

STATISTICAL ANALYSIS OF TRANSCRIPT COUNTS

Jan Ruijter

Transcript count libraries

assumption:

every mRNA copy has the same chance of ending up in the library

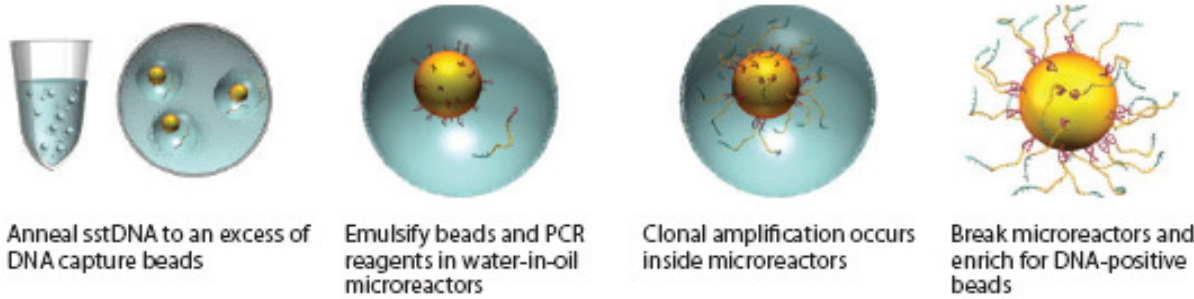
$$\frac{n_{\text{specific mRNA}}}{N_{\text{total mRNA}}} = \frac{n_{\text{specific counts}}}{N_{\text{total counts}}}$$

The diagram shows two dashed boxes. The left box is labeled **CELL** and contains the fraction $\frac{n_{\text{specific mRNA}}}{N_{\text{total mRNA}}}$. The right box is labeled **LIBRARY** and contains the fraction $\frac{n_{\text{specific counts}}}{N_{\text{total counts}}}$. An equals sign is placed between the two boxes, indicating that the two fractions are equal.

Sequencing libraries

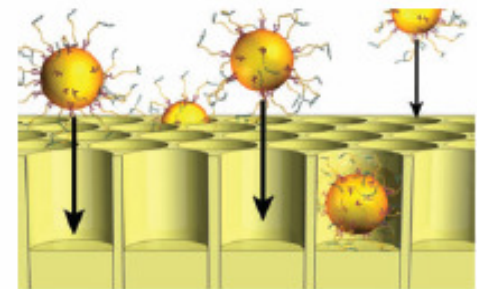
Emulsion PCR

8 hours



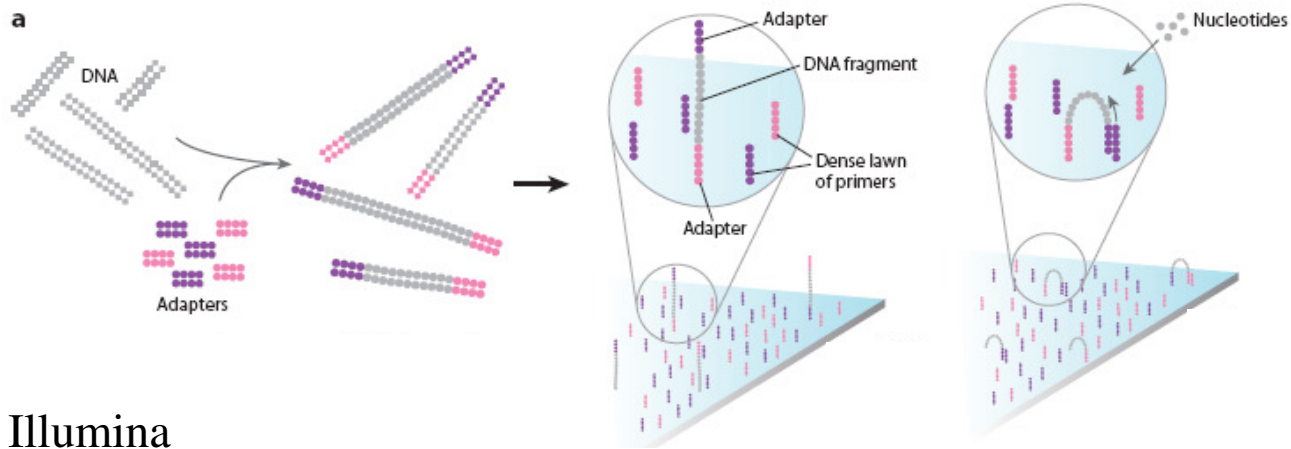
Sequencing

7.5 hours

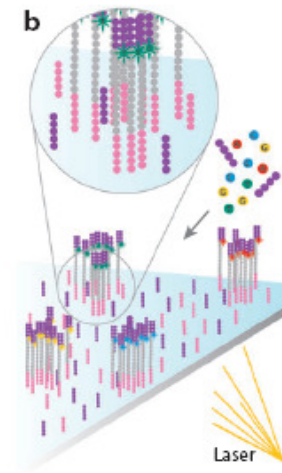


Roche 454

a

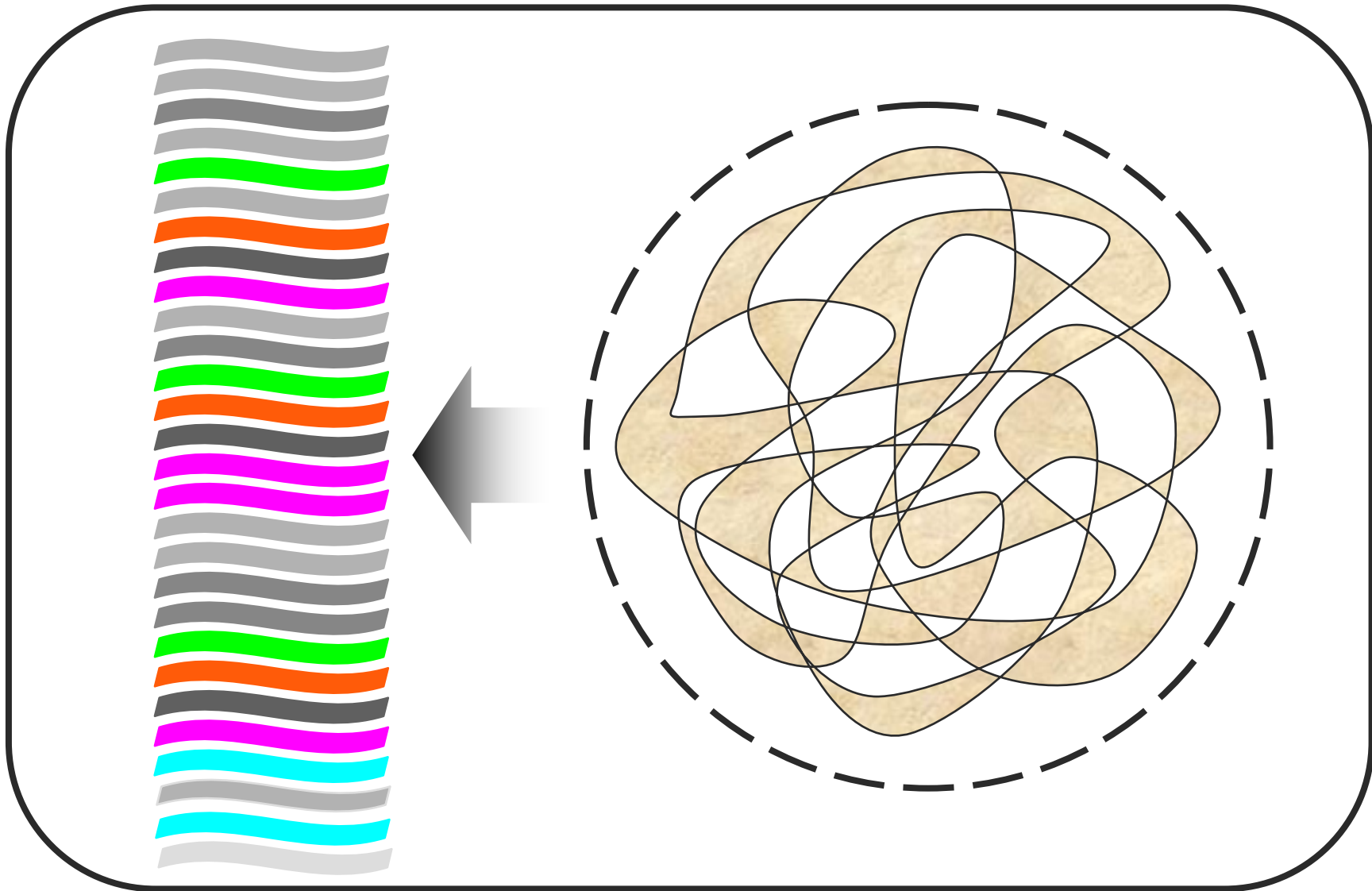


b

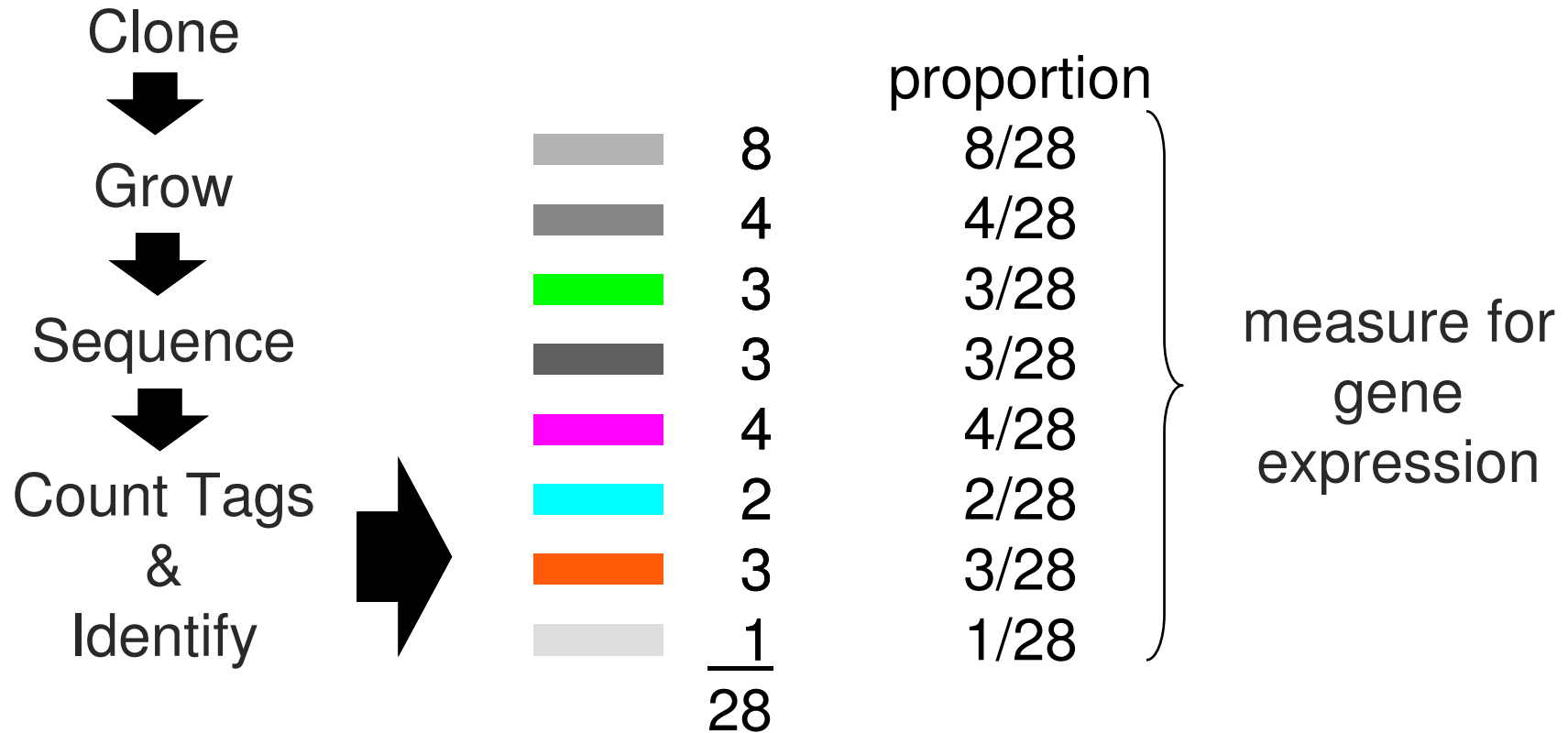


Illumina

Introduction to SAGE



Introduction to SAGE



SERIAL ANALYSIS OF GENE EXPRESSION

Comparison of two SAGE Libraries

library Name	normal_cerebellum	BB542_whitematter
species	Hs	Hs
library Type	Normal	Normal
tissue Type	Brain	Brain

Library	sage_lib13	sage_lib67
Tag		
AGAAAGATGT	2	1
AACGACCTCG	16	35
AACTGCTTCA	2	2
ACCCTTCCCT	2	7
AAGGAATCGG	0	3
AATAAAGCTA	39	47
AAGCATTAAA	16	55
ACAACAAAGA	35	42
ACAACACTAC	23	48
AAATAAAGCC	45	8
AAATAAAAGA	2	0
ACTTTTTGGC	4	3
"	"	
Total	51280	94876

test per tag

Tag	AATAAAGCTA	
	Lib 13	Lib 67
n	39	47
N	51280	94876

Hypothesis testing

'direct'

H_0 (Null) hypothesis:
there is no difference

'indirect'

calculate chance P to
find observed difference
when H_0 is true

calculate test statistic T
with observed difference

determine chance P
to find observed T
when H_0 is true

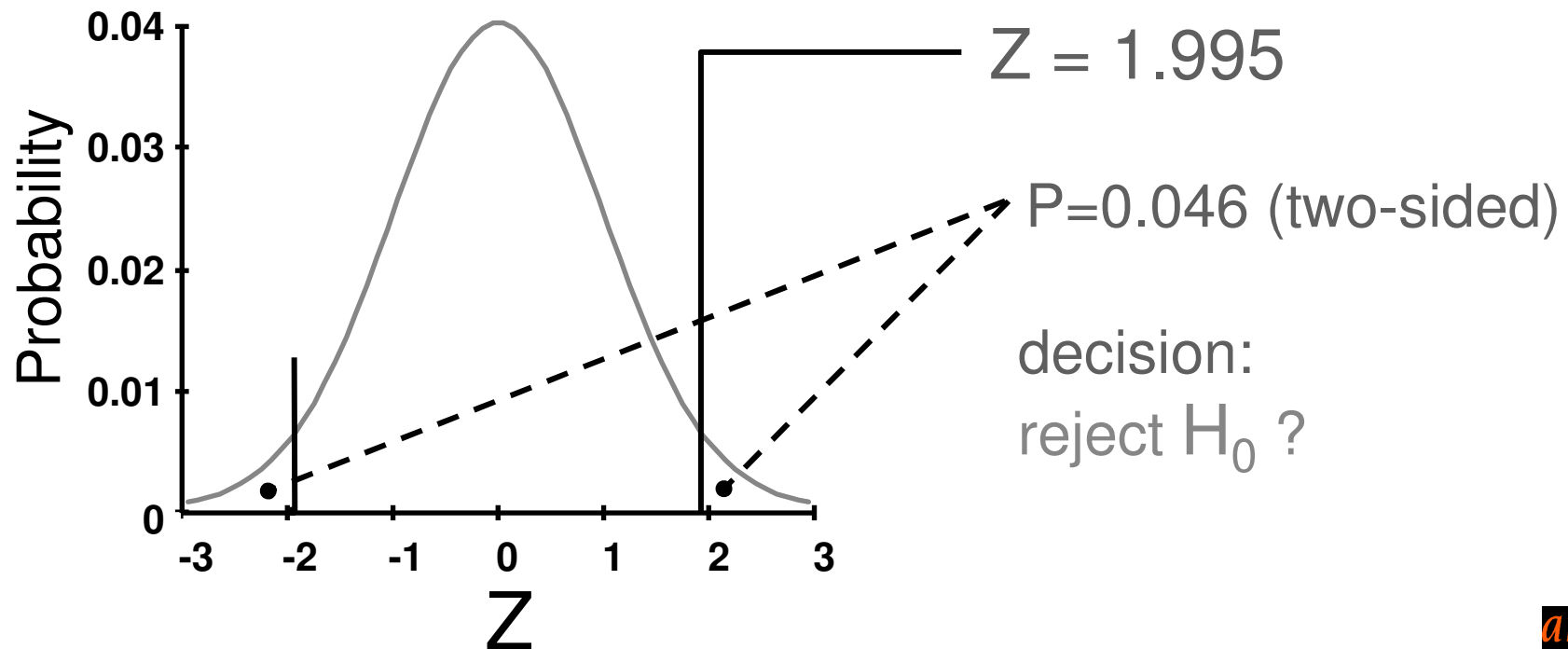
decision rule:
reject H_0 when P is smaller
than significance level α

Hypothesis testing: Z-test

Tag	AATAAAGCTA	
	Lib 13	Lib 67
n	39	47
N	51280	94876
p	0.00076	0.00050

$$H_0: p_1 = p_2 = p_0$$

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(1/N_1 + 1/N_2)}}$$



Tests used for comparison of two libraries

Z-test
(Kal et al. 1999)

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(1/N_1 + 1/N_2)}}$$

Chi-squared test

$$Chi^2 = \sum \left\{ (n - n_0)^2 / n_0 \right\}$$

Madden et al. 1997

$$Z = \frac{n_1 - n_2}{\sqrt{n_1} + \sqrt{n_2}}$$

Fisher's Exact test

$$P(n_1, n_2) = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2}}{\binom{N_1 + N_2}{n_1 + n_2}}$$

Audic and Claverie
(1997)

$$P(n_2 | n_1) = \frac{\left(\frac{N_2}{N_1}\right)^{n_2} \frac{(n_1 + n_2)!}{n_1! n_2! \left(1 + \frac{N_2}{N_1}\right)^{(n_1 + n_2 + 1)}}}{P < \alpha / 2}$$

SAGE300

P from Monte Carlo simulation

reject H_0 when

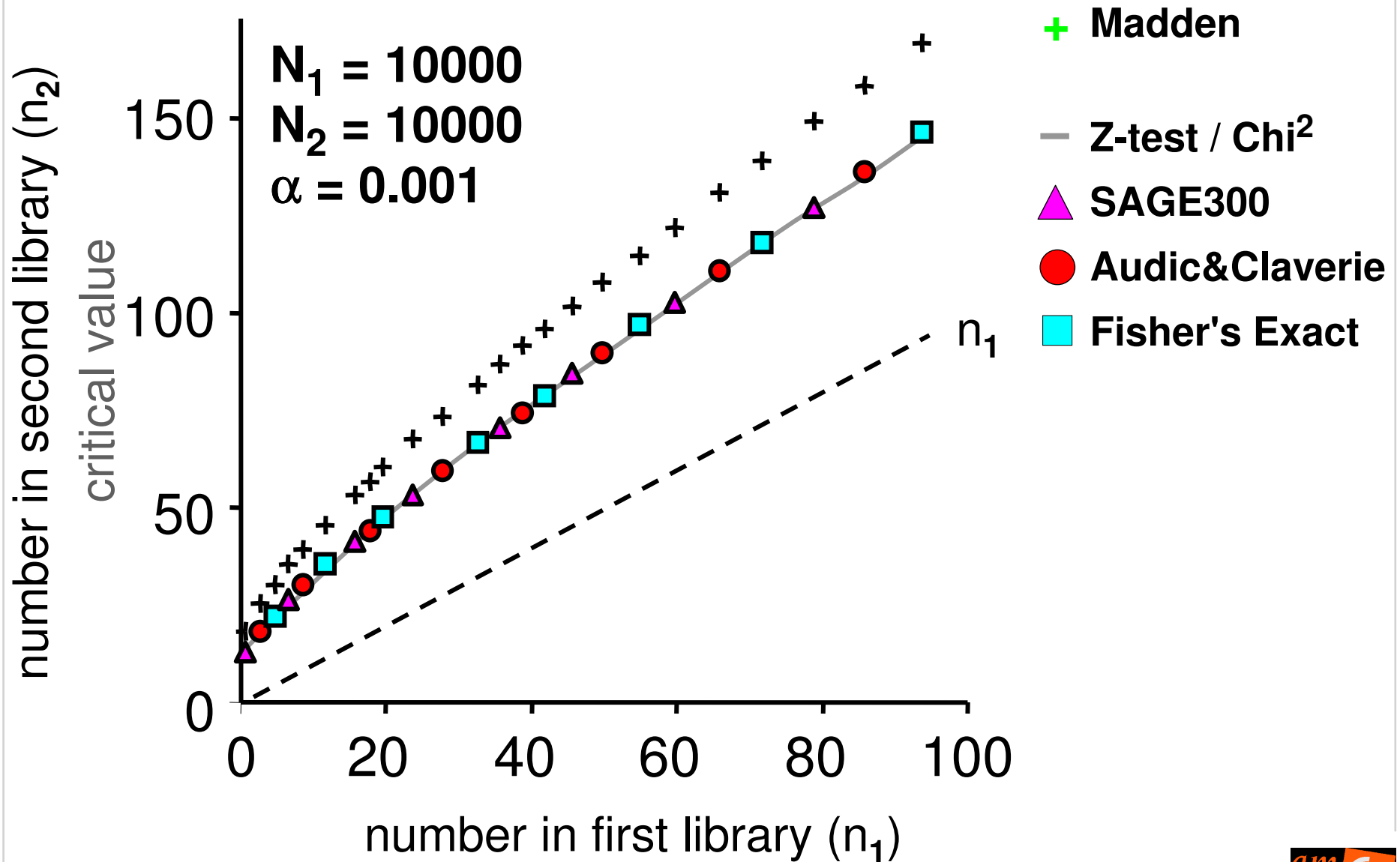
$$Z > Z_{\alpha/2}$$

or

$$Z < -Z_{\alpha/2}$$

$$P < \alpha / 2$$

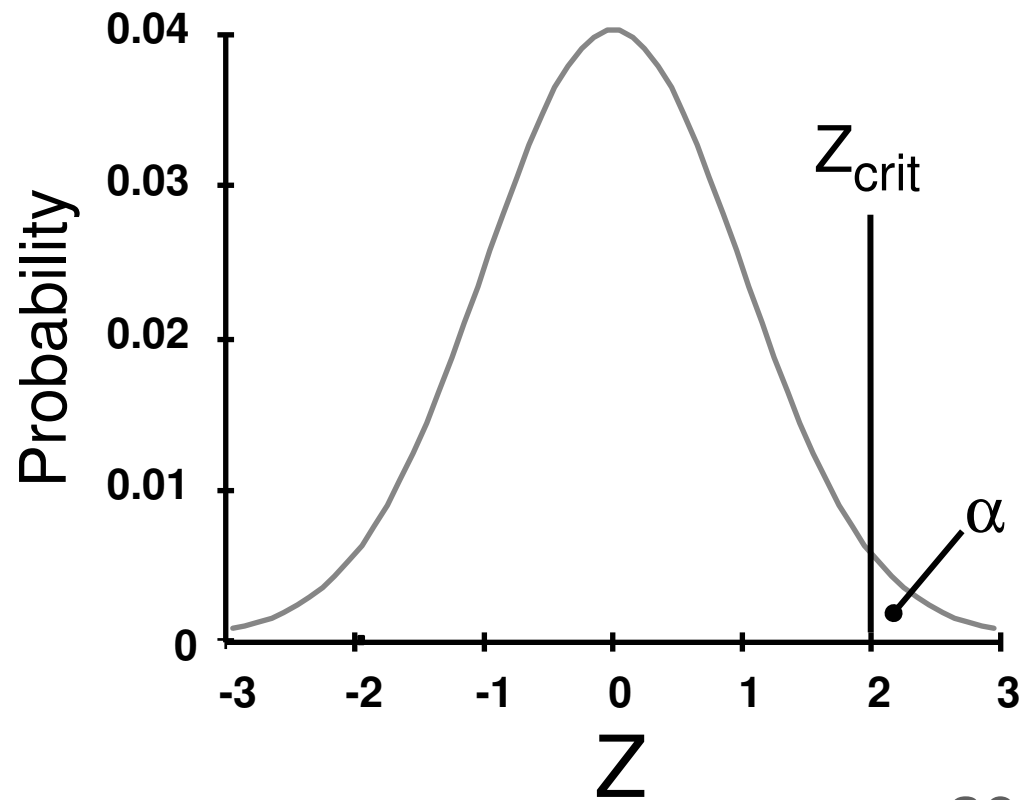
Comparison of tests to compare two libraries



From two to many

Common approach:

pair-wise test of all pairs of libraries



α = chance that H_0 is rejected while H_0 is true

k libraries:

$$P_{\text{error}} = 1 - (1 - \alpha)^{k(k-1)/2}$$

accumulation of Type I error

From two to many

Other approaches:

When subsets of libraries are known:

t-test between two groups of proportions

ignores library sizes

treats all libraries as equally precise

Z-test between pooled libraries

artificially large library size

ignores variation between libraries

Comparison of many libraries

H_0 (Null) hypothesis:

ALL libraries have the same tag abundance

H_1 alternative hypothesis:

at least one library has a different tag abundance

remaining question:

- which library / libraries deviate(s)

specific questions:

- how much does each library differ from the others
- are there homogeneous subsets of libraries

Comparison of many libraries

Two step approach:

Step 1: test overall H_0 : all libraries are equal

when H_0 is rejected

Step 2a: determine deviating libraries

or

Step 2b: test against known subset

or

Step 2c: search for homogeneous subsets

Sample data set

AATAAAGCTA (synuclein, beta)

Library	Tissue	Specific	Other	Total
Lib 13	N	39	51241	51280
Lib 30	MC	2	48552	48554
Lib 37	A	1	80264	80265
Lib 41	G	3	61883	61886
Lib 42	G	1	70086	70087
Lib 47	A	1	77003	77004
Lib 56	M	5	38928	38933
Lib 57	N	56	48489	48545
Lib 67	N	47	94829	94876
Lib 68	N	46	58780	58826
Lib 107	G	2	62673	62675
Lib 112	N	81	77887	77968
Lib 122	NC	2	52259	52261
Lib 125	N	52	63156	63208
Lib 127	A	3	38631	38634

15 "brain" libraries

7 Normal

8 Tumor:

Glioblastoma

Astrocytoma

Medulloblastoma

Cell line

G-statistic

G-statistic: $G = 2 \cdot \ln (\text{Likelihood ratio})$

AATAAAGCTA

Library	Specific	Other
Lib 13	39	51241
Lib 30	2	48552
"	"	"
Lib 125	52	63156
Lib 127	3	38631

likelihood of observed results

likelihood of observed results
when H_0 is true

G-statistic

G-statistic: $G = 2 \sum_{Libs} \sum_{s-o} (n \cdot \ln(n/n_o))$

AATAAAGCTA

Library	Specific	Other
Lib 13	39	51241
Lib 30	2	48552
"	"	"
Lib 125	52	63156
Lib 127	3	38631

n = observed number
n₀ = expected number
when H₀ is true

G-statistic: overall test, intrinsic H₀

G-statistic: $G = 2 \sum_{Libs} \sum_{s-o} (n \cdot \ln(n/n_o))$

AATAAAGCTA

Library	Specific	Other	Total
Lib 13	39	51241	51280
Lib 30	2	48552	48554
"	"	"	"
Lib 125	52	63156	63208
Lib 127	3	38631	38634
Total	<u>341</u> +		<u>925002</u> +

expected values (n₀)

19	51261
18	48536
"	"
23	63185
14	38620

} G_{intrinsic} = 453

compare to χ^2 distribution

abundance according to H₀ :

$$p_0 = \frac{341}{925002} = 0.00037$$

P < 0.00000

decision: reject H₀

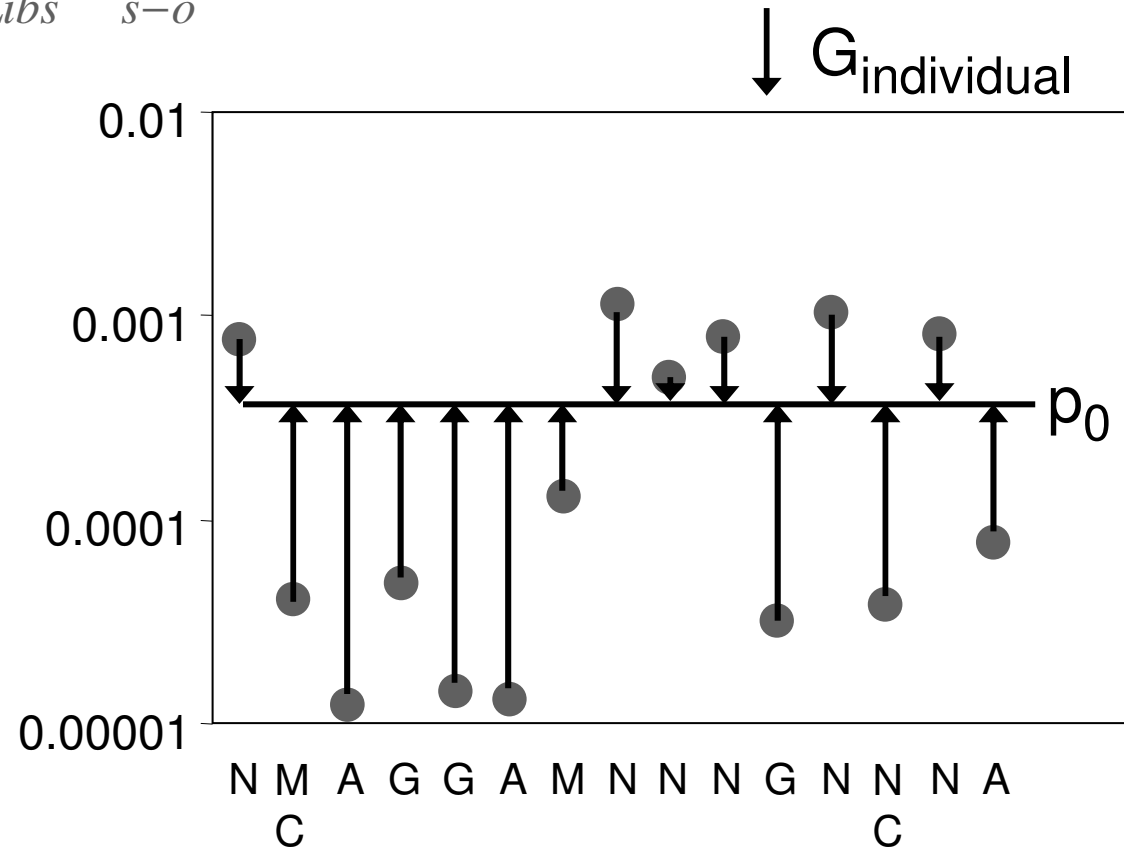
$$G_{\text{intrinsic}} = \sum G_{\text{individual}}$$

G-statistic: $G = \sum_{\text{Libs}} 2 \sum_{s-o} (n \cdot \text{Ln}(n/n_o))$

AATAAGCTA

Library	Specific	Other
Lib 13	39	51241
Lib 30	2	48552
"	"	"
Lib 125	52	63156
Lib 127	3	38631

$$p_0 = \frac{341}{925002} = 0.00037$$



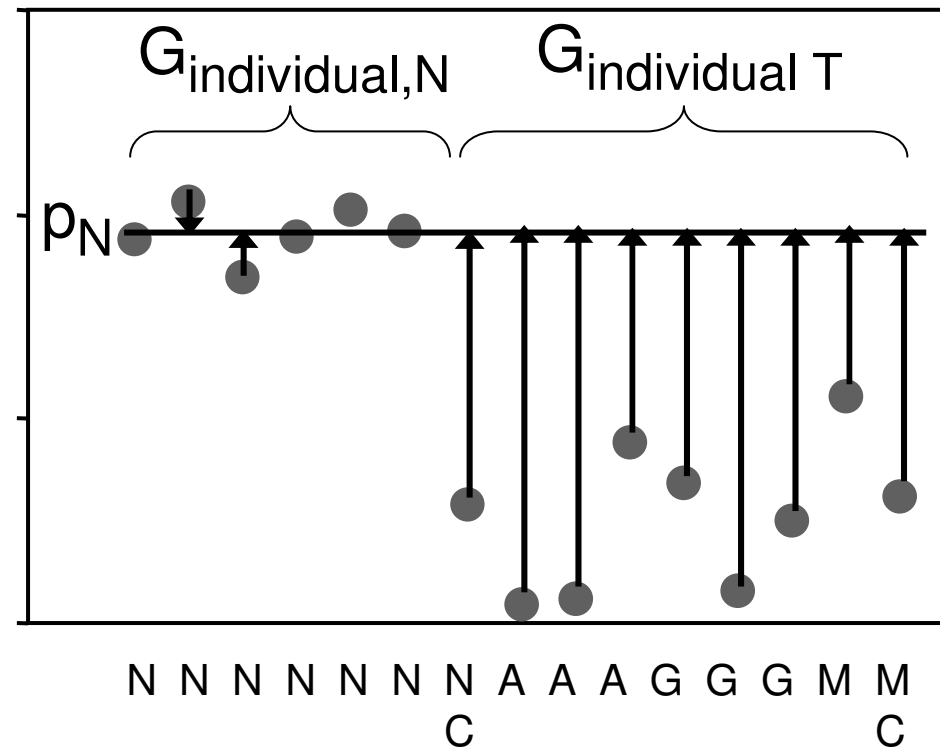
Supervised test: standard subset

G-statistic: $G_{extrinsic} = \sum_{libs} G_{individual, p_N}$

standard subset:
normal brains

calculate $p_N = \Sigma n / \Sigma N$

calculate $G_{individual}$
compared to p_N



Supervised test: comparison of subsets

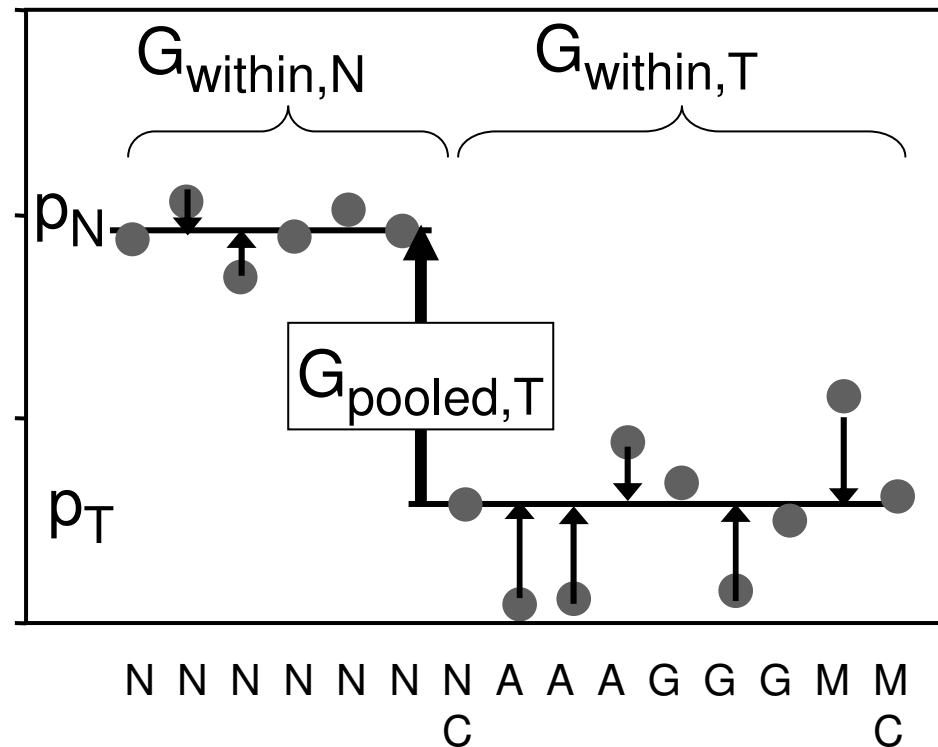
G-statistic: $G_{extrinsic} = \sum_{N+T} G_{within} + G_{pooled, p_N - p_T}$

standard subset:
normal brains

calculate p_N and p_T

calculate $G_{individual}$
compared to p_N or p_T

calculate G_{pooled}
= difference of p_N and p_T



Supervised test: decision rules

G-statistic: $G_{extrinsic} = \sum_{N+T} G_{within} + G_{pooled, p_N - p_T}$

Subsets differ when:

overall test:

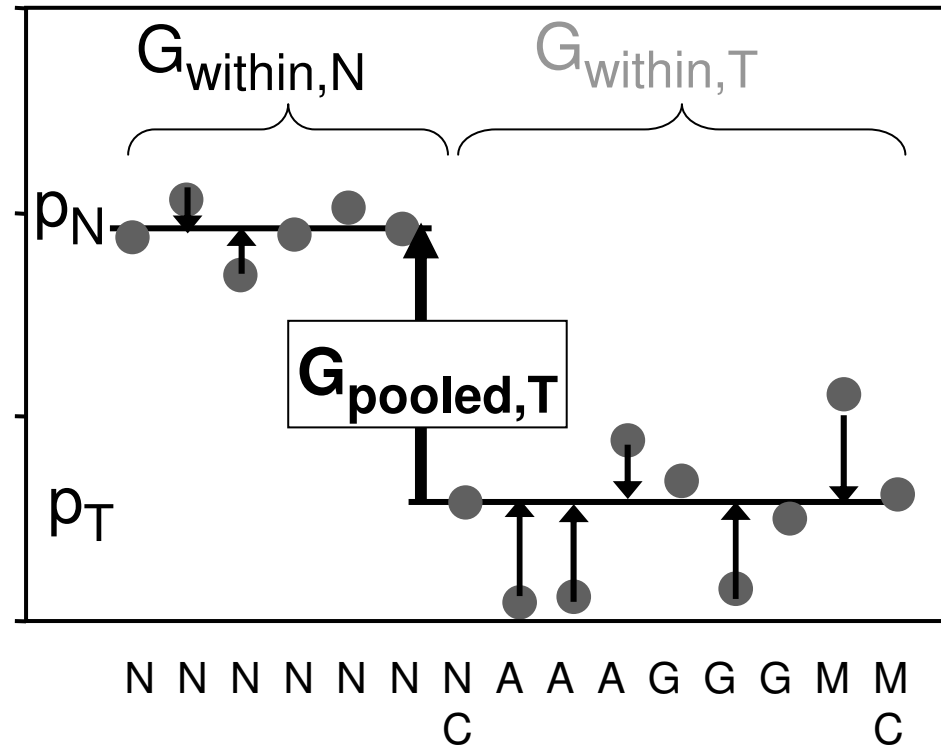
- $G_{intrinsic}$ $P < \alpha$

supervised test:

- $G_{within, standard}$ $P > \alpha$

- G_{pooled} $P < \alpha$

- $G_{within, other set}$ $P > \alpha$



Supervised versus Unsupervised

Requires:

extrinsic information

Subsets differ when:

overall test:

- $G_{\text{intrinsic}} \quad P < \alpha$

supervised test:

- $G_{\text{within,standard}} \quad P > \alpha$

- $G_{\text{pooled}} \quad P < \alpha$

- $G_{\text{within,other set}} \quad P > \alpha$

When not available:

Search for:

homogeneous subsets

combination of subsets
with:

- lowest $\sum G_{\text{within}}$

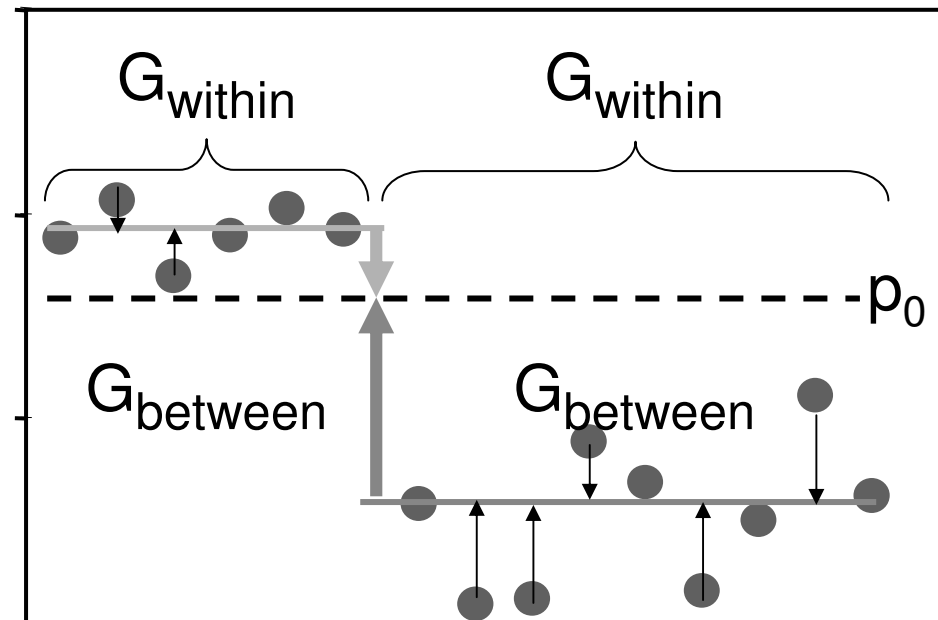
- highest $\sum G_{\text{between}}$

Unsupervised clustering of libraries

G-statistic: $G_{intrinsic} = \sum_{N+T} G_{within} + \sum_{N+T} G_{between}$

G_{within} : variation within subset

$G_{between}$: difference between subset and p_0



$$G_{intrinsic} = \sum(G_{individual}) = \sum(G_{within}) + \sum(G_{between})$$

Unsupervised clustering of libraries

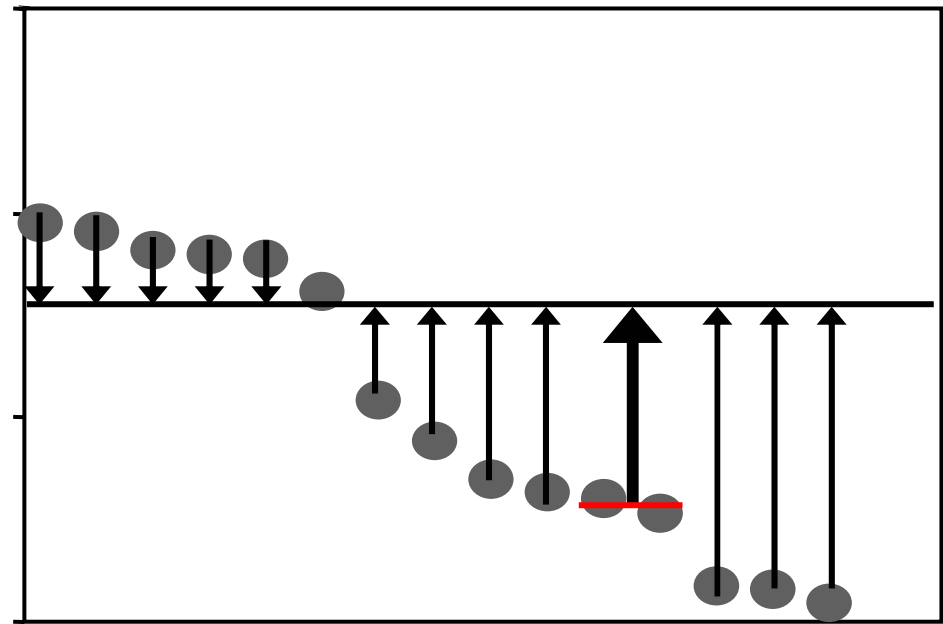
G-statistic: $G_{intrinsic} = \sum_{N+T} G_{within} + \sum_{N+T} G_{between}$

combination of subsets with:

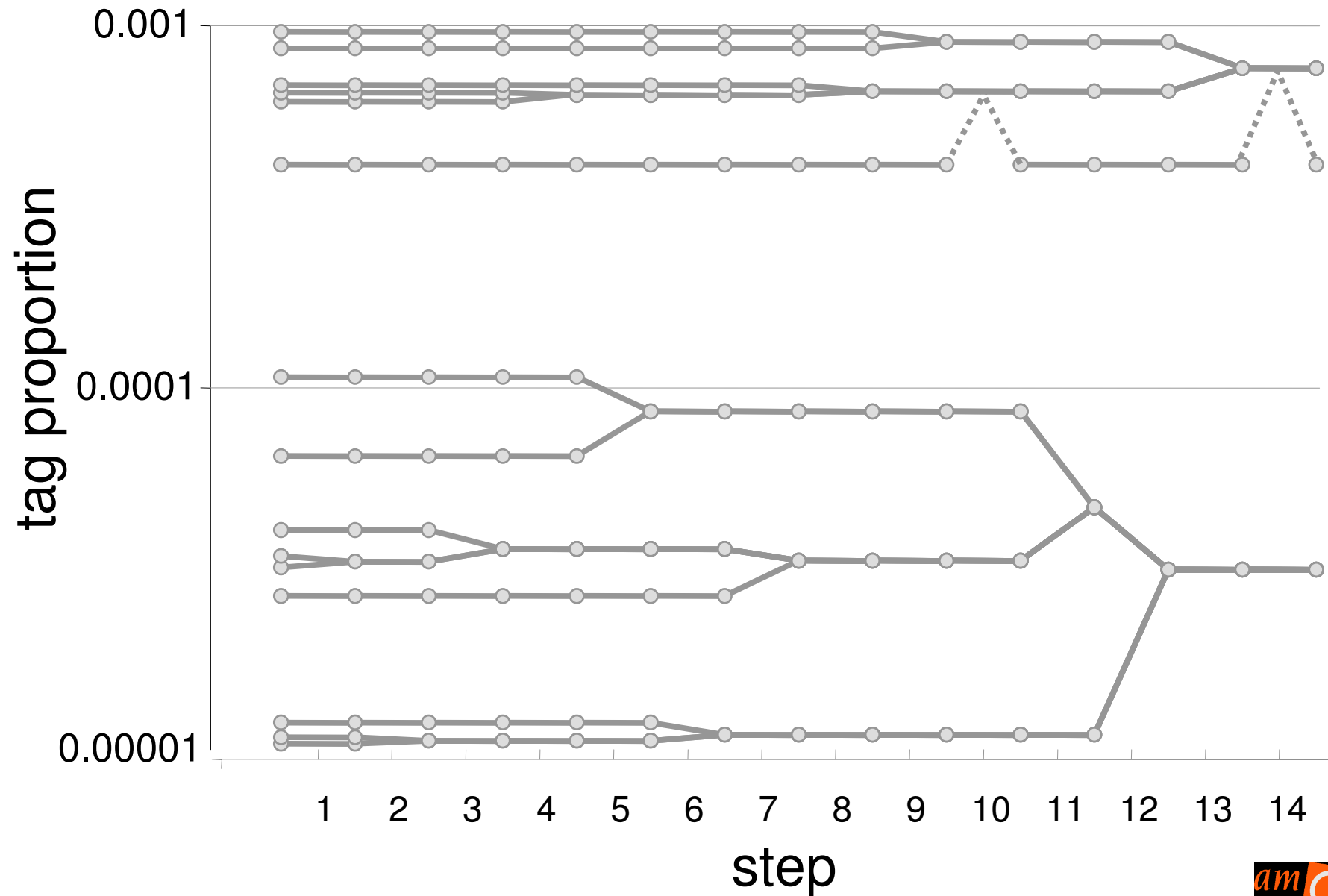
- lowest $\sum G_{within}$
- highest $\sum G_{between}$

Stepwise method:

- sort libraries by proportion
- start with closest pair
- add library to subset if $G_{within,subset} < G_{critical}$
- continue until all libraries are in a subset or $\sum G_{between} < G_{critical}$



Unsupervised clustering of libraries



Statistical comparison of SAGE libraries: *from two to many*

In summary:

two

all tests for comparing two libraries

lead to the same decisions

(Ruijter, van Kampen, Baas. **Physiol Genomics** 11, 2002)

many

the G-statistic allows the testing of
the heterogeneity between libraries

as well as

the determination of deviating libraries

and / or

the search for homogeneous subsets

(Schaaf, Ruijter et al. **FASEB J** 19, 2005)

Exercises: room L- 005

Programs:

Start - **MIK** - Sagestat - G-test
SAGEstat

Excel file with exercises:

<http://amc-app1.amc.sara.nl/EDUwiki/>

on

"Introduction to bioinformatics" page

under

"Documents" and / or "Course exercises"

copy Excel file to your desktop

after the exercise:

use Webmail to mail the results to yourself