

STATISTICAL ANALYSIS OF TRANSCRIPT COUNTS

Jan Ruijter

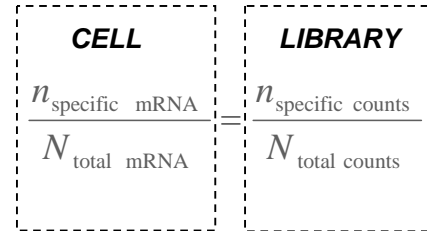
Introduction Bioinformatics, February 2010

Anatomy & Embryology 

Transcript count libraries

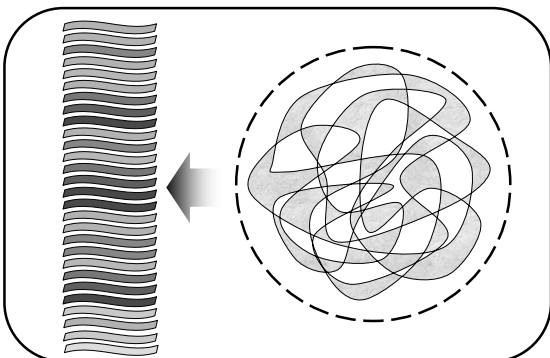
assumption:

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



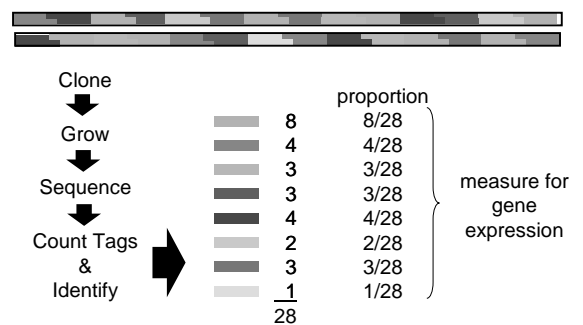
Anatomy & Embryology 

Introduction to SAGE



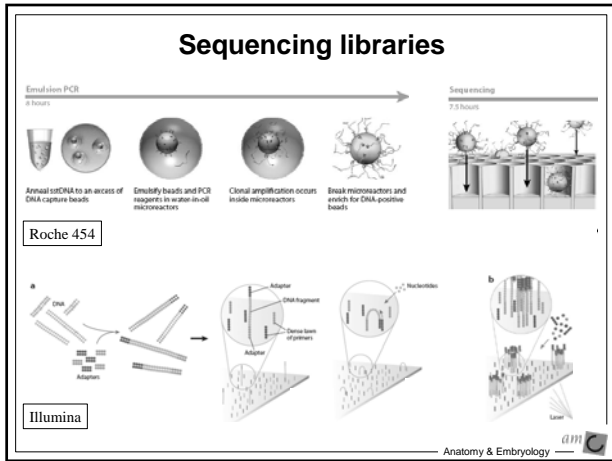
Anatomy & Embryology 

Introduction to SAGE



SERIAL ANALYSIS OF GENE EXPRESSION

Anatomy & Embryology 



Comparison of two SAGE Libraries

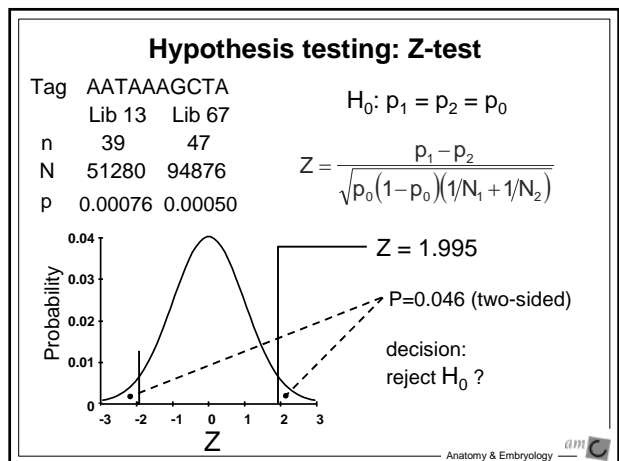
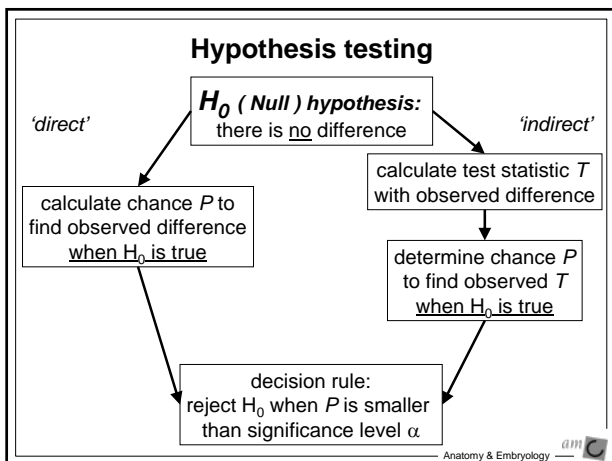
library Name	normal_cerebellum	BB542_whitematter
species	Hs	Hs
library Type	Normal	Normal
tissue Type	Brain	Brain

Library	sage_lib13	sage_lib67
Tag		
AGAAAGATGT	2	1
AACGACCTCG	16	35
AACTGCTCA	2	2
ACCCTTCCT	2	7
AAGGAATCGG	0	3
AATAAAGCTA	39	47
AAGCATTAAA	16	55
ACAACAAGA	35	42
ACAACACTAC	23	48
AAATAAAGCC	45	8
AAATAAAGA	2	0
ACTTTTGCC	4	3
Total	51280	94876

test per tag

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

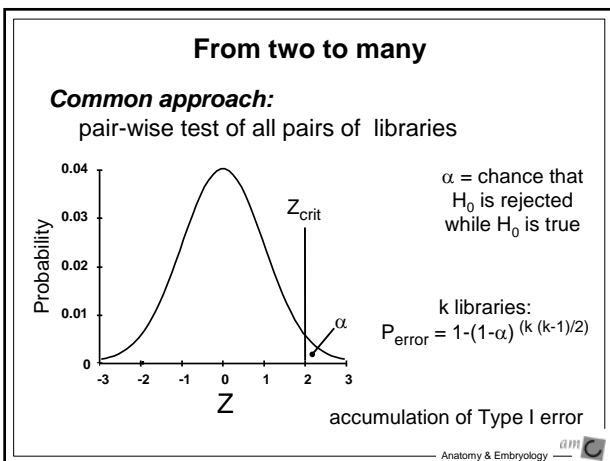
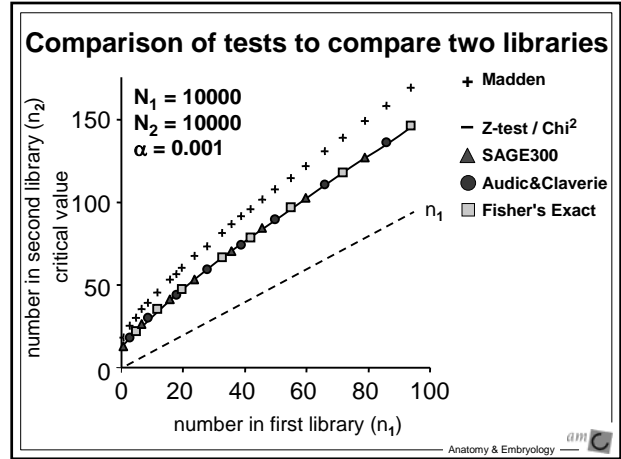
Anatomy & Embryology



Tests used for comparison of two libraries

Z-test (Kal et al. 1999)	$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(1/N_1 + 1/N_2)}}$	} reject H_0 when $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$
Chi-squared test	$Chi^2 = \sum \{(n - n_0)^2 / n_0\}$	
Madden et al. 1997	$Z = \frac{n_1 - n_2}{\sqrt{n_1 + n_2}}$	} $P < \alpha/2$
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)	$P(n_2 n_1) = \frac{\binom{N_2}{n_2} \binom{N_1}{n_1 + n_2 - n_2}}{\binom{N_1 + N_2}{n_1 + n_2}}$	
SAGE300	P from Monte Carlo simulation	

am Anatomy & Embryology



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

am Anatomy & Embryology

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: $G = 2 \cdot \text{Ln}(\text{Likelihood ratio})$

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

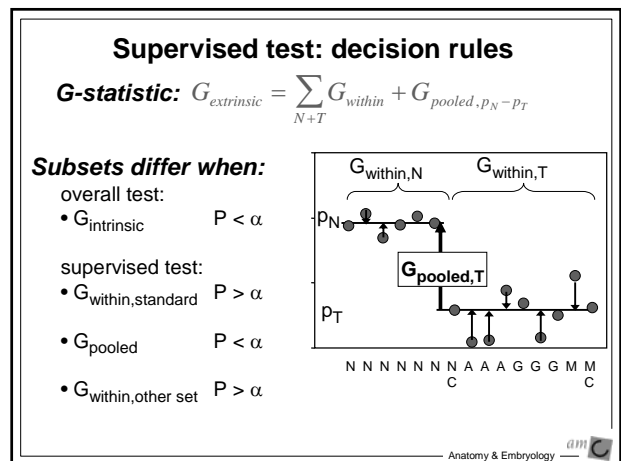
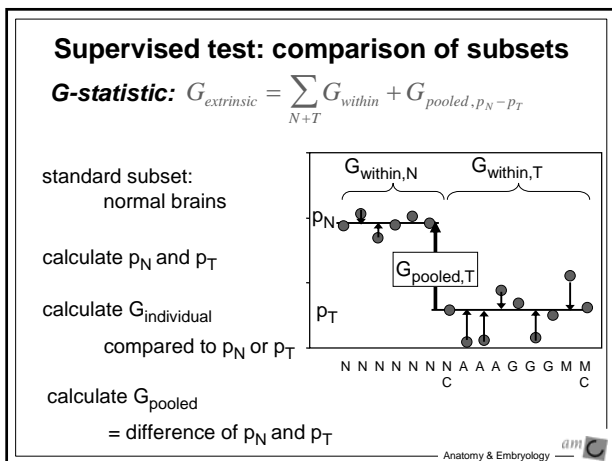
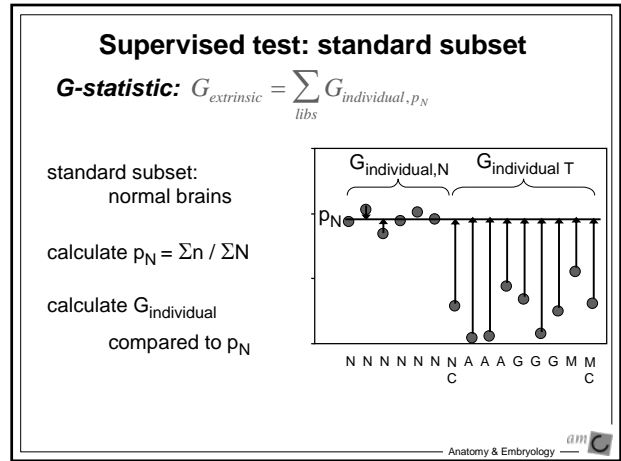
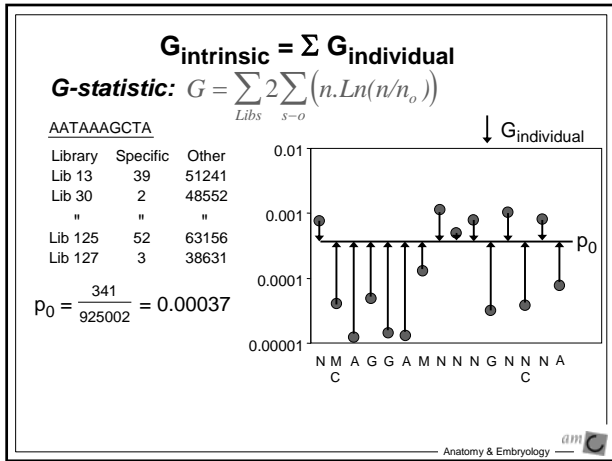
G-statistic

G-statistic: $G = 2 \sum_{\text{Libs}} \sum_{s=0} (n \cdot \text{Ln}(n/n_0))$

AATAAAGCTA

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

n = observed number
 n_0 = expected number
when H_0 is true



Supervised versus Unsupervised

Requires:

extrinsic information

Subsets differ when:

overall test:
 • $G_{\text{intrinsic}}$ $P < \alpha$

supervised test:
 • $G_{\text{within,standard}}$ $P > \alpha$

• G_{pooled} $P < \alpha$

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

When not available:

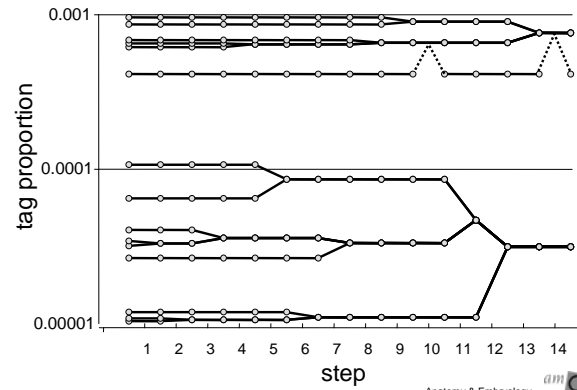
Search for:
 homogeneous subsets

combination of subsets
 with:

• lowest ΣG_{within}

• highest $\Sigma G_{\text{between}}$

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)