IPA®

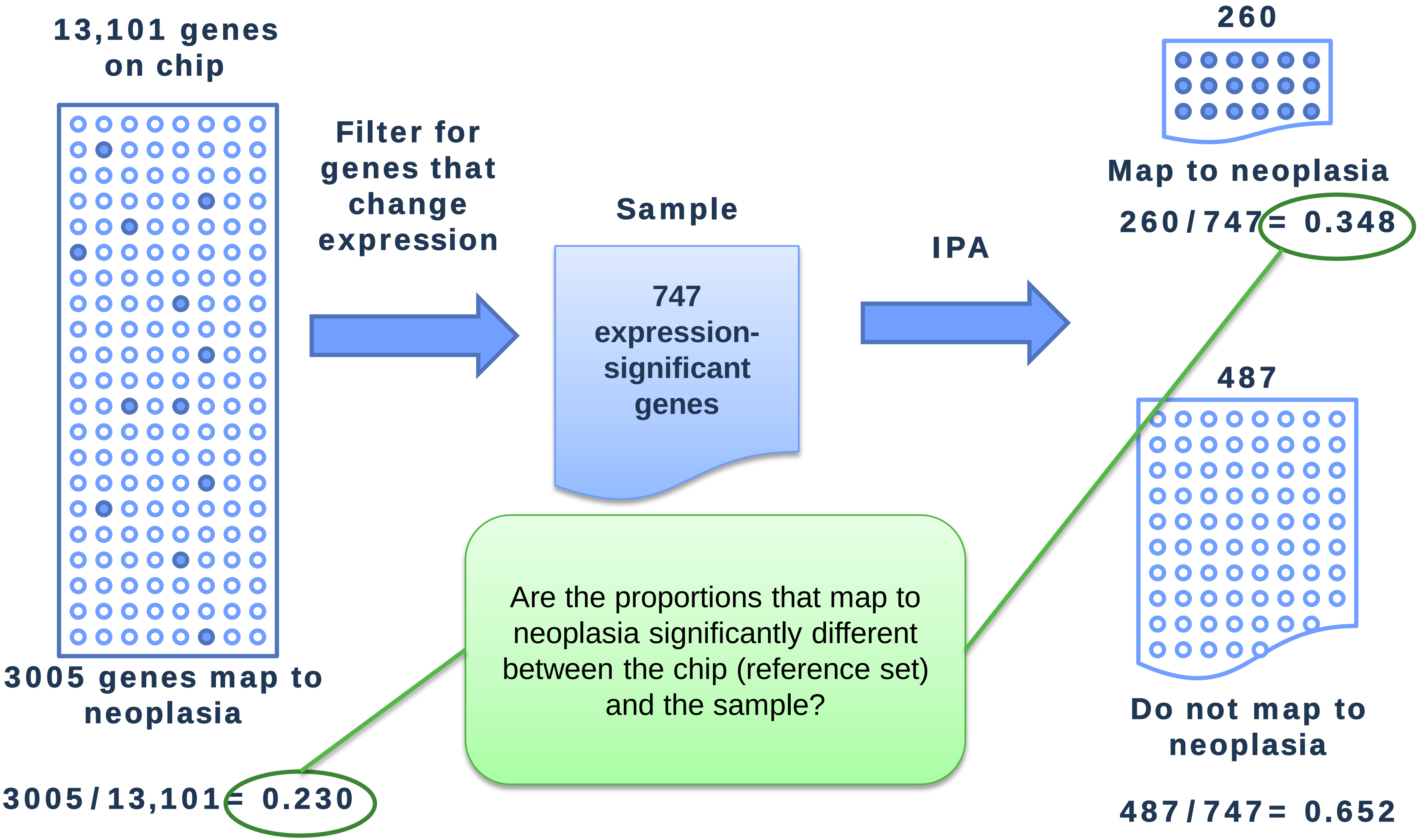
Integrate and understand complex ‘omics data

Statistics in IPA

How the Fisher's Exact Test is Calculated

- Is the proportion of genes in my sample mapping to a gene set (those that are significant) similar to the proportion of all measureable genes (reference set) that map in the gene set?
  - If the proportions are similar, there is no biological effect

Mapping Colorectal Cancer Expression Data to the Function "Neoplasia"



13,101 genes on chip

Filter for genes that change expression

Sample

747 expression-significant genes

IPA

260

Map to neoplasia

260 / 747 = 0.348

487

Do not map to neoplasia

487 / 747 = 0.652

3005 genes map to neoplasia

3005 / 13,101 = 0.230

Are the proportions that map to neoplasia significantly different between the chip (reference set) and the sample?

Calculating the Fisher's Exact Test

- A 2x2 contingency table is created based on the total population, the sample, and how many genes map to the function/pathway. This table is used to calculate the Fisher's Exact Test

	Neoplasia	Not Neoplasia	
In Sample	k	n - k	n
Not in Sample	m - k	N + k - n - m	N - n
	m	N - m	N

m= Total that map to function/pathway
N= Total
k= Number that map to function/pathway in sample
n= Total sample

Calculating the Fisher's Exact Test

- Numbers based on the colorectal cancer data mapping to neoplasia

	Neoplasia	Not Neoplasia	
In Sample	260	487	747
Not in Sample	2745	9609	12354
	3005	10096	13101

3005 = Total that map to neoplasia on chip
13101 = Total on chip
260 = Number that map to neoplasia in sample
747 = Total sample

Fisher's Exact Test p-value = 2.13 E-14

What Can We Say About Our Colorectal Cancer Data Set And Neoplasia?

- We can conclude that the proportion, or over representation, of genes mapping to neoplasia is not likely the result of sampling and is likely an effect of the disease

What Universe Is Used

- For Downstream Effects Analysis (function analysis):
  - Same as above, all genes with function-annotation, plus all genes within canonical pathways, including members of groups and complexes, plus members of tox lists. Note that there is overlap between these sets and the union is taken as the reference set.
- For canonical Pathway analysis:
  - All genes with function-annotation, plus all genes within canonical pathways, including members of groups and complexes, plus members of tox lists. Note that there is overlap between these sets and the union is taken as the reference set.
- For Upstream Analysis
  - All genes and chemicals with downstream targets (E, T, and PD relationship types), and those targets, subject to filtering as above
- For networks:
  - All genes with one or more molecular connections.

How Do I Choose The Reference Set?

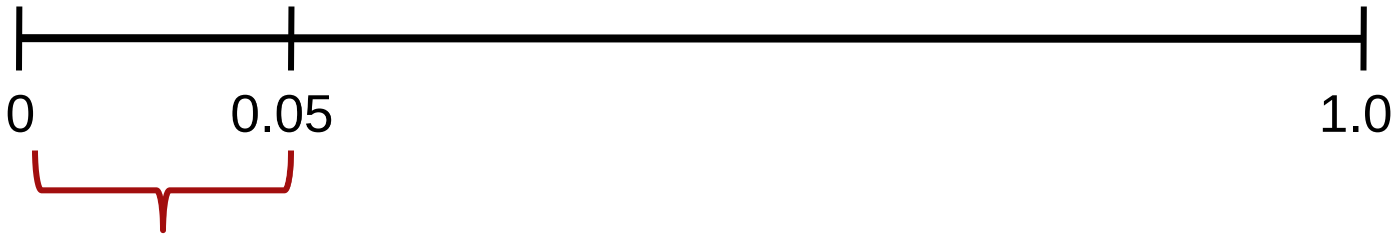
- If you are using a standard vendor platform supported by IPA, then that platform should be selected as your reference set.
- If you do not know the platform or the data were taken from different platforms, select a reference set that best estimates the entire population you evaluated.
  - For gene expression data, select the “Ingenuity Knowledge Base (genes only)”
    - This setting uses all function- and pathway-eligible genes in the knowledge base, approximately 14,500.
  - For metabolomics, select the “Ingenuity Knowledge Base (endogenous chemicals only)”
  - You have the option to having your uploaded data set used as the reference set (User Data Set)

What About TaqMan or Similar Focus Array?

- Low density arrays are problematic because the genes that are being measured are usually not randomly chosen to start with, but are typically selected based on a priori function or pathway knowledge
- Let’s assume a inflammatory cytokine array
  - If you select the Ingenuity Knowledge Base as your reference, your p-values for inflammation functions and pathways will be artificially low (significant) because the array was heavily biased for these genes.
  - If you upload every gene on the array, and select the “User Data Set” reference option, your p-values are statistically accurate, but inflammatory functions and pathways may not appear significant because the likelihood of having a random sample with similar proportions to inflammation processes is extremely high.

Multiple Testing Correction

- Benjamini-Hochberg method of multiple testing correction
- Based on the Fisher’s exact test p-value
- Calculates false discovery Rate
  - Threshold indicates the fraction of false positives among significant functions

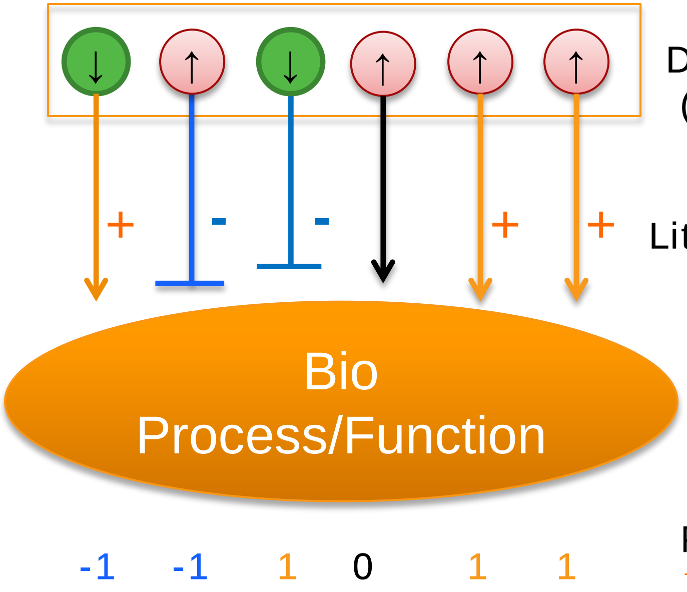


5% (1/20) may be a false positive

Which p-Value Calculation Should I Use?

- “What is the significance of function X in relation to my dataset?”
  - Use Fisher’s Exact test result
- “What are all significant functions within this dataset?”
  - Use Benjamini-Hochberg multiple testing correction

Downstream Effect z-score



Differential Gene Expression
(Uploaded Data)

Literature-based effect genes have on process or function

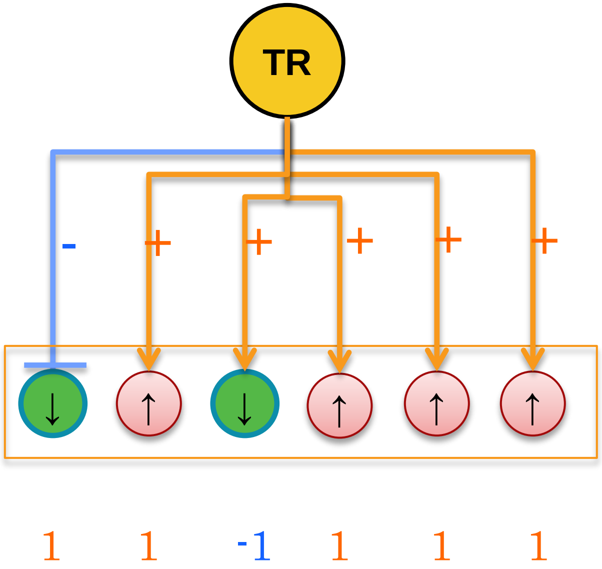
Predicted Effect-
1: Increasing (correlated), -1: inhibited (anti-correlated)

$$z = \frac{x}{\sigma_x} = \frac{\sum_i x_i}{\sqrt{N}} = \frac{N_+ - N_-}{\sqrt{N}} = \frac{1}{\sqrt{5}} = .447$$

- “z-score” is statistical measure of correlation between relationship direction and gene expression.
- z-score > 2 or < -2 is considered significant

Actual z-score is weighted by relationship, relationship bias, data bias

Activation z-score



Every TR is analyzed

Literature-based effect TR has on downstream genes

Differential Gene Expression (Uploaded Data)

Predicted activation state of TR:
1: activated (correlated), -1: inhibited (anti-correlated)

$$z = \frac{x}{\sigma_x} = \frac{\sum_i x_i}{\sqrt{N}} = \frac{N_+ - N_-}{\sqrt{N}} = \frac{4}{\sqrt{6}} = 2.04$$

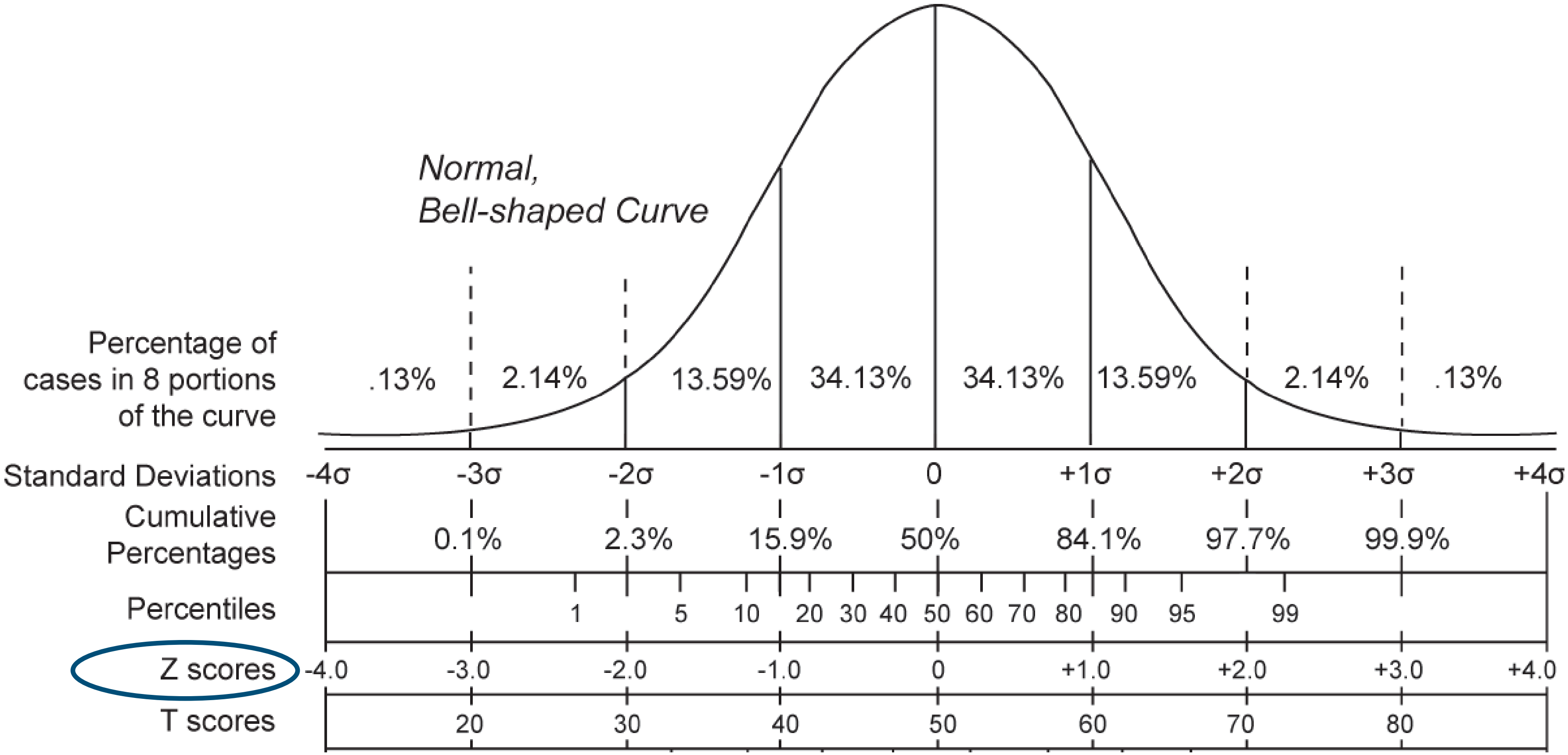
- z-score is statistical measure of correlation between relationship direction and gene expression.
- z-score > 2 or < -2 is considered significant

Actual z-score *can* weighted by relationship, relationship bias, data bias

z-scores and Normal Distribution

A set of genes chosen at random should be about equally likely to have an increasing or decreasing effect, thus, about 50% each direction, or a z=0.

A z-score represents the non-randomness of directionality within a gene set


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80