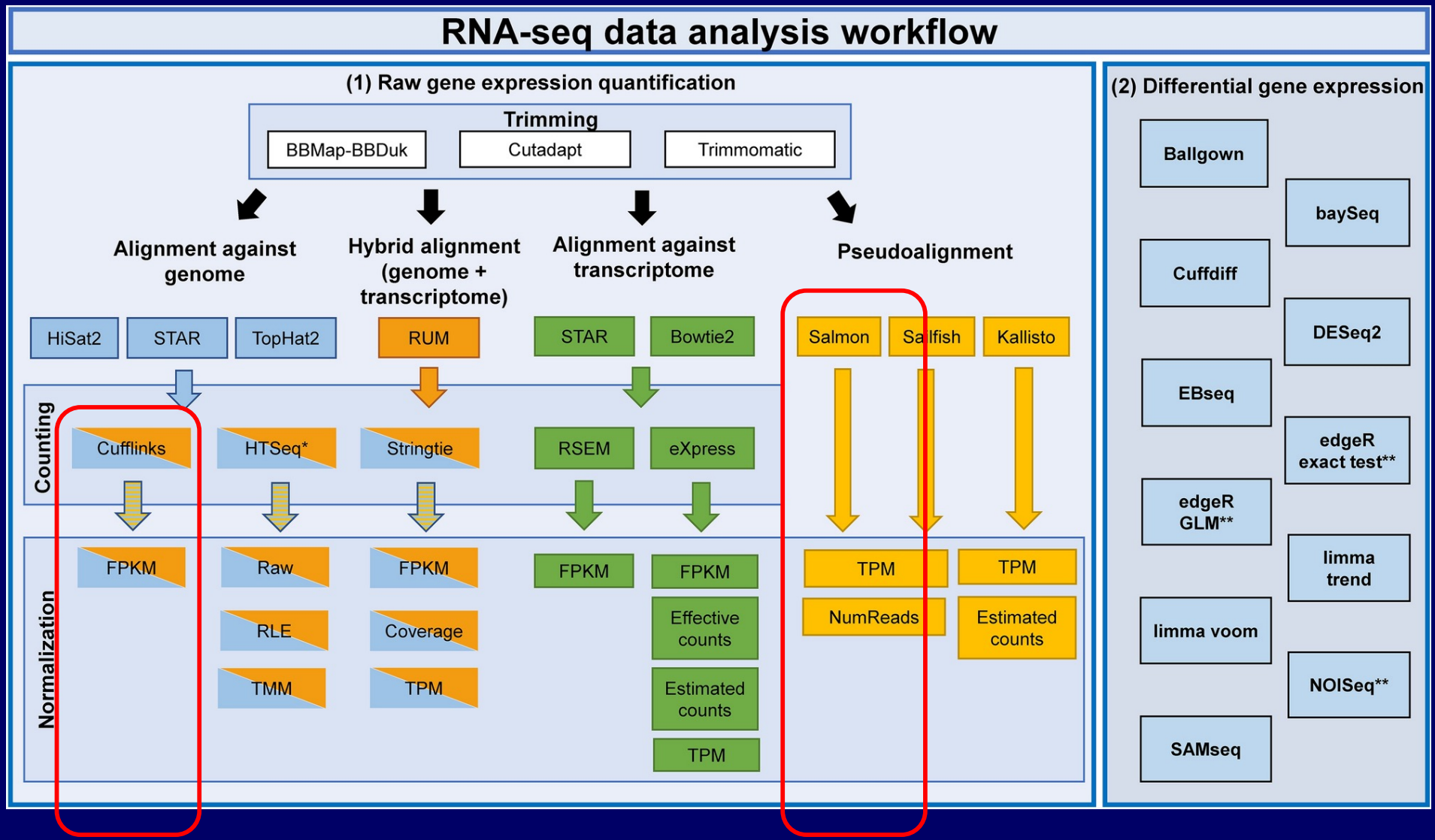
RNA-seq: Getting counts



RNA-Seq: Getting counts

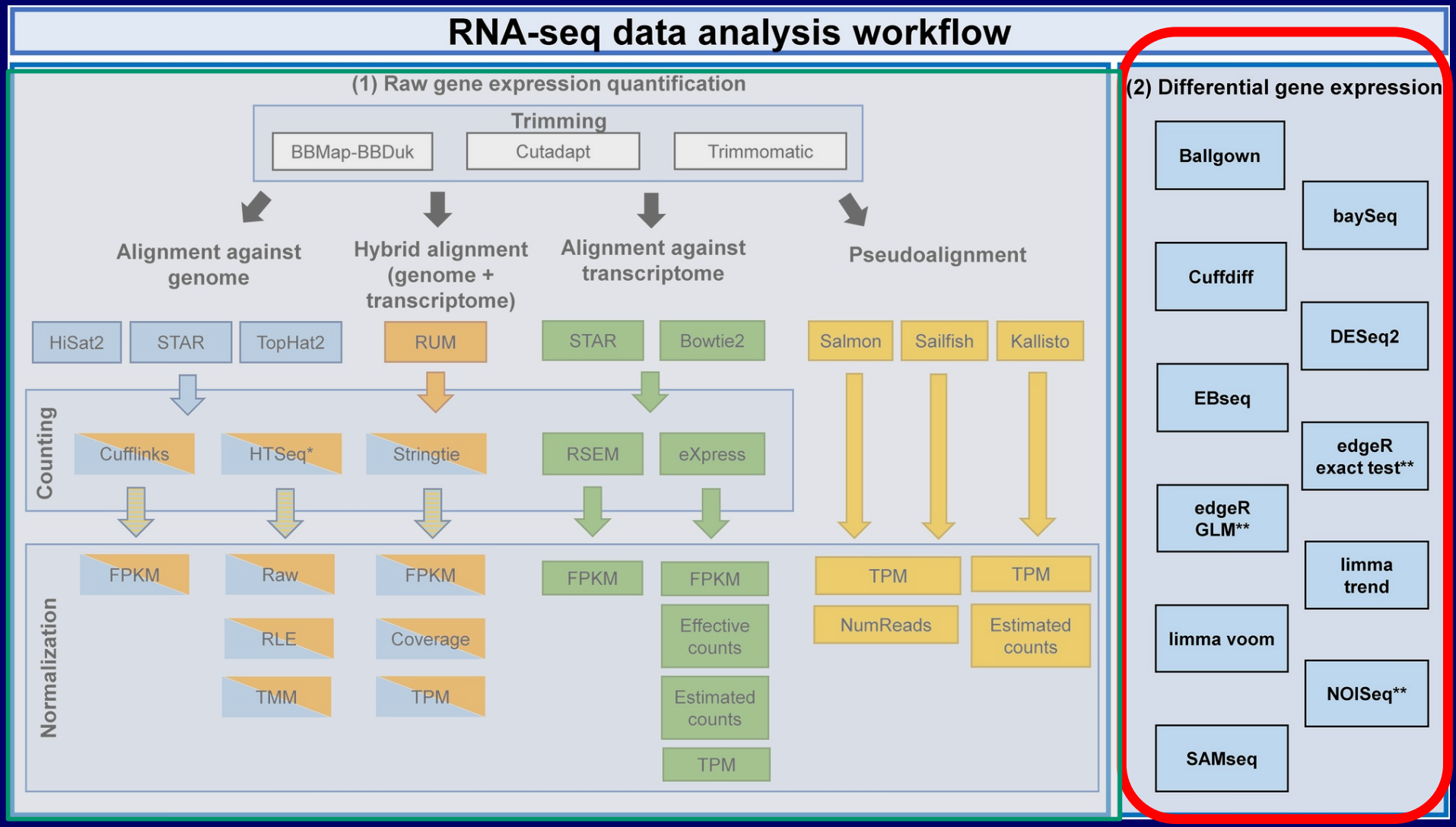
- Raw – counts (reads) per gene.
- Normalized
  - FPKM (Fragments Per Kilobase gene length and per Million reads)
  - TPM (Transcripts Per Million)
- Depending on the which program will be used for identifying DEGs.
  - DESeq (DESeq2) requires raw counts
  - CuffLinks generated normalized counts as well as models for CuffDiff.

RNA-Seq Overview

Four major steps, semi-independent of each other.

I. Mapping → produce SAM/BAM or counts data.
II. Quantification → produce RPKM for each gene/transcript.
III. Identifying DEG (Differentially expressed genes) → gene list.

RNA-seq: Identify DEGs



Many options at this stage. Personal favorites – Cuffdiff and DESeq2

Identification of Differentially Expressed Genes (DEGs)

```
module load cufflinks

## Frist merge the gtf files for samples to be compared.
ln /ufrc/gms6014/share/genome/dm6/annotation/genes.gtf dm6.gtf
ln /ufrc/gms6014/share/genome/dm6/sequence/genome.fa dm6.fa

cuffmerge -g dm6.gtf -s dm6.fa -p 2 WG_assemblies.txt
```

./WG_young_1.clout/transcripts.gtf
./WG_young_2.clout/transcripts.gtf
./WG_old_1.clout/transcripts.gtf
./WG_old_2.clout/transcripts.gtf

Identification of Differentially Expressed Genes (DEGs)

```
module load cufflinks

## Frist merge the gtf files for samples to be compared.
ln /ufrc/gms6014/share/genome/dm6/annotation/genes.gtf dm6.gtf
ln /ufrc/gms6014/share/genome/dm6/sequence/genome.fa dm6.fa

cuffmerge -g dm6.gtf -s dm6.fa -p 2 WG_assemblies.txt
```

./WG_young_1.clout/transcripts.gtf
./WG_young_2.clout/transcripts.gtf
./WG_old_1.clout/transcripts.gtf
./WG_old_2.clout/transcripts.gtf

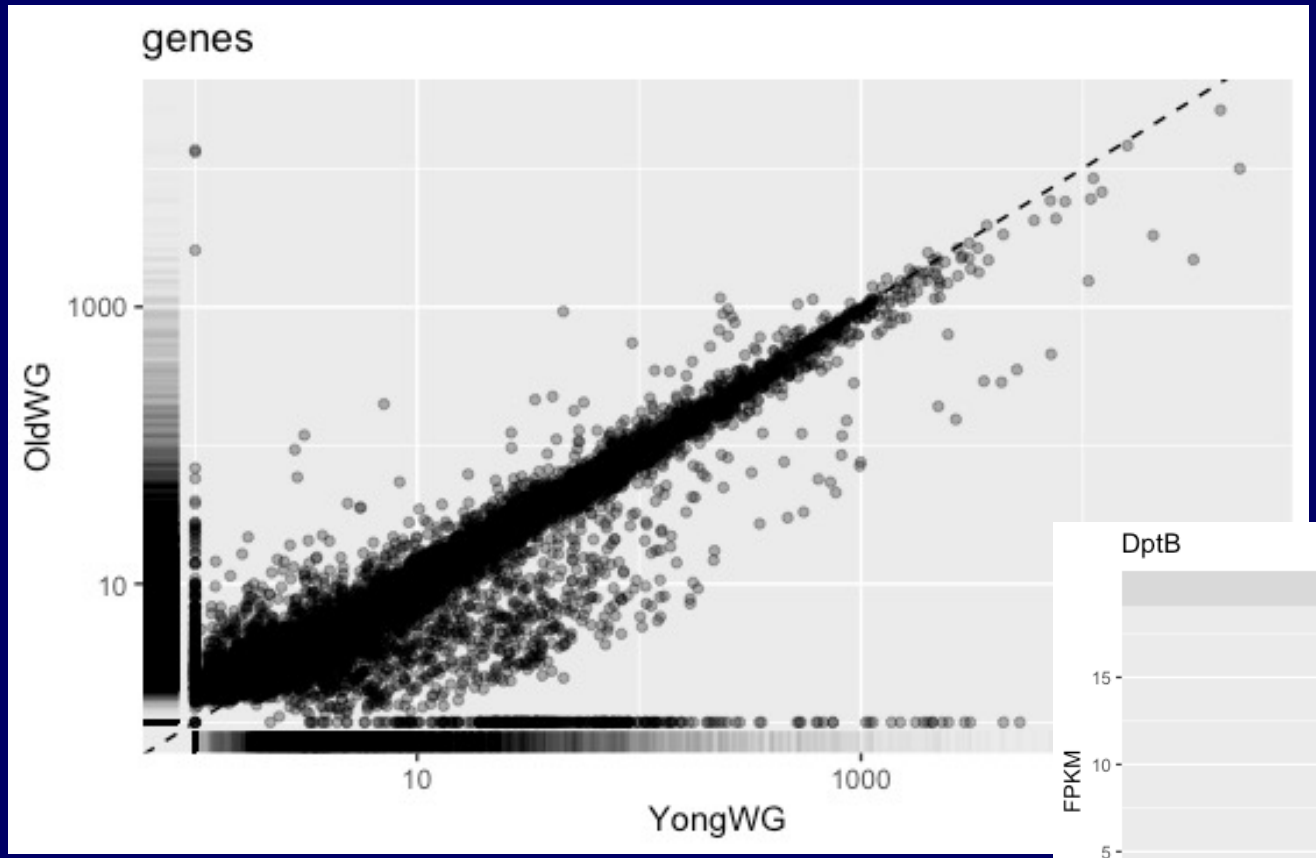
Identification of differentially expressed genes (DEGs)

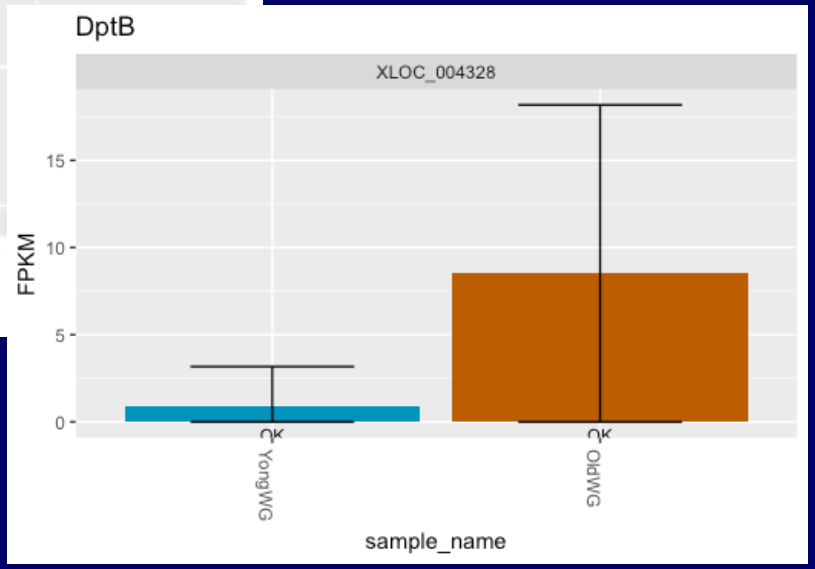
```
module load cufflinks

cuffdiff -o Old_v_Young -b ./index/Dm6.44.fa  -u Merged/merged.gtf  -p 2 -L youngWG,oldWG \
 ./starMap/WG_young_1Aligned.sortedByCoord.out.bam,./starMap/WG_young_2Aligned.sortedByCoord.out.bam \
 ./starMap/WG_old_1Aligned.sortedByCoord.out.bam,./starMap/WG_old_2Aligned.sortedByCoord.out.bam
```

Practice – observe CuffDiff results

- Transfer the CuffDiff result folder ("`Old_v_Young`") from HiPerGator to your own computer.
- Observe (you may force it to be opened by Excel by adding .xlxs extension):
  - Gene_exp.diff
  - Gene_exp.tracking.
- Further explore of results with the R package "cummeRbund" .

Observe CuffDiff results with cummeRbund





RNA-Seq Overview

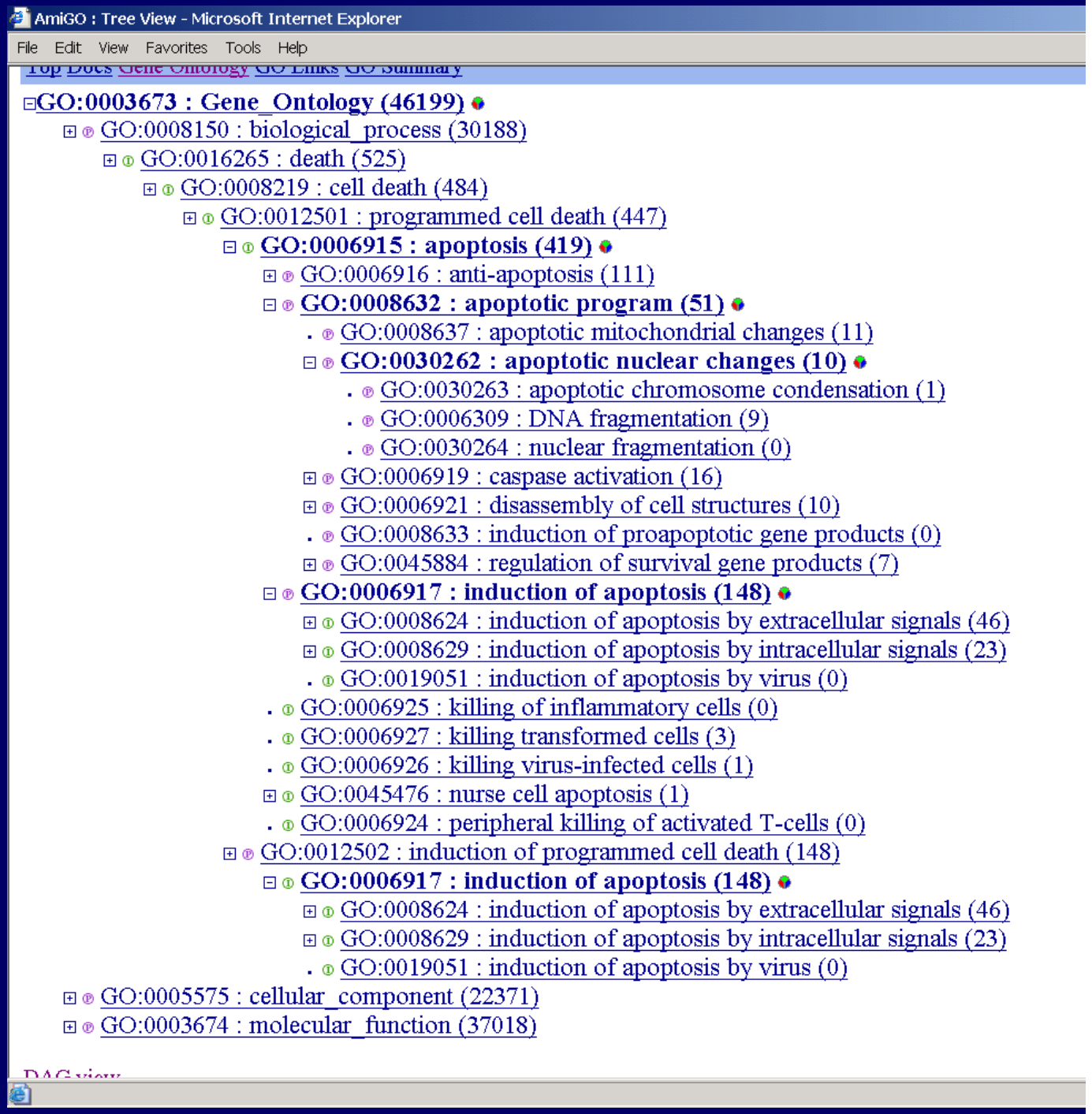
Four major steps, semi-independent of each other.

I. Mapping → produce SAM/BAM or counts data.
II. Quantification → produce RPKM for each gene/transcript.
III. Identifying DEG (Differentially expressed genes) → gene list.
IV. Identifying affected biological processes/pathways.

Functional Analysis of HTS data

- Gene Ontology – http://www.geneontology.org/
- Regulatory pathways.
- Modeling & Systems Biology.

Gene Ontology – hierarchical framework of terms / concepts



Gene Ontology

Goal – "produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing" – GO consortium (~2001)

Ontology:

" The branch of metaphysics that deals with the nature of being" – The American Heritage Dictionary

Implications of Gene Ontology (I)

Monitoring biological processes or molecular functions beyond individual gene.

Example:

1.) Which biological process (mol. Function) is activated/suppressed following a treatment?

Gene Expression Profile Differences between the two lung cancer cell lines A549 and H23

extracellular (GO:0005576)	1.91E-08	169
Cell Communication	1.32E-07	690
plasma membrane (GO:0005886)	1.34E-07	511
Complement and coagulation cascades - Homo sapiens	1.73E-07	20
Metabolism	2.10E-06	174
carbohydrate metabolism (GO:0005975)	2.45E-06	207
cell adhesion molecule activity (GO:0005194)	0.000102	113
Structural Protein	0.000231	271
extracellular matrix (GO:0005578)	0.000235	53
Cell Growth and Maintenance	0.000569	590
Cell Adhesion	0.000917	100

development	1.40E-07	596
cell differentiation (GO:0030154)	6.60E-05	186
regulation of gene expression, epigenetic (GO:0040029)	7.71E-05	442
cell growth (GO:0016049)	8.37E-05	307
transcription regulator activity (GO:0030528)	0.000307	319
extracellular (GO:0005576)	0.000515	153

Implications of Gene Ontology (II)

Basis for cross genome comparison and integrating knowledge from different model systems.

Term	Fly Genes	Worm Genes	Mouse Genes	Human Genes	Sacc. Yeast Genes	Pombe Yeast Genes	Weed Genes
cell cycle	265	182	294	717	424	622	181
cell cycle dependent actin filament reorganization	2	0	0	0	4	0	0
DNA replication and chromosome cycle	134	128	67	175	172	73	146
endomitotic cell cycle	0	0	1	1	0	3	0
M phase	171	39	69	181	213	253	3
mitotic cell cycle	133	140	102	314	239	202	141
nuclear migration	1	0	0	0	13	0	0
regulation of cell cycle	42	4	136	383	87	65	3
schizogony	0	0	0	0	0	0	0
second mitotic wave (sensu Drosophila)	1	0	0	0	0	0	0

Tools associated with GO

- A comprehensive list at GO web site.
- Tools for browsing, AmiGO, QuickGO at EBI, etc.
- Tools for identifying over represented GOs/pathways**, etc.**

Using GO to gain comprehensive understanding of cellular differences

Practice: Load a gene list to identify over-represented GO