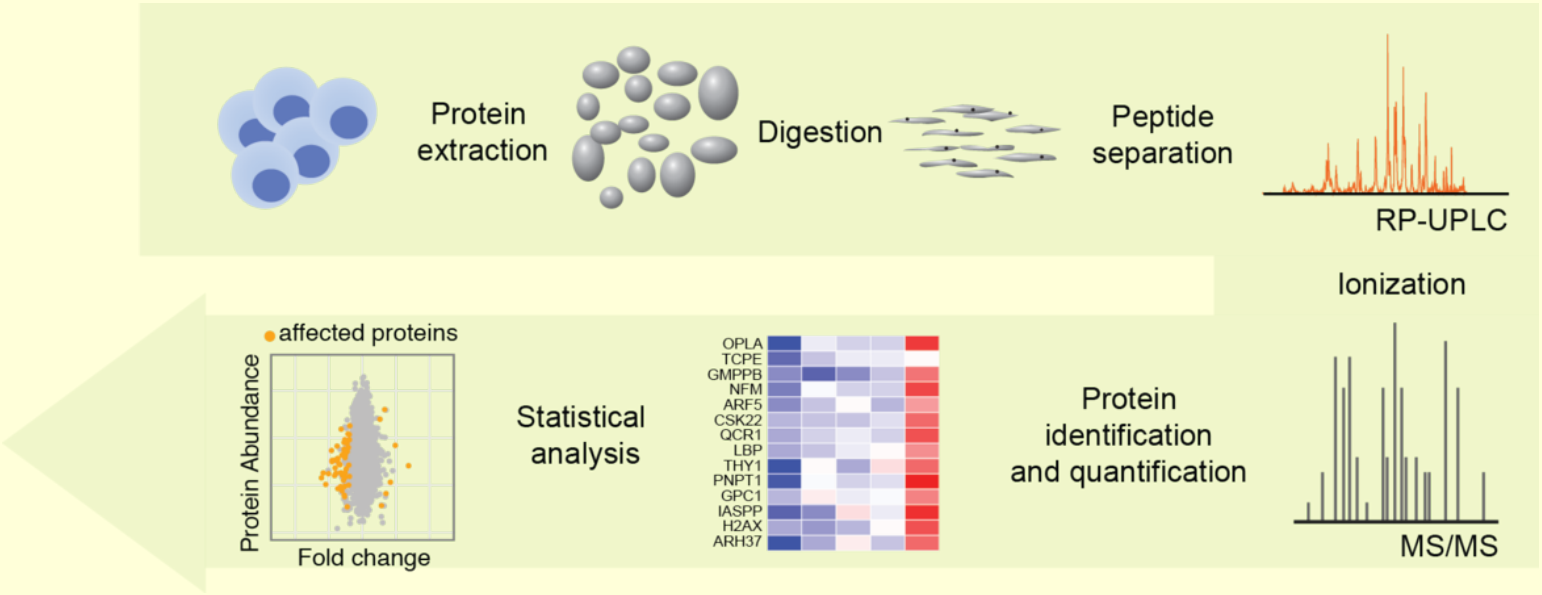
- Whole proteins are too large for mass spectrometry so each protein band is first digested with an enzyme called trypsin which cleaves the polypeptide after every lysine (K) and arginine (R) to yield smaller peptides




Because each protein has a unique sequence, there will be a unique series of peptides generated from trypsin digestion.

The mixture is mass analyzed and identified if the sequences are in the database

# Proteomics



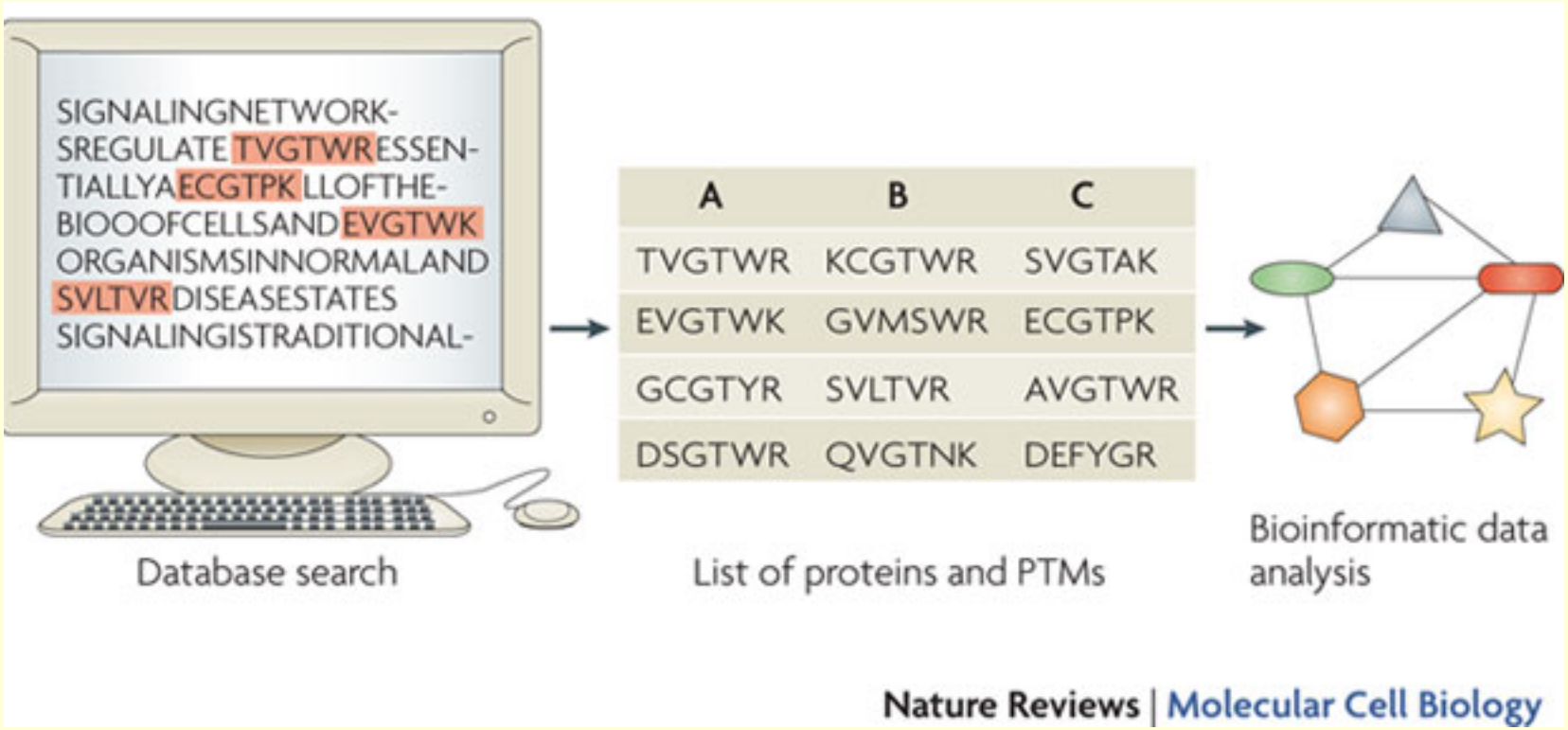

from http://www.leibniz-fli.de/research/core-facilities-services/cf-proteomics/

# Proteomics for measuring gene expression

1) Since each possible oligopeptide has a uniuqe profile on MS, the exact sequence of the oligopeptide can be identified, based on its profile.

2) By comparing the oligopeptides identified with known proteins in a database, the identities of each protein in the population can be determined.

3) Because MS is highly accurate, we also get a measurement of the amount of each protein.




from http://www.nature.com/nrm/journal/v11/n6/fig_tab/nrm2900_F2.html

# Metabolomics

- Similarly many of the proteins in a cell are enzymes which catalyze reactions leading to a large set of products.

- The set of primary and secondary products produced from the enzymes of metabolism in the cell at any instance is referred to as the metabolome.

    - primary metabolism:  eg amino acids, sugars, nucleotides

    - Secondary metabolism:  eg. carotenoids, phenolic cmps

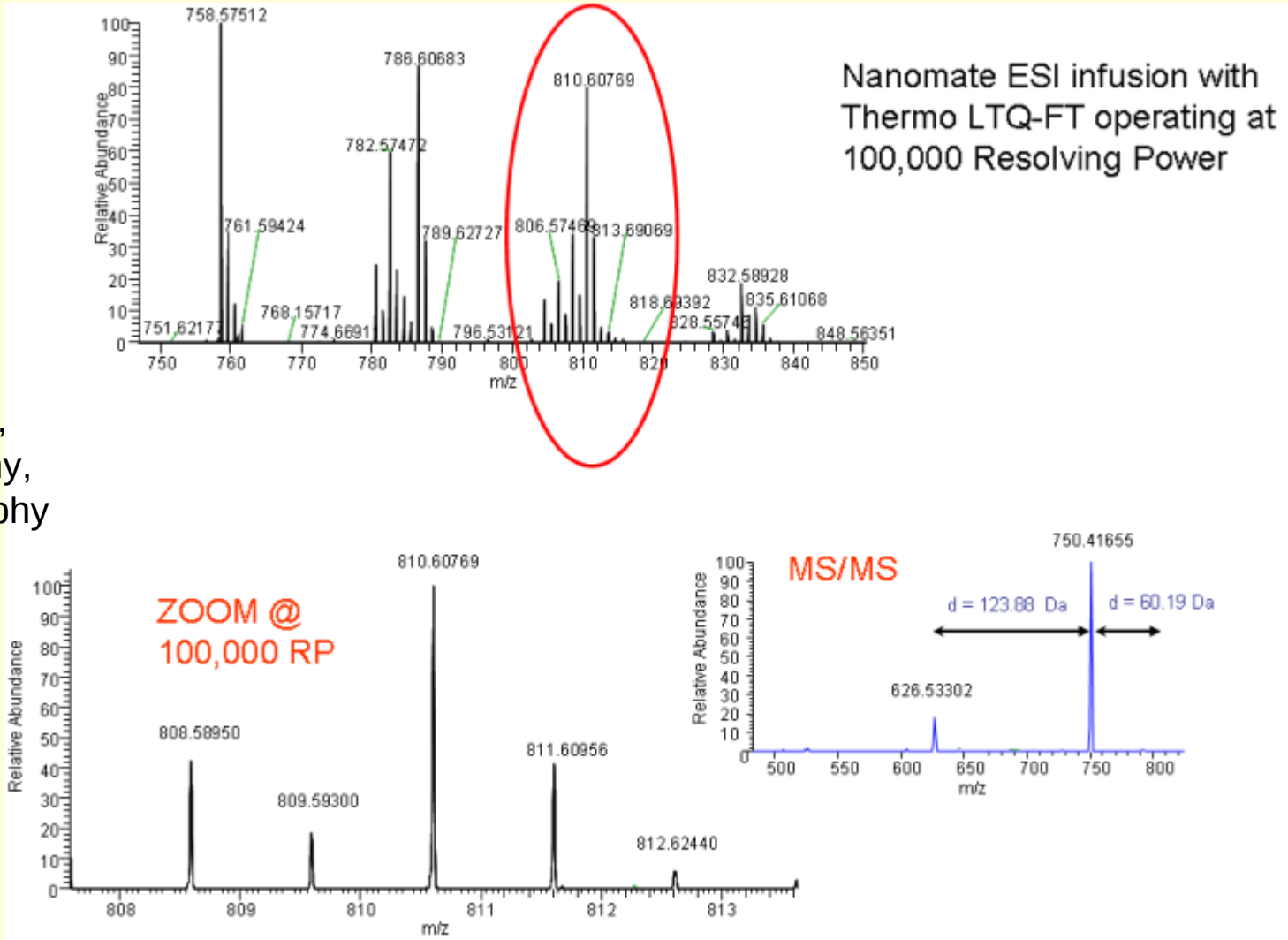
10,000+  metabolites in a cell at any instant

**Metabolomics**: A family of techniques for separating metabolites
from a cell based on charge, molecular weight, and other characteristics.

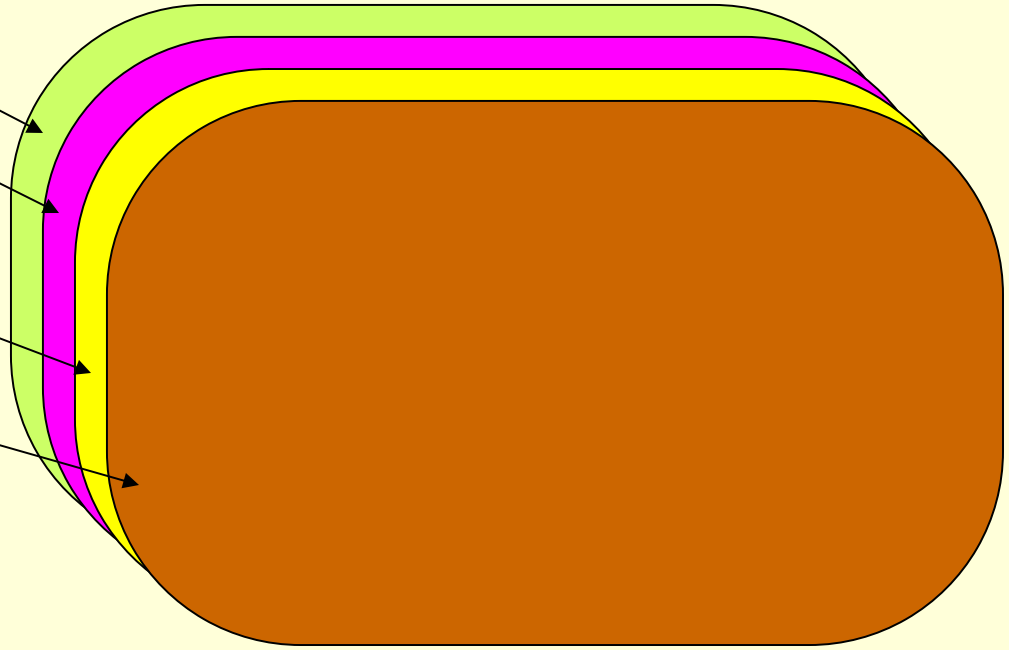
# Metabolomics

Each type of molecule has a distinct set of characteristics.

Profiles from methods such as Mass spectrometry, gas chromatography, liquid chromotography are compared to a database of molecules, to identify each metabolite.



- Genome
- Transcriptome
- Proteome
- Metabolome



Information slices of the same cells at the same time under the same conditions