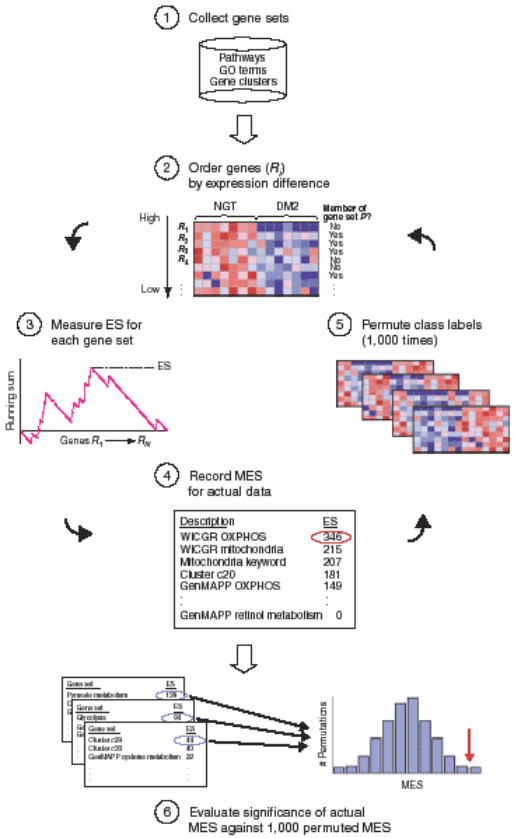
Statistics for Genomic Data Analysis

Annotation; gene set testing



http://moodle.epfl.ch/course/view.php?id=15271

Identifying differential expression

- Preprocess data
  - Image analysis
  - Quality assessment
  - Normalization
- Single gene linear modeling and empirical Bayesian shrinkage
  - lmFit
  - eBayes
  - topTable -> (possibly long) list of DE genes

Limitations of single gene tests

- If expression changes are not loare, might not be able to detect significance after controlling for multiple tests
- Difficulty *interpreting* a resulting (possibly long) gene list
- Single gene testing ignores information on *functional annotation*

Problem: relating list of DE genes to biology

- What to do with the list?
  - Select some genes for validation
  - Follow-up experiments on some genes
  - Publish a huge table with the results
  - Literature search to learn about genes on the list
- Maybe further/different data analysis will help?

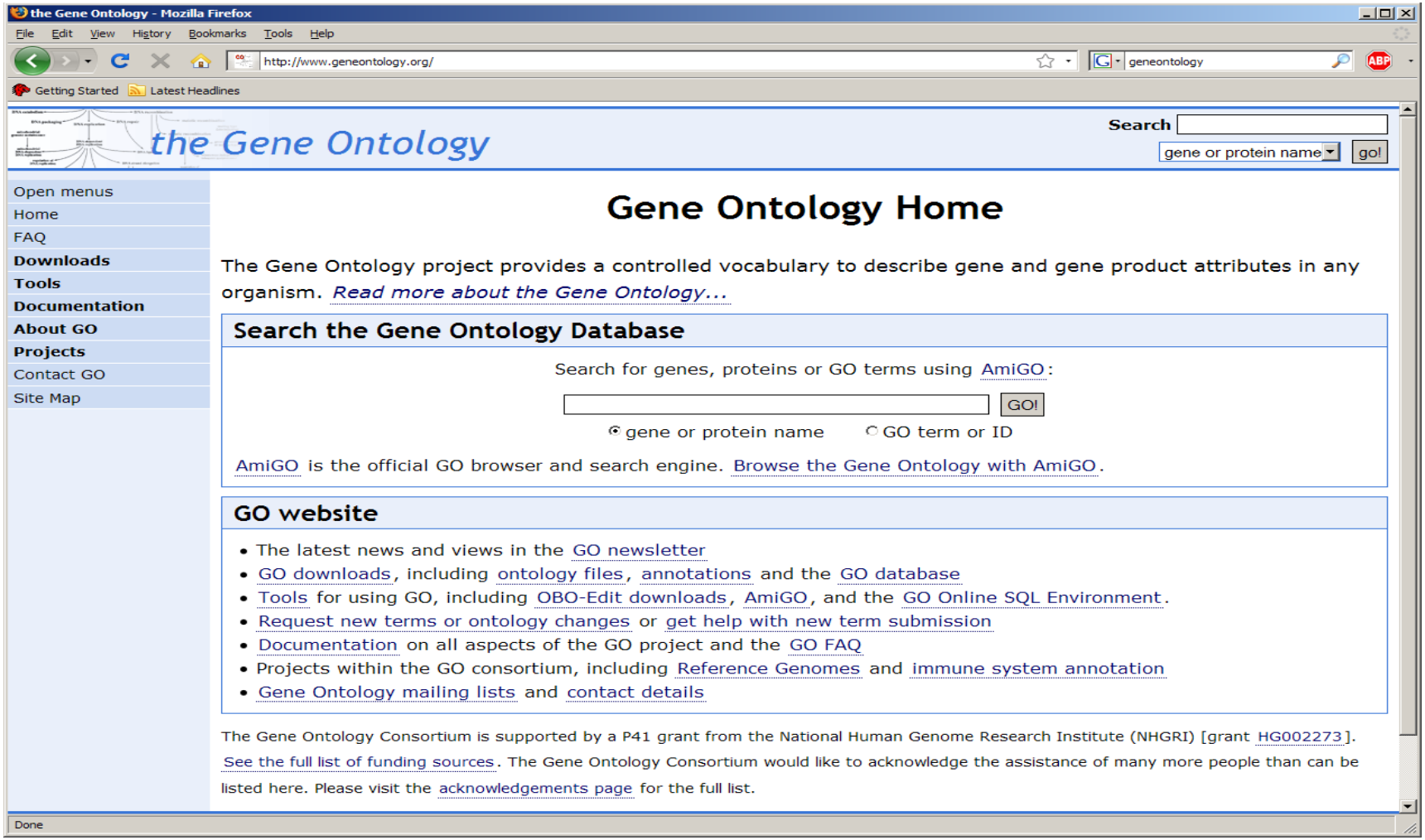
Sets of genes

- *Sets* of genes defined *a priori*
- There are usually many sets of genes of (potential) interest
  - genes in particular *pathways*
  - genes having a certain *function*
- Genes can have multiple functions
- Pathway databases (*e.g.* KEGG)
- Gene ontologies (GO)
- Protein `knowledgebase' SwissProt

The Gene Ontology Consortium

- Coordinates GO development
- GO is a set of 3 *ontologies* for gene products
  - Molecular function
  - Biological process
  - Cellular component
- An ontology is a *restricted structured vocabulary* of terms used to represent domain knowledge
- The *leaves* are more specific terms, their parents are less specific

Gene Ontology



Molecular function

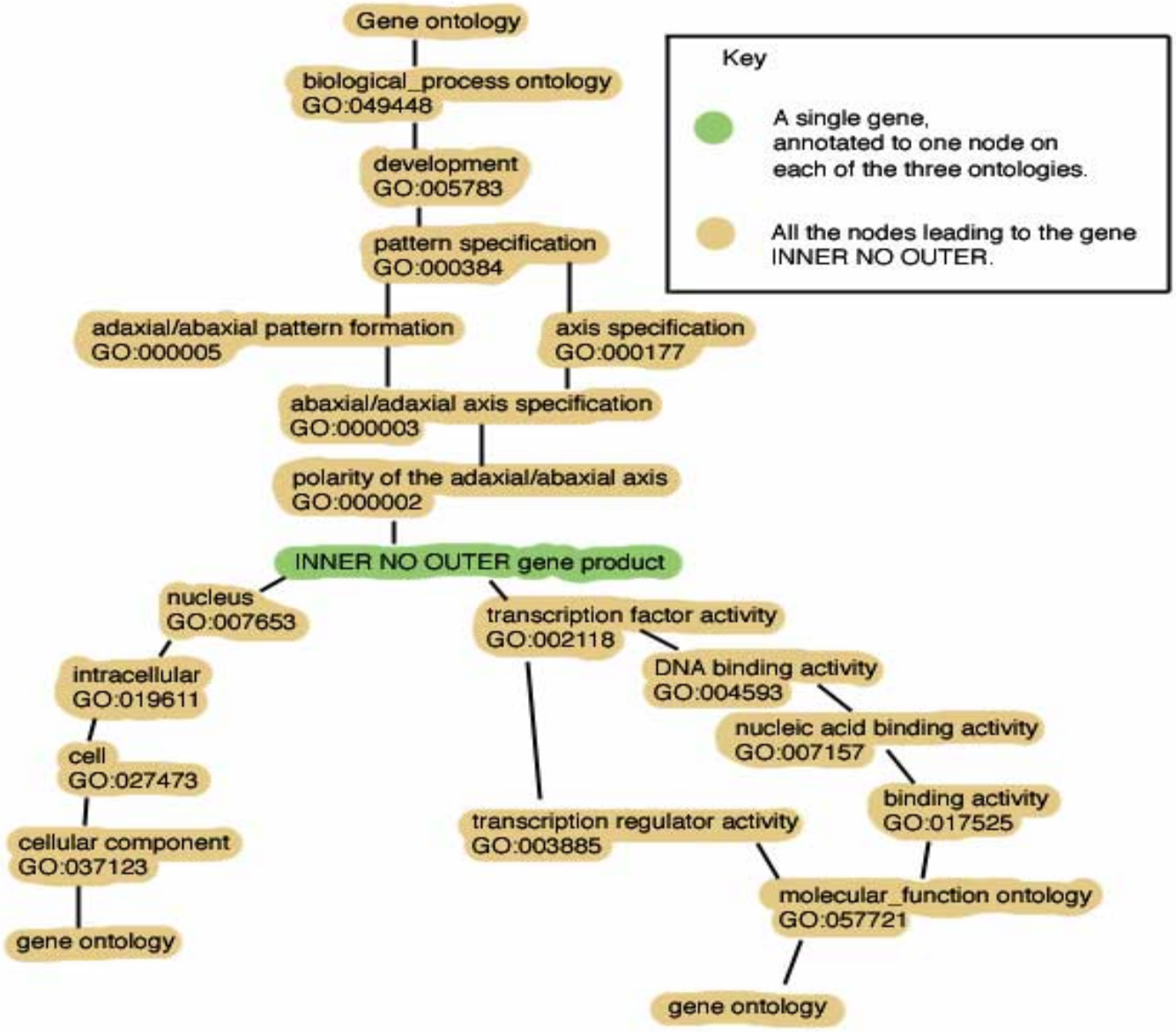
- Defined to be what a gene product does at the *biochemical level*
- Describes (only) the *capability* of the gene product
- Does not say anything about where the function is carried out, under what circumstances, how it works, *etc.*
- Examples: transporter, enzyme

Biological process

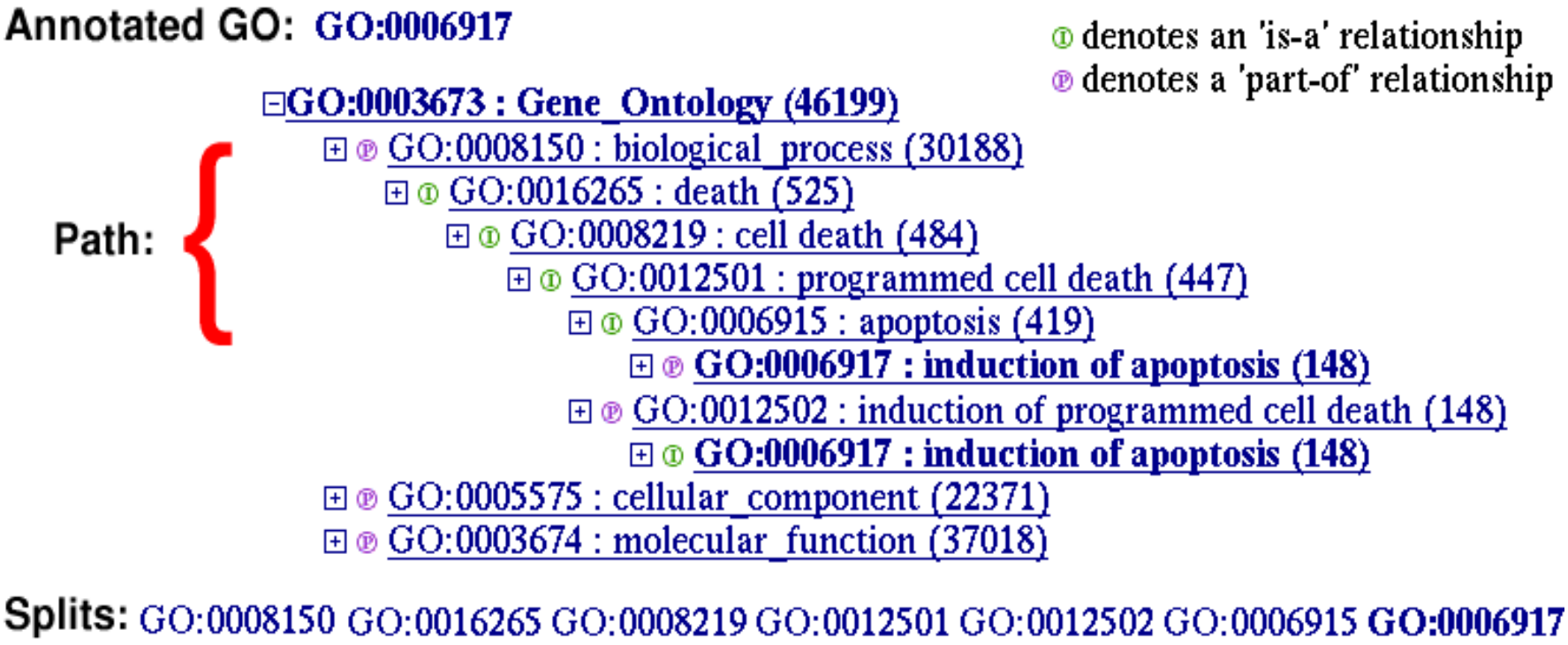
- A biological objective to which the gene product contributes
- Accomplished by assemblies of molecular functions
- Not the same as a pathway (which has dependencies and dynamics distinct from a biological process)
- Not always easy to distinguish from molecular function
- Example: signal transduction

Cellular component

- Component of a cell that is part of a larger structure
- Examples: telomere, nucleus


Gene ontology
biological_process ontology
GO:049448
development
GO:005783
pattern specification
GO:000384
adaxial/abaxial pattern formation
GO:000005
axis specification
GO:000177
abaxial/adaxial axis specification
GO:000003
polarity of the adaxial/abaxial axis
GO:000002
INNER NO OUTER gene product
nucleus
GO:007653
intracellular
GO:019611
cell
GO:027473
cellular component
GO:037123
gene ontology
transcription factor activity
GO:002118
DNA binding activity
GO:004593
nucleic acid binding activity
GO:007157
binding activity
GO:017525
transcription regulator activity
GO:003885
molecular_function ontology
GO:057721
gene ontology
Key
A single gene, annotated to one node on each of the three ontologies.
All the nodes leading to the gene INNER NO OUTER.

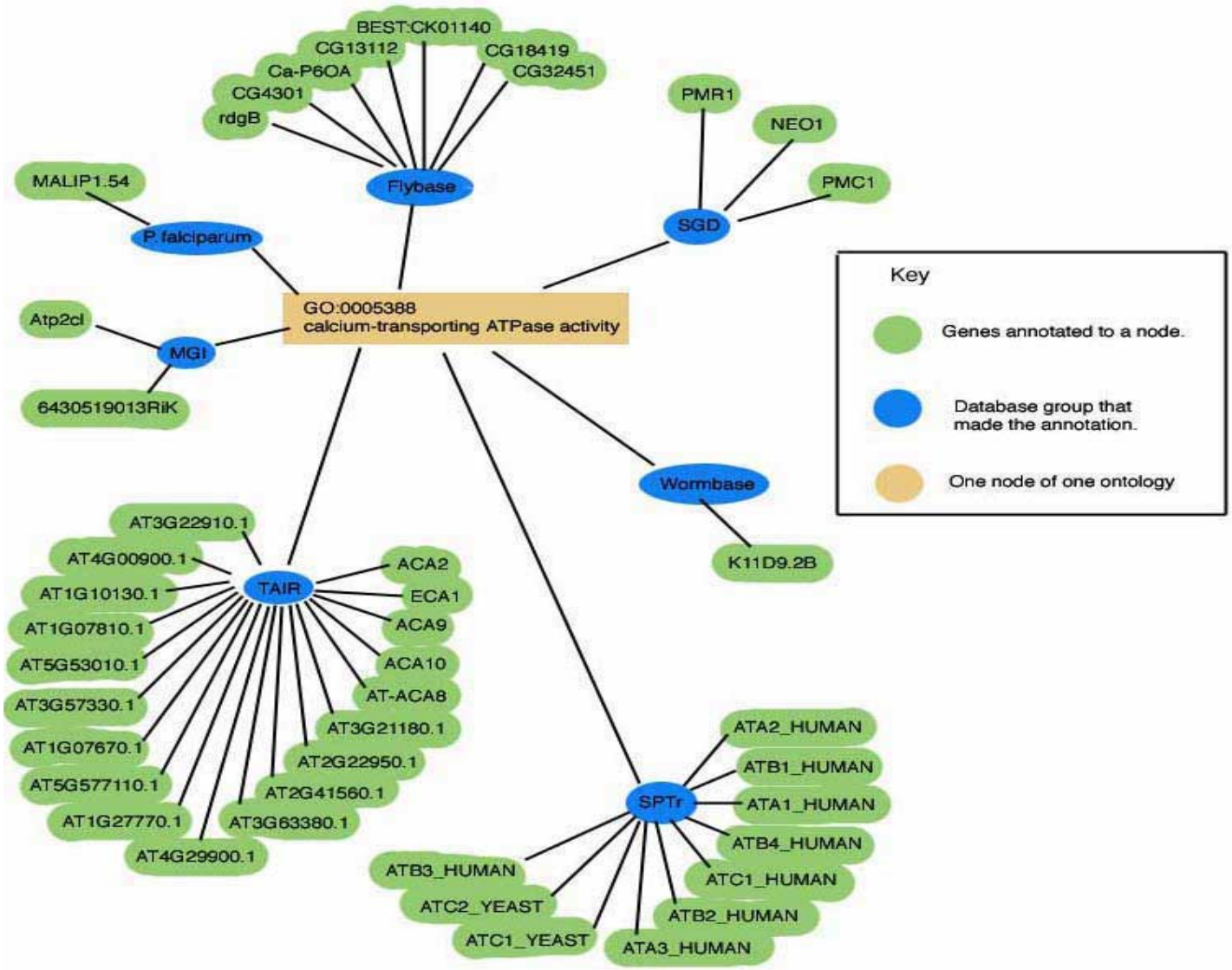
Structure of a GO annotation



Each gene can have several annotated GOs, and each GO can have several splits

Annotation of genes to a node

Each node is connected to other, related nodes



GO and microarray gene sets

Hypothesis: Functionally related, differentially expressed genes should accumulate in the corresponding GO-group

Problem: to find a method which scores accumulation of differential gene expression in a node of the GO

Strategies for gene set testing

- Hypergeometric testing (Fisher's exact test)
- Gene Set Enrichment Analysis (GSEA)
- Main difference: hypergeometric requires a definition of DE vs. not (often based on a *p*-value), whereas GSEA takes a continuous measure and computes a global summary for the gene set

A lady tasting tea

- Exact test developed for the following setup:
- A lady claims to be able to tell whether the tea or the milk is poured first
- 8 cups, 4 of which are tea first and 4 are milk first (and the lady knows this)
- Thus, the margins are known in advance
- Want to assess the chance of observing a result (table) *as or more extreme*

Fisher's Exact Test

- Method of testing for association when some *expected values are small*
- Measures the chances we would see differences of this magnitude or larger if there were *no association*
- The test is *conditional on both margins* – both the row and column totals are considered to be fixed

More about Fisher's exact test

- Fisher's exact test computes the probability, given the observed marginal frequencies, of obtaining exactly the frequencies observed and any configuration more extreme
- '*More extreme*' means any configuration with a *smaller probability of occurrence* in the same direction (one-tailed) or in both directions (two-tailed)

Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

	+	-	
A	3		5
B			10
	8	7	15

	+	-	
A	0		5
B			10
	8	7	15

	+	-	
A	4		5
B			10
	8	7	15

	+	-	
A	1		5
B			10
	8	7	15

	+	-	
A	5		5
B			10
	8	7	15

Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

.326

	+	-	
A	3		5
B			10
	8	7	15

.392

.007

	+	-	
A	0		5
B			10
	8	7	15

	+	-	
A	4		5
B			10
	8	7	15

.163

.093

	+	-	
A	1		5
B			10
	8	7	15

	+	-	
A	5		5
B			10
	8	7	15

.019

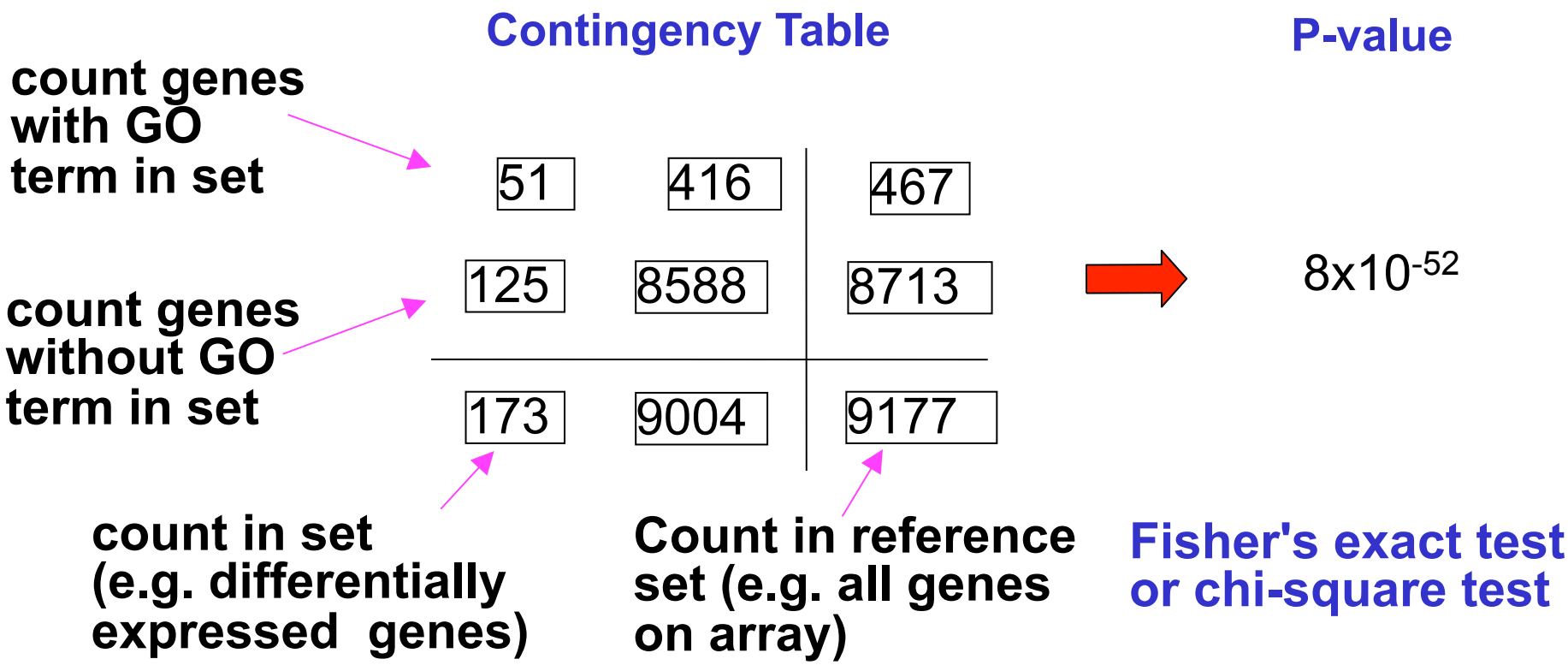
Where do these probabilities come from??

- With both margins fixed, there is only 1 cell that can vary
- The probabilities come from the hypergeometric distribution
- This distribution gives probabilities for the number of 'successes' in a sample of size n drawn without replacement from a population of size N comprised of a known number of 'successes'

Hypergeometric testing in BioC

- **Category** package
  - function **hyperGTest**
- **GOstats** package

Is a GO term is specific for a set?



Problems with Fisher's test

- The exact test was developed for the case of fixed marginals
- In this case the probability (p-value) computed by the Fisher test is *exact* (unlike the chi-square test, which relies on approximations)
- However, this setup is unrealistic for most studies – even if we know how many samples we will get in each group, we generally cannot fix in advance *both margins*
- Other methods have also been proposed to deal with this problem

(BREAK)

Original GSEA (Mootha *et al*)

- Genes involved in ***ox****idative* ***phos****phorylation* coordinately down-regulated in human diabetes
- Affy data on 22,000 genes in skeletal muscle biopsy samples from 43 males, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance and 18 with Type 2 diabetes (DM2)
- Computed *t*-statistic for each gene
- *No significant difference* found between NGT and DM2 after adjusting for multiple testing
- Their idea: test 149 *a priori* defined *gene sets* for association with disease phenotypes

The 149 gene sets

- *Sets of metabolic pathways:*
  - manually curated pathways (standard textbook literature reviews, and LocusLink)
  - Netaffx annotations using GenMAPP
- *Sets of coregulated genes:*
  - SOM clustering of the mouse expression atlas

Original GSEA calculation (I)

- For each gene set S, compute Kolmogorov-Smirnov running sum
- Order genes according to some criterion (e.g. a two-sample *t*-test)
- Beginning with the top ranking gene, the running sum increases when a gene in set S is encountered and decreases otherwise
- The enrichment score (ES) is defined to be the maximum value of the running sum

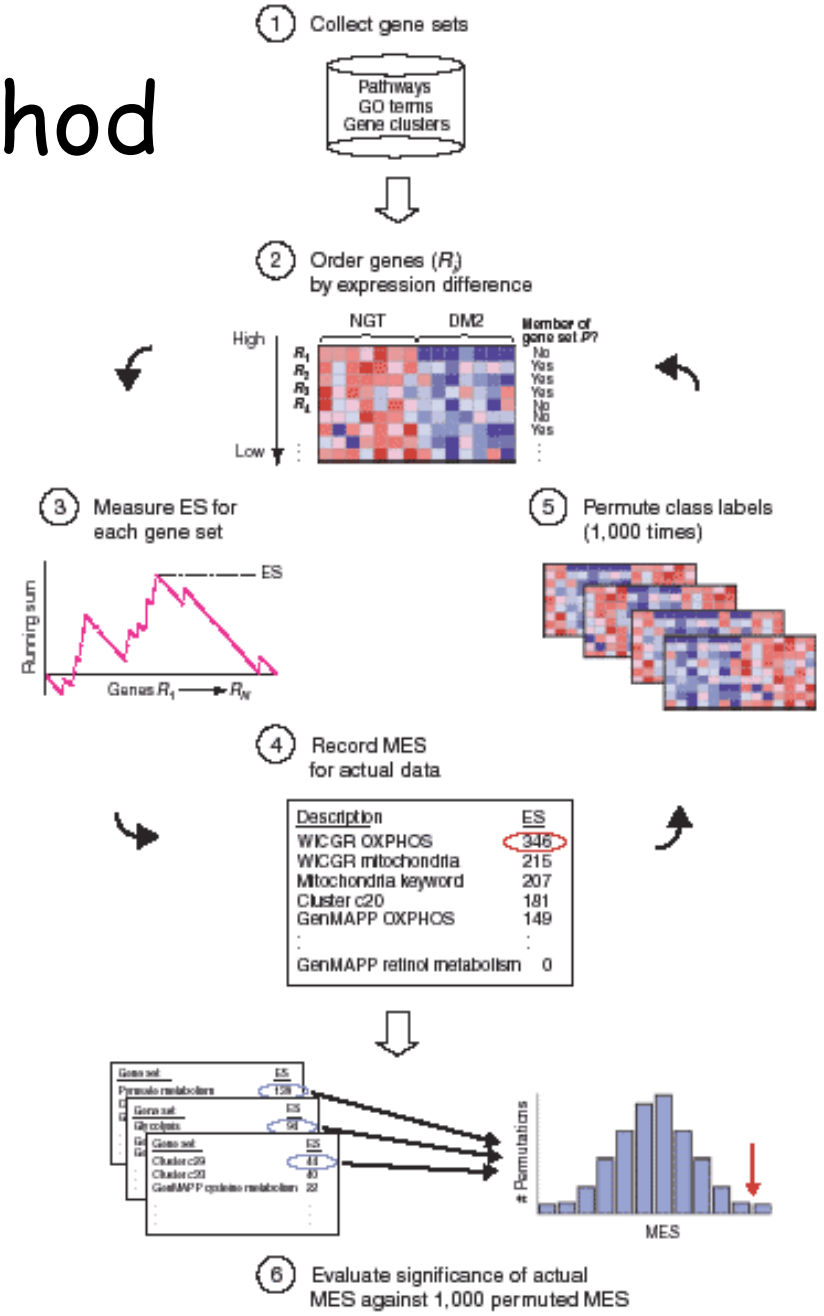
Original GSEA calculation (II)

- Obtain maximal ES (MES) over all sets S
- For each of B permutations of the class labels, ES and MES values are computed
- The observed MES is then compared to the B values of MES that have been computed, via permutation
- Several modifications have been proposed

Mootha *et al.* method

ES=enrichment score
for each gene
= scaled K-S dist

A set called **OXPHOS**
got the largest ES score,
with *p=0.029* on *1,000*
permutations.

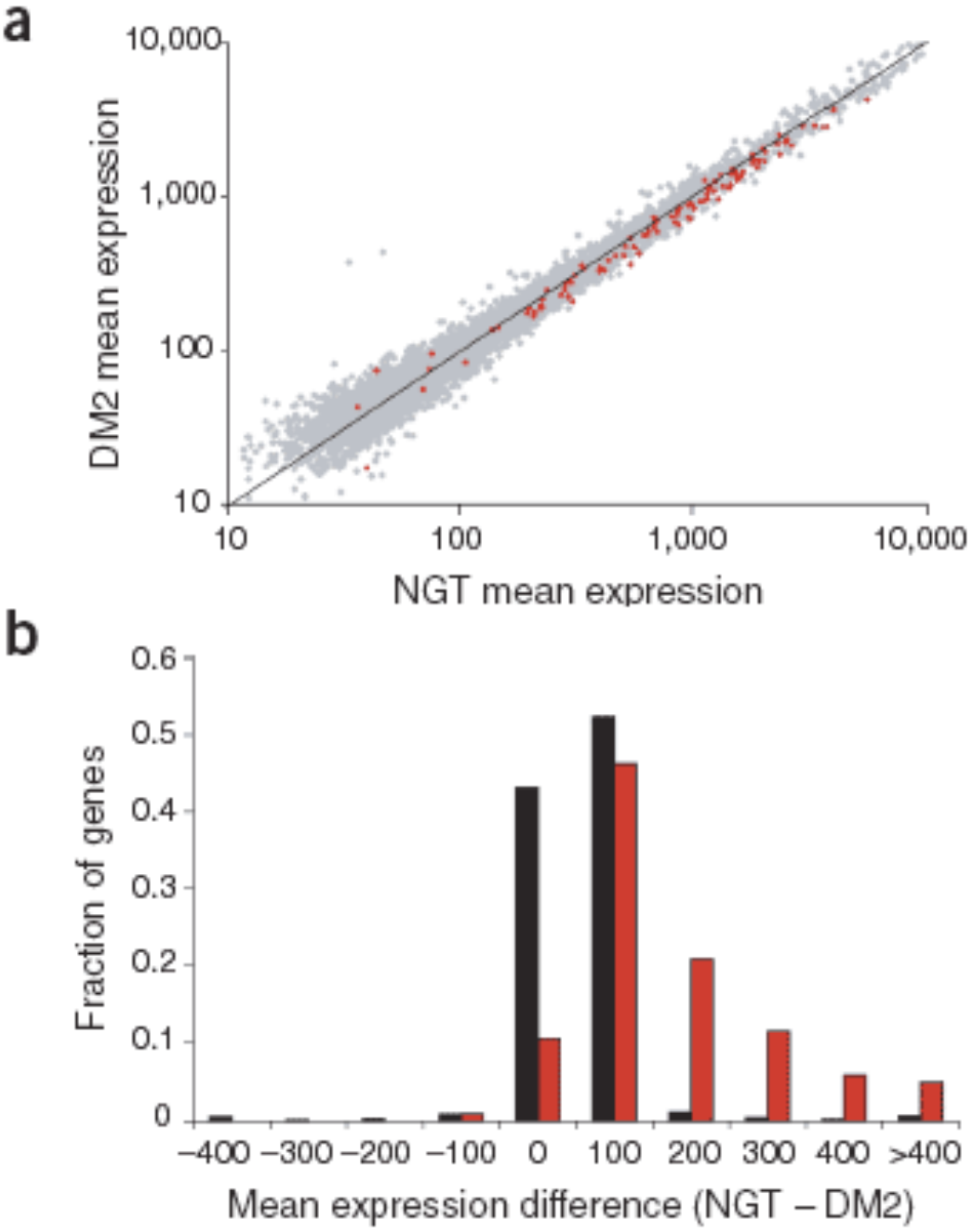


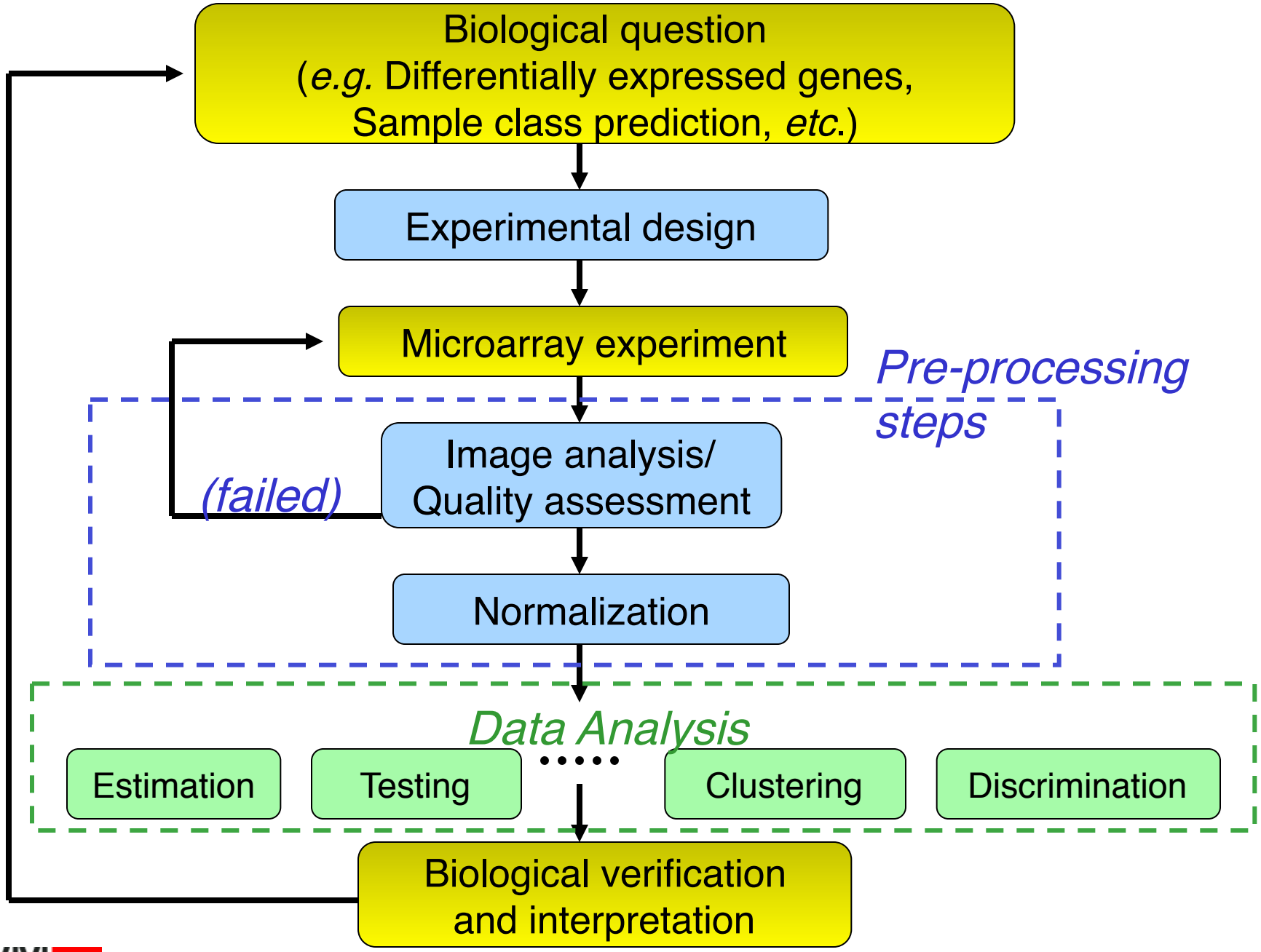
The result

OXPHOS

(A small difference for many genes)

All genes
OXPHOS




Biological question
(e.g. Differentially expressed genes,
Sample class prediction, etc.)
Experimental design
Microarray experiment
Pre-processing steps
(failed)
Image analysis/
Quality assessment
Normalization
Data Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Review of major points

- Normalization
- Affymetrix gene expression
- Affymetrix quality assessment
- Identifying DE
- Cluster analysis

Review of major points – Normalization

- The *purpose* of normalization is to correct for systematic differences which do not represent true biological variation
- Normalize as much as necessary, and as little as possible!
- Standard RMA normalization is quantile normalization

Review of major points – Affymetrix GeneChip expression

- Summarizes fluorescence intensities for all probes within a probeset (which represents a single sequence, or 'gene')
- 3 steps to a measure of expression: bg correction, normalization, summarization
- Best to do this *robustly*
- Best to normalize a set of chips (rather than adjusting each chip to a baseline chip)
- RMA – bg, quantile normalization, chip + probe effect model fit by median polish

Review of major points – Affymetrix GeneChip quality assessment

- Affymetrix quality measures (.rpt file) relate to hybridization quality
- Useful to consider instead quality of expression measure (since that is what data analysis/decisions are based on)
- Robust regression weights can be used to assess quality

Review of major points – Identifying differential expression

- Fold change (or log FC = M) intuitive for biologists, but ignores variability of replicate gene expression measurements across arrays
- t-statistic takes variability into account (too much when only a small number of replicates)
- Linear modeling and empirical Bayes moderated t-statistic (mod t) or B statistic perform better for detecting truly DE genes
- p-value adjustment
- Volcano plot to display (B vs. M, or |mod t| vs. M)

Review of major points – Cluster analysis

- Can cluster *samples* or *genes* or *both*
- Visualization: *dendrogram* (for hierarchical methods or *heatmap*
- There are many things that can vary in a cluster analysis: make choices/decisions based on the *aim* of the analysis and the *types of differences* you are interested in detecting