# So you have a cluster, now what???

- Hypothesis: genes with common expression patterns share biological functions.
- How do researchers determine the function of a gene?
  - Phenotypes
    - Correlate expression and mutants with phenotype
  - Biochemical
    - Enzymatic activity
      - Kinase
    - Substrate binding
      - DNA, RNA, proteins
  - Structural
    - Crystallography → functional domains
      - HTH
    - Superstructures
      - Tubulin
  - Computationally – homology to genes with known function

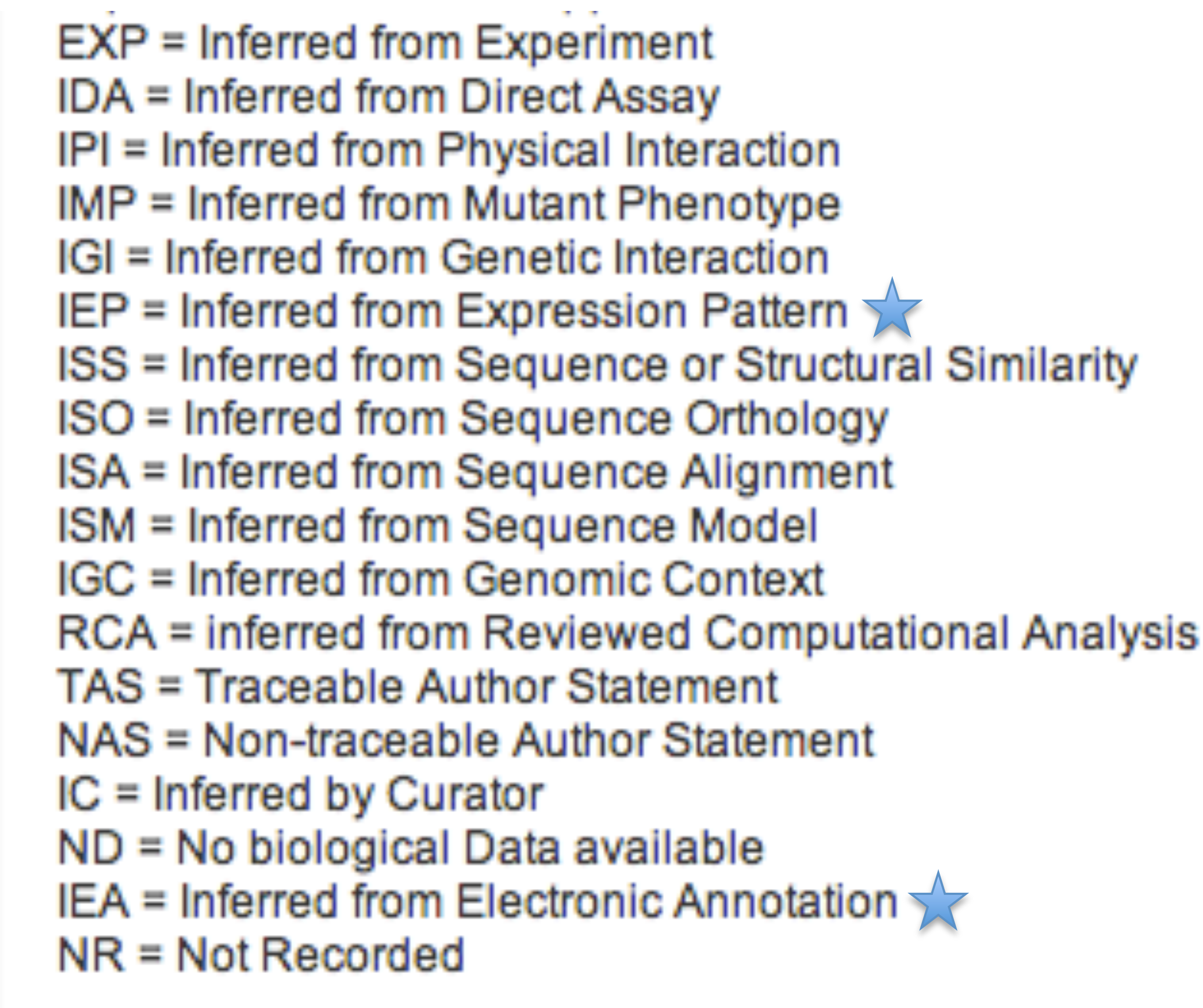
# Gene Ontology

- What is ontology?
  - The study of ‘being’ or ‘existence’
  - An attempt to classify and describe fundamental units of organization
  - In biology, GO classifies functions of gene products, or proteins
- Types
  - COGs – clusters of orthologous groups
  - arCOGs – COGs for archaea, which have unusual gene functions

# GO databases

- GO Consortium is a joint project of three model organism databases:
  - FlyBase
  - Mouse Genome Informatics (MGI)
  - Saccharomyces Genome Database (SGD)
- Has expanded in the last few years
  - http://www.geneontology.org/ - go there

# GO evidence codes

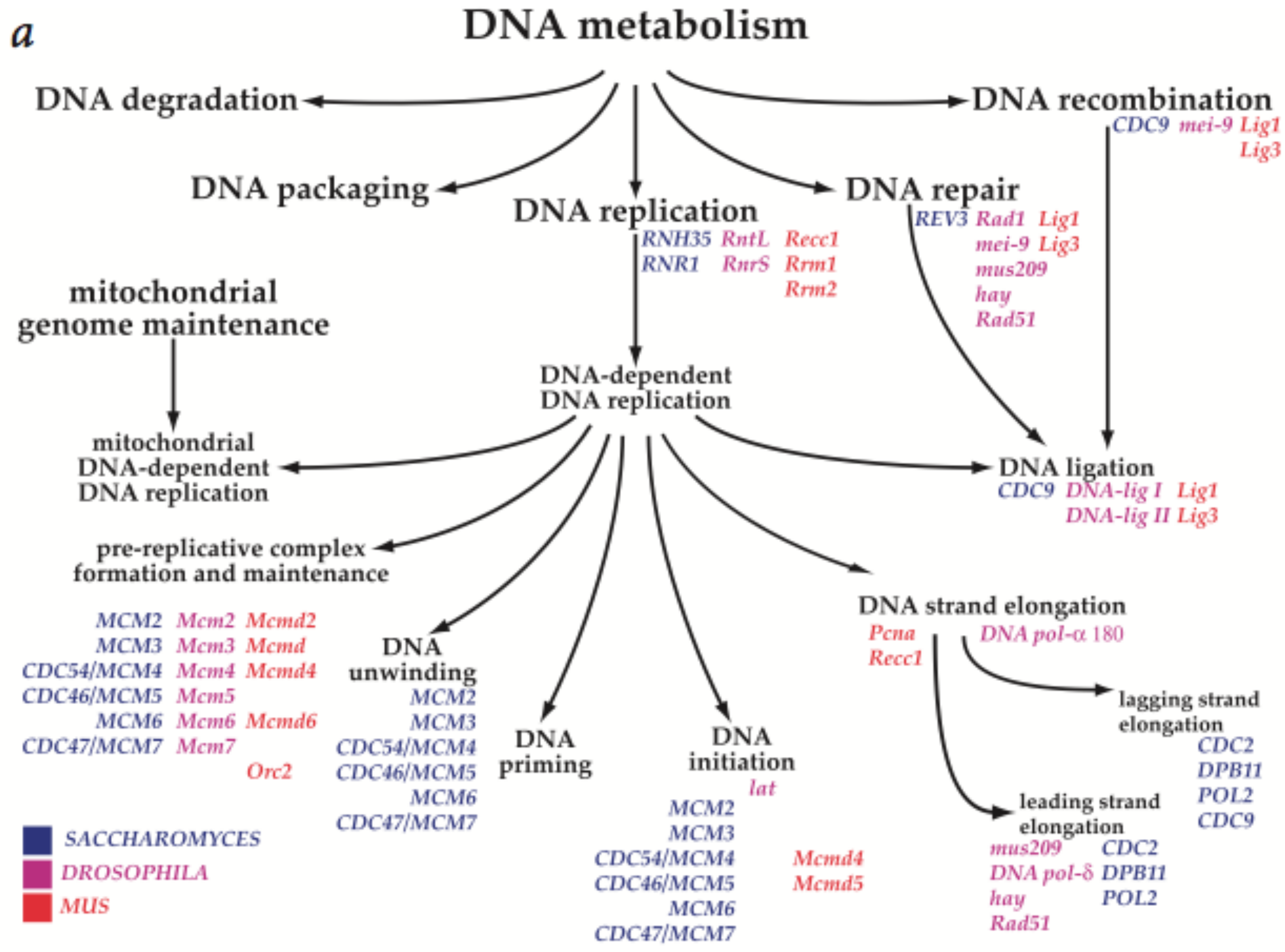


EXP = Inferred from Experiment
IDA = Inferred from Direct Assay
IPI = Inferred from Physical Interaction
IMP = Inferred from Mutant Phenotype
IGI = Inferred from Genetic Interaction
IEP = Inferred from Expression Pattern ★
ISS = Inferred from Sequence or Structural Similarity
ISO = Inferred from Sequence Orthology
ISA = Inferred from Sequence Alignment
ISM = Inferred from Sequence Model
IGC = Inferred from Genomic Context
RCA = inferred from Reviewed Computational Analysis
TAS = Traceable Author Statement
NAS = Non-traceable Author Statement
IC = Inferred by Curator
ND = No biological Data available
IEA = Inferred from Electronic Annotation ★
NR = Not Recorded

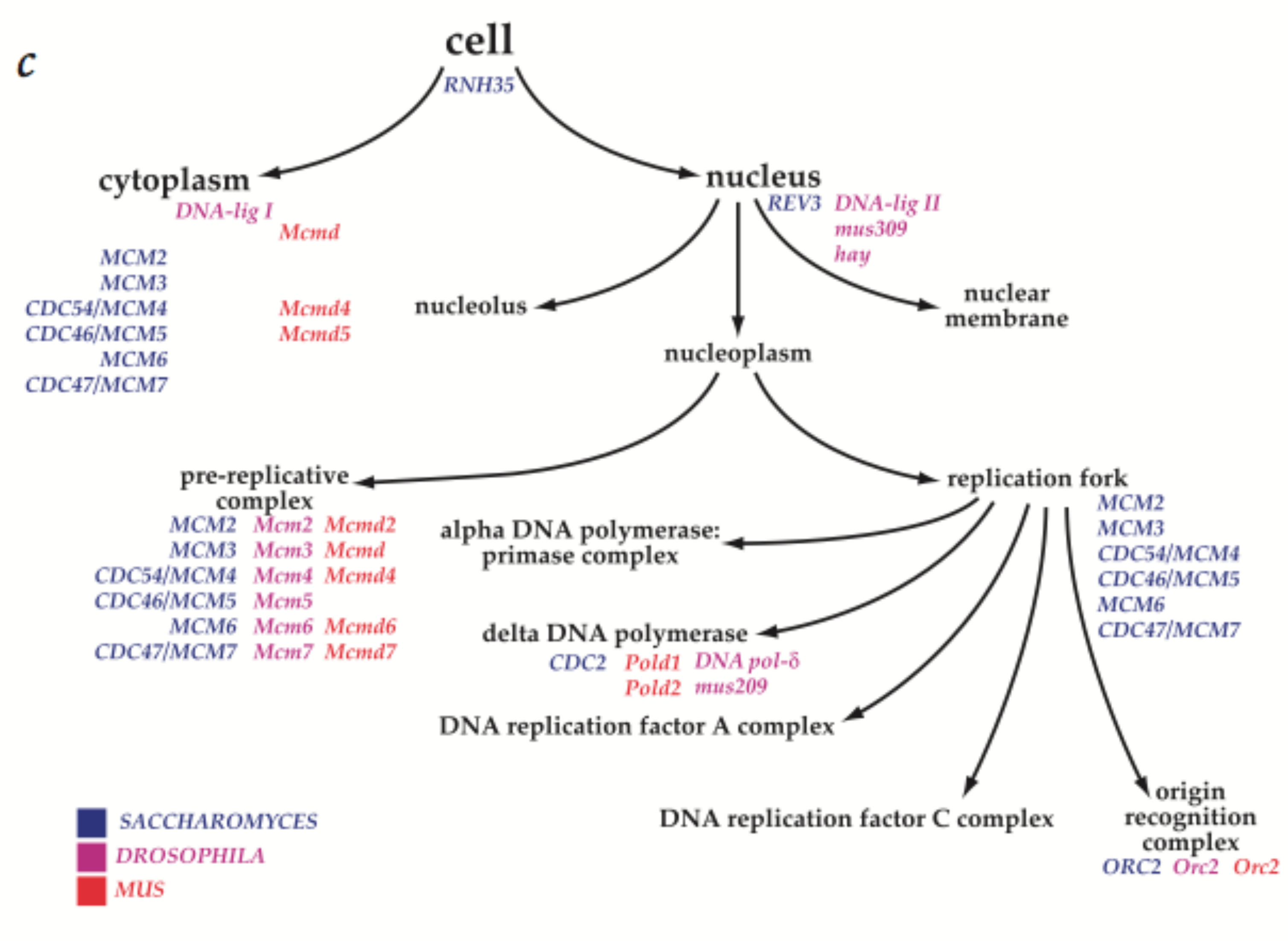
# Gene Ontology

- <u>3 ways to describe a gene (organized as acyclic graphs):</u>
- Biological process
  - Molecular events with a defined beginning or end, related to the function of integrated living units (cells, tissues, etc)
- Cellular component
  - The part of a cell or extracellular environment
- Molecular function
  - The elemental activities of a gene at the molecular level

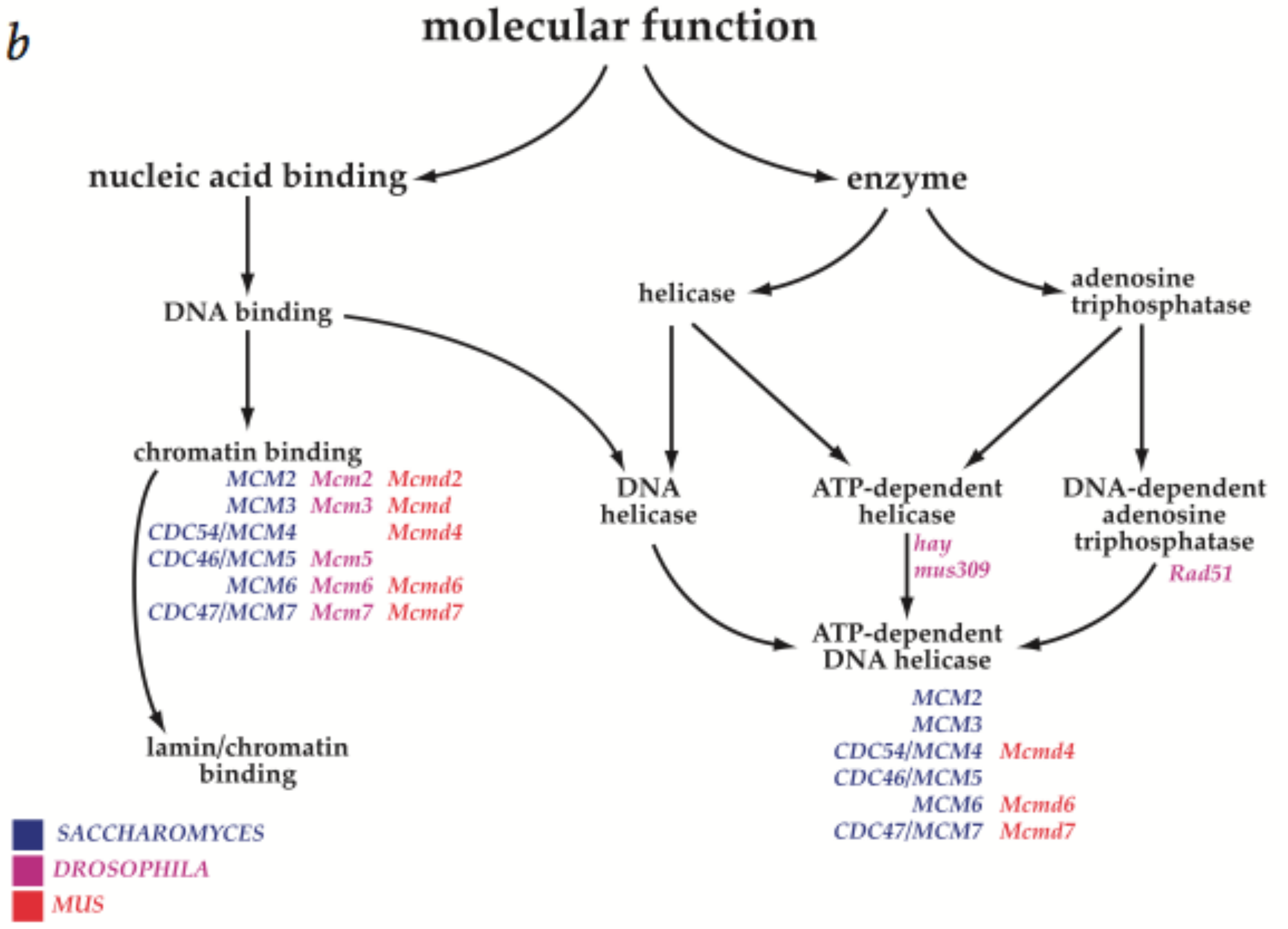
# Biological Process



# Cellular Component



# Molecular function



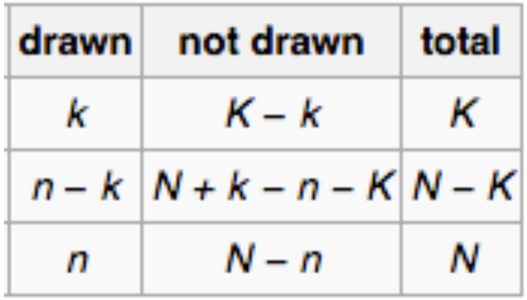
# STATISTICS

- **Are genes in a particular process/function represented above random chance?**
  - ~6000 genes in the yeast genome
  - Biological process→ cellular amino acid metabolic process = 242 genes = 3.8%
- **Gene Set Enrichment Analysis**
  - Hypergeometric test
  - p-value

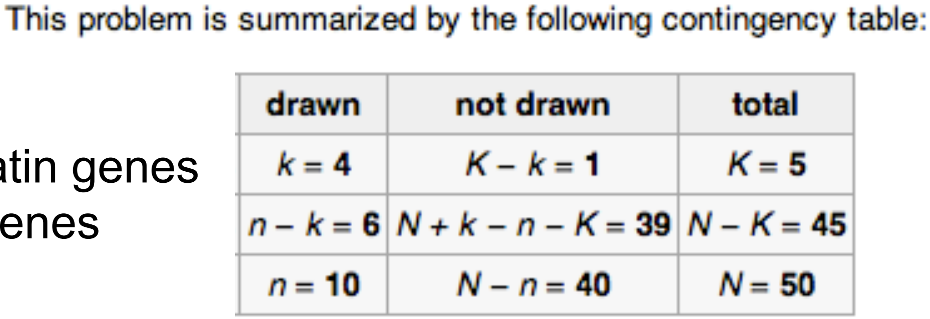
# HYPERGEOMETRIC TEST

- **Probability of:**
  - k successes from
  - n draws in a population of size
  - N containing a total of
  - K successes
- **In cluster analysis, probability of:**
  - Finding a certain number (k) of genes
  - Of a number n of a certain type (GO term, COG category)
  - In a population of N genes (in this case, the whole genome)
  - Out of all the genes (K) in our cluster

# HYPERGEOMETRIC TEST

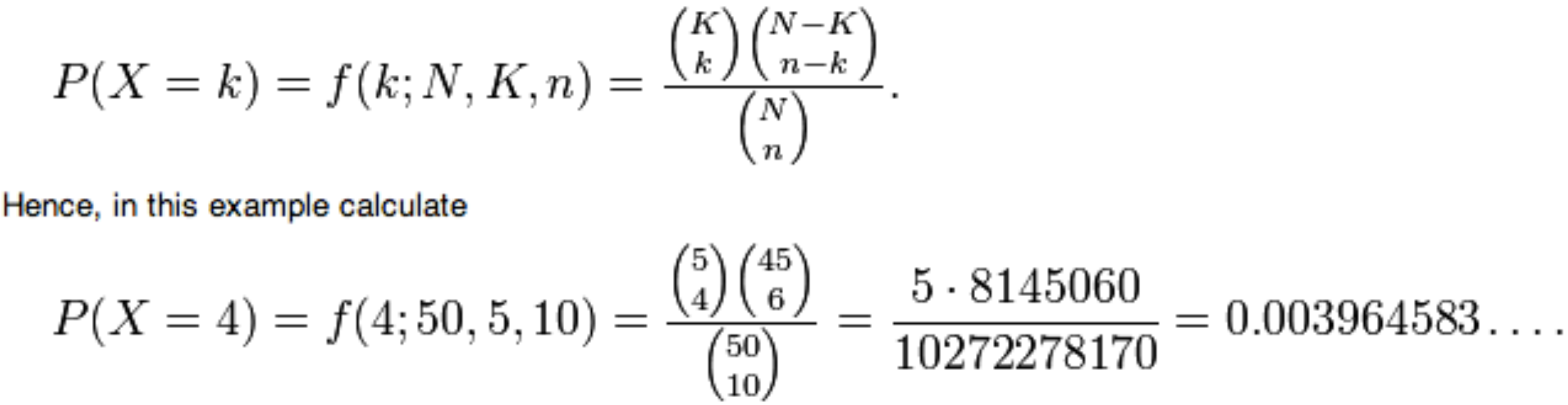


| | drawn | not drawn | total |
|---|---|---|---|
| Chromatin genes | $k$ | $K - k$ | $K$ |
| Other genes | $n - k$ | $N + k - n - K$ | $N - K$ |
| Total | $n$ | $N - n$ | $N$ |



This problem is summarized by the following contingency table:

| | drawn | not drawn | total |
|---|---|---|---|
| Chromatin genes | $k = 4$ | $K - k = 1$ | $K = 5$ |
| Other genes | $n - k = 6$ | $N + k - n - K = 39$ | $N - K = 45$ |
| Total | $n = 10$ | $N - n = 40$ | $N = 50$ |



$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Hence, in this example calculate

$$P(X = 4) = f(4; 50, 5, 10) = \frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}} = \frac{5 \cdot 8145060}{10272278170} = 0.003964583\ldots.$$

# Interpreting the results

- My cluster is enriched for functions in the glyoxylate cycle with a p-value of 0.001.
- Is this significant?
- What does it mean biologically? How do I find out?
- My cluster is also enriched for functions in glycolysis. Which “annotation” of my cluster is “right”?