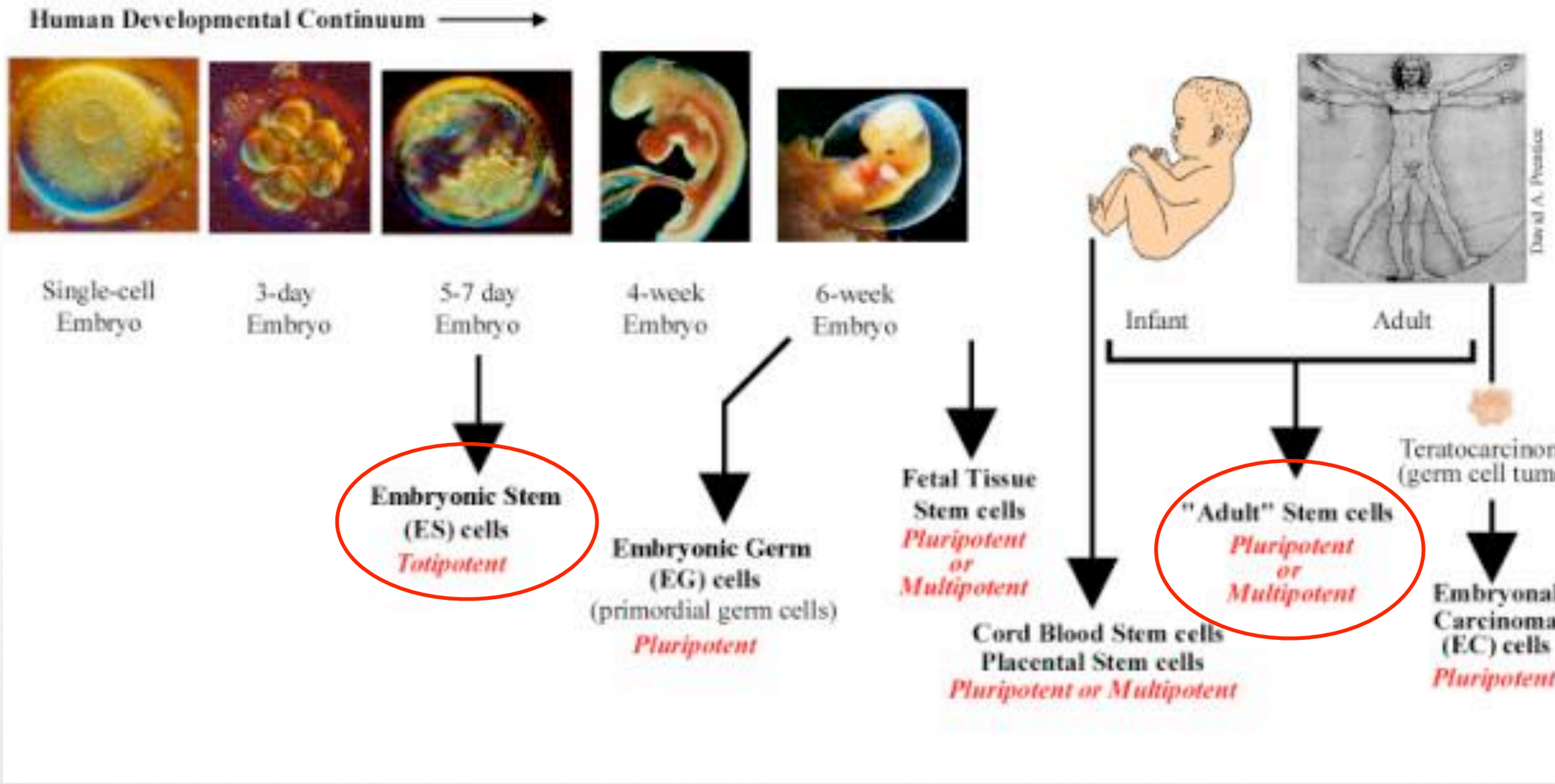
Core stemness

Martina Koeva

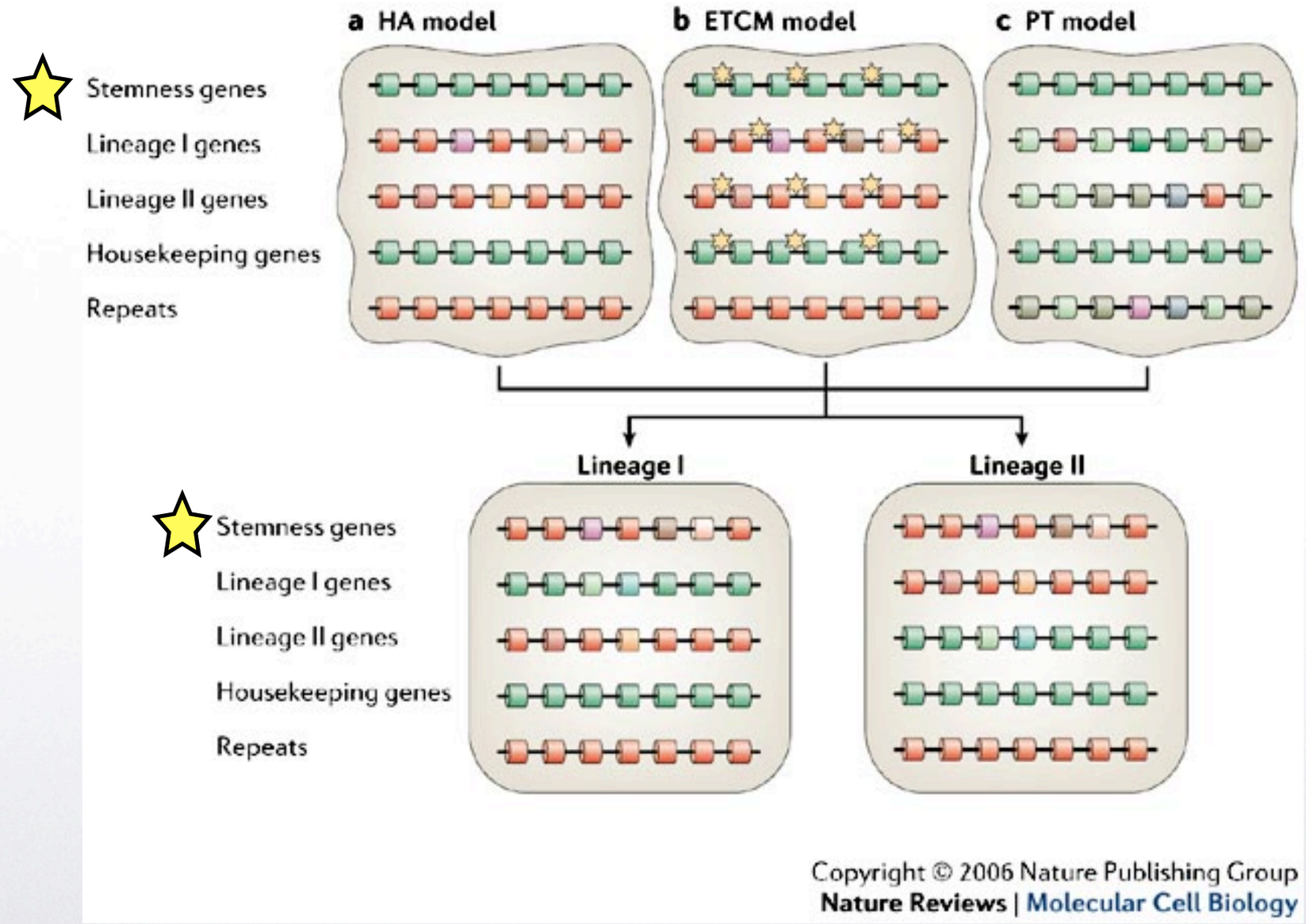
Lab meeting - 09/24/2008

Stem Cells



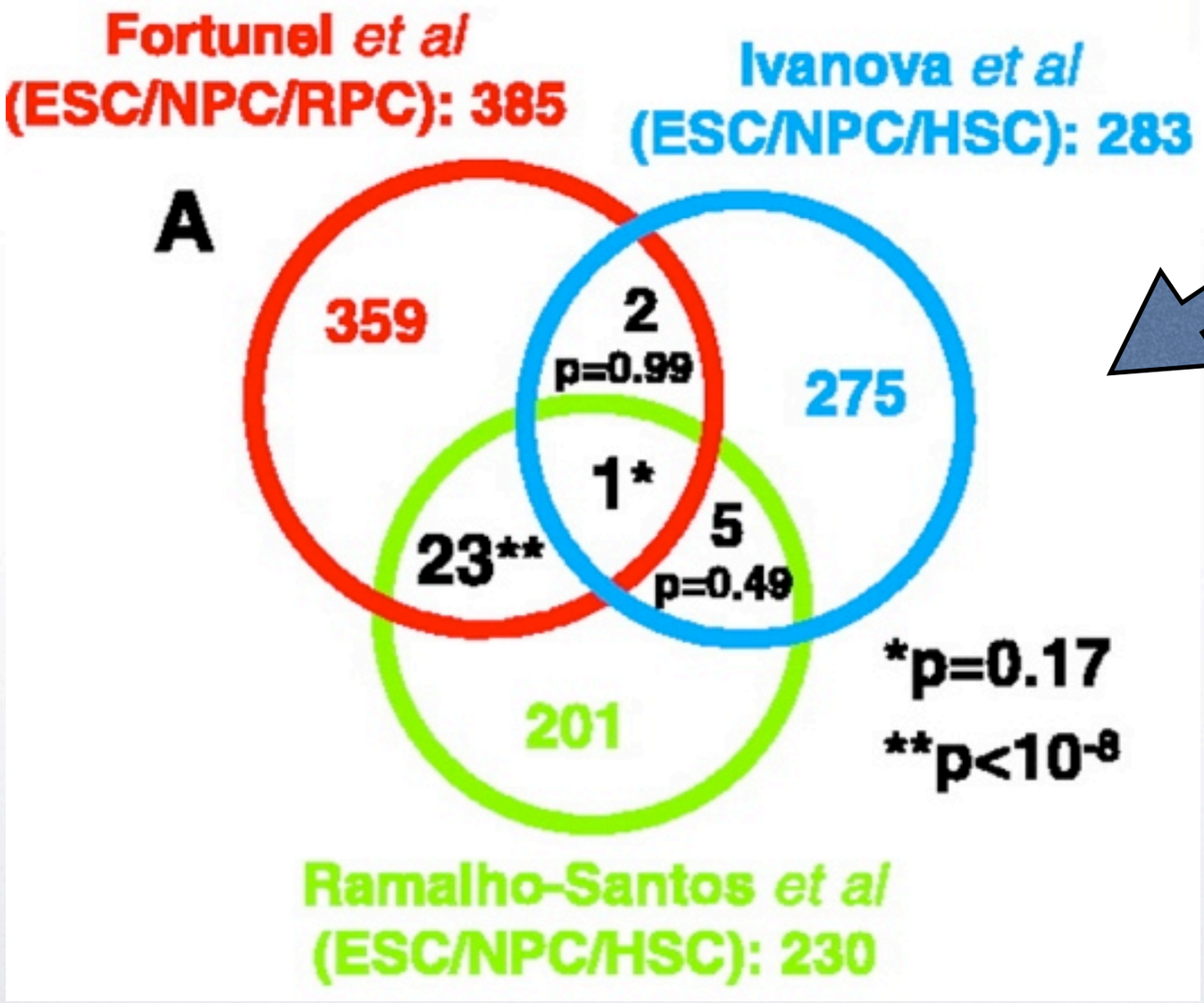
http://robby.nstemp.com/photo6.html

The concept of stemness



Copyright © 2006 Nature Publishing Group
Nature Reviews | Molecular Cell Biology

The stemness gene



Itga-6

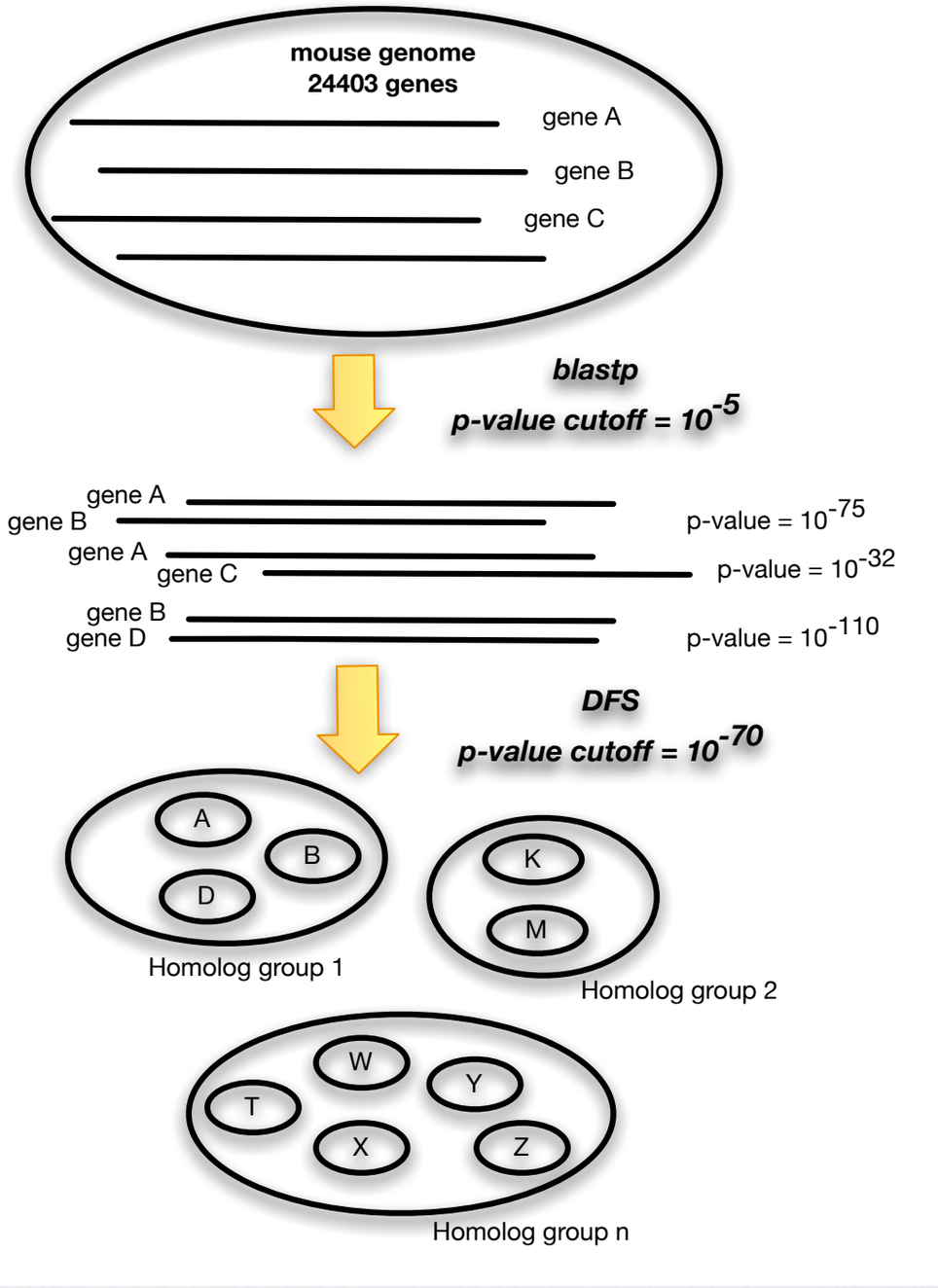
Our hypothesis

- We hypothesize that different stem cells may employ common mechanisms or pathways even though they express distinct repertoires of genes
- Global approach: identify common *homolog groups* or *pathways*

Methods

-blastp

-entire mouse proteome

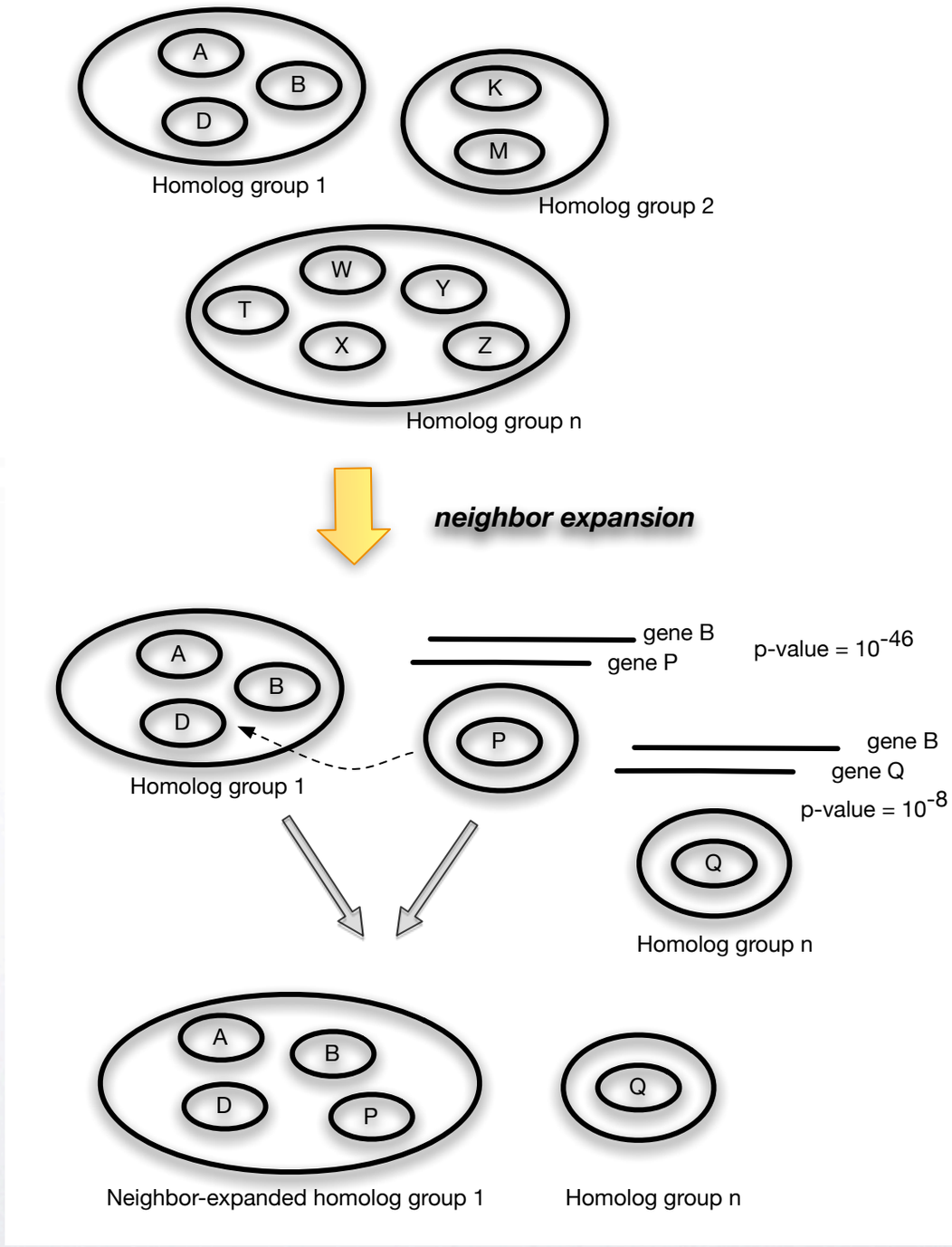


Methods

neighbor expansion

=

homolog group boosting with singletons

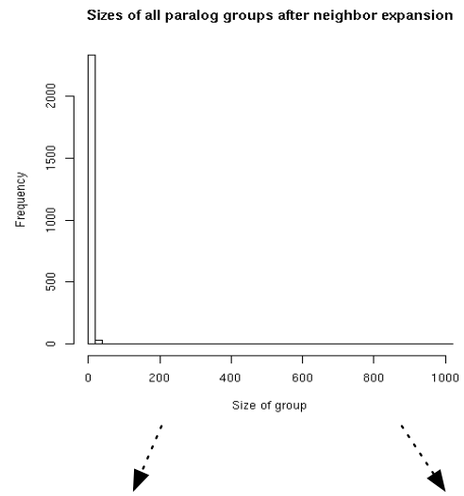


Neighbor expansion

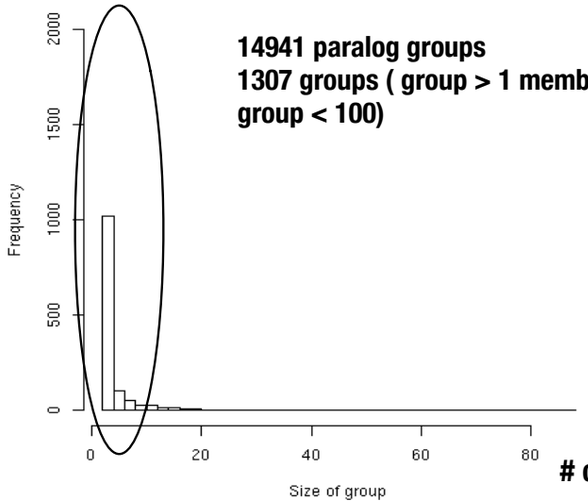
Statistics

Before expansion:
Mean = 5.9 genes/group,
Median = 3
Std = 30.5
Min = 2 genes/group (singletons excluded)
Max = 1008 genes/group

After expansion:
Mean = 4.998 genes/group
Median = 3
Std = 23.8
Min = 2 genes/group (singletons excluded)
Max = 1018 genes/group



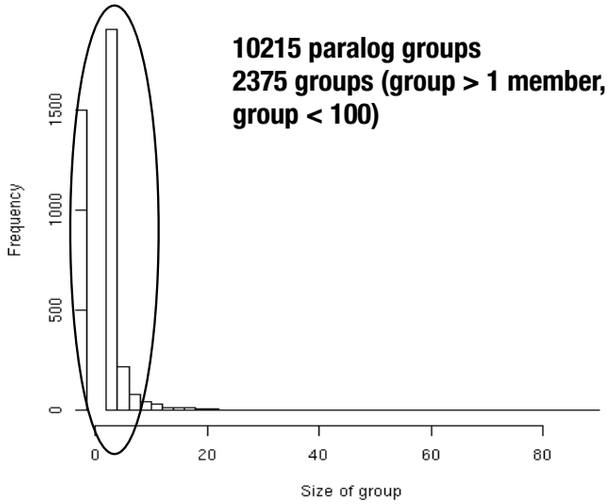
Sizes of paralog groups with less than 100 members from strict cutoff



14941 paralog groups
1307 groups (group > 1 member, group < 100)

neighbor expansion

Sizes of paralog groups with less than 100 members after neighbor expansion

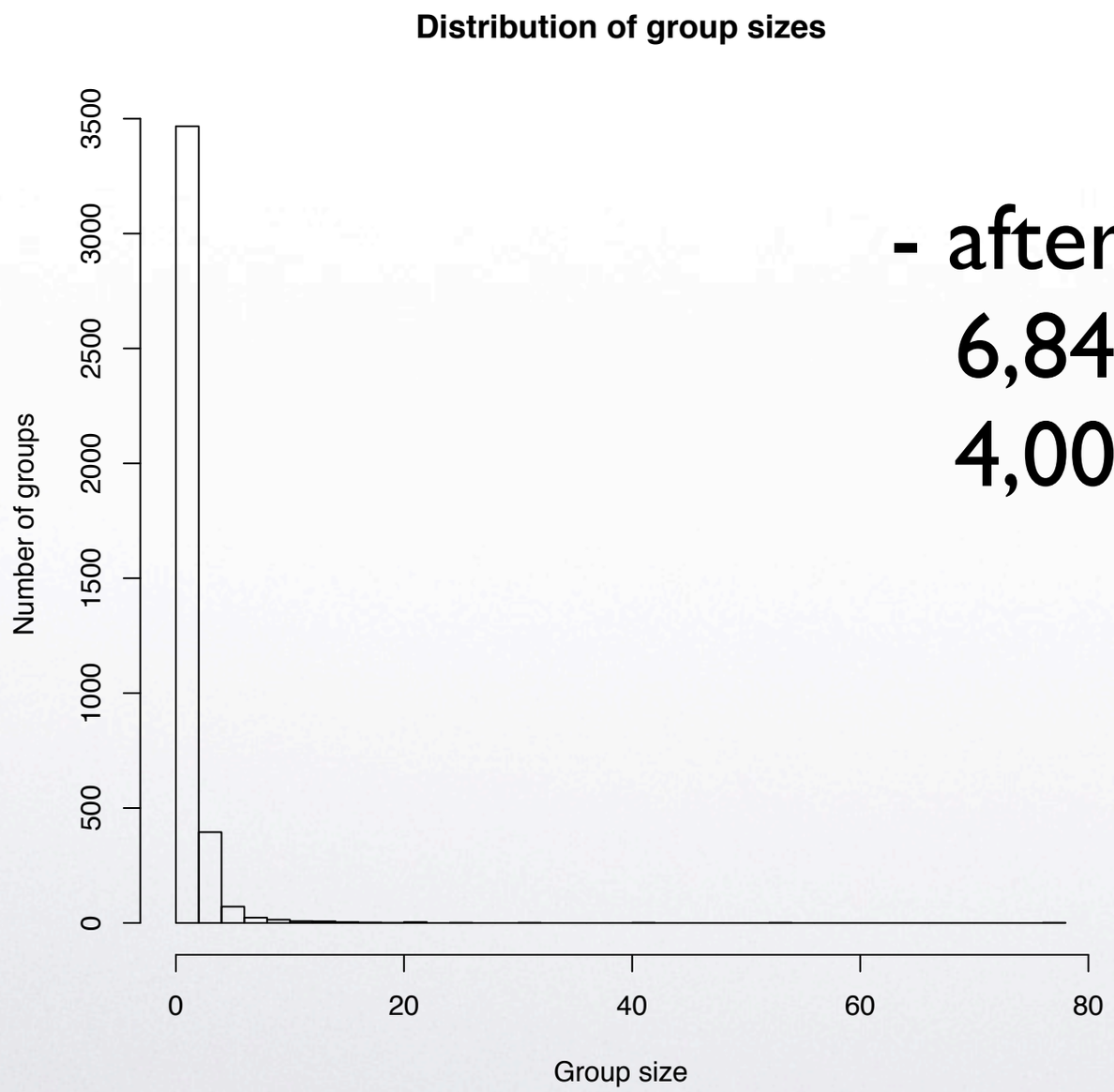


10215 paralog groups
2375 groups (group > 1 member, group < 100)

# of paralog groups

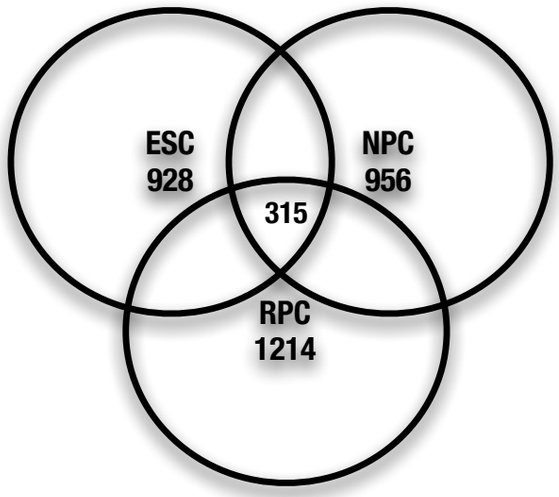
# of small paralog groups

Restricted expanded groups

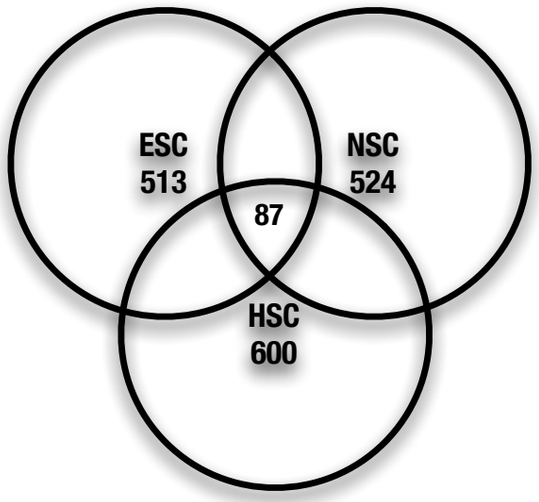


- after restriction to tested genes
  6,843 genes
  4,001 homolog groups

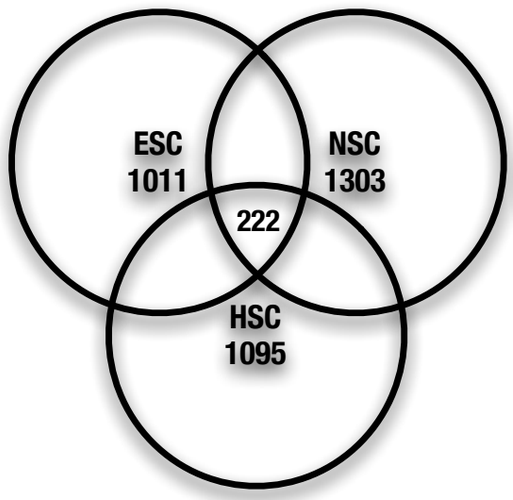
Common groups



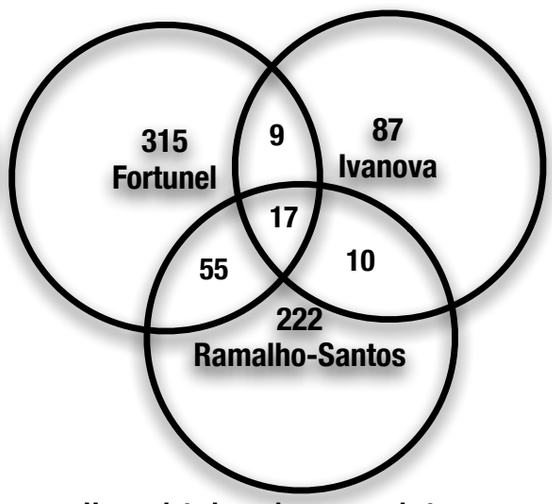
Fortunel dataset - upregulated paralog groups



Ivanova dataset - upregulated paralog groups



Ramalho-Santos dataset - upregulated paralog groups



Upregulated paralog groups between all tested datasets

17 common homolog groups to all experiments

Scoring

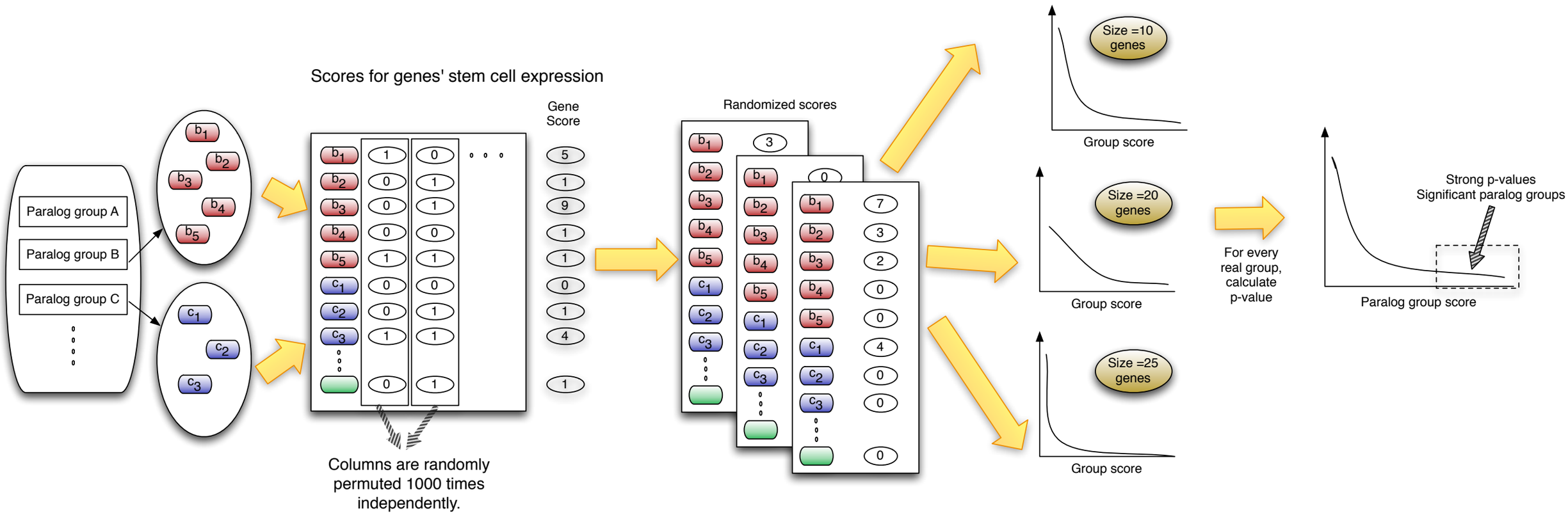
- Whole score types
  - average score
  - max score
  - a gene-weighted score
- Partial score types
  - between-study score
  - within-study score
  - between-tissue score
  - within-tissue score

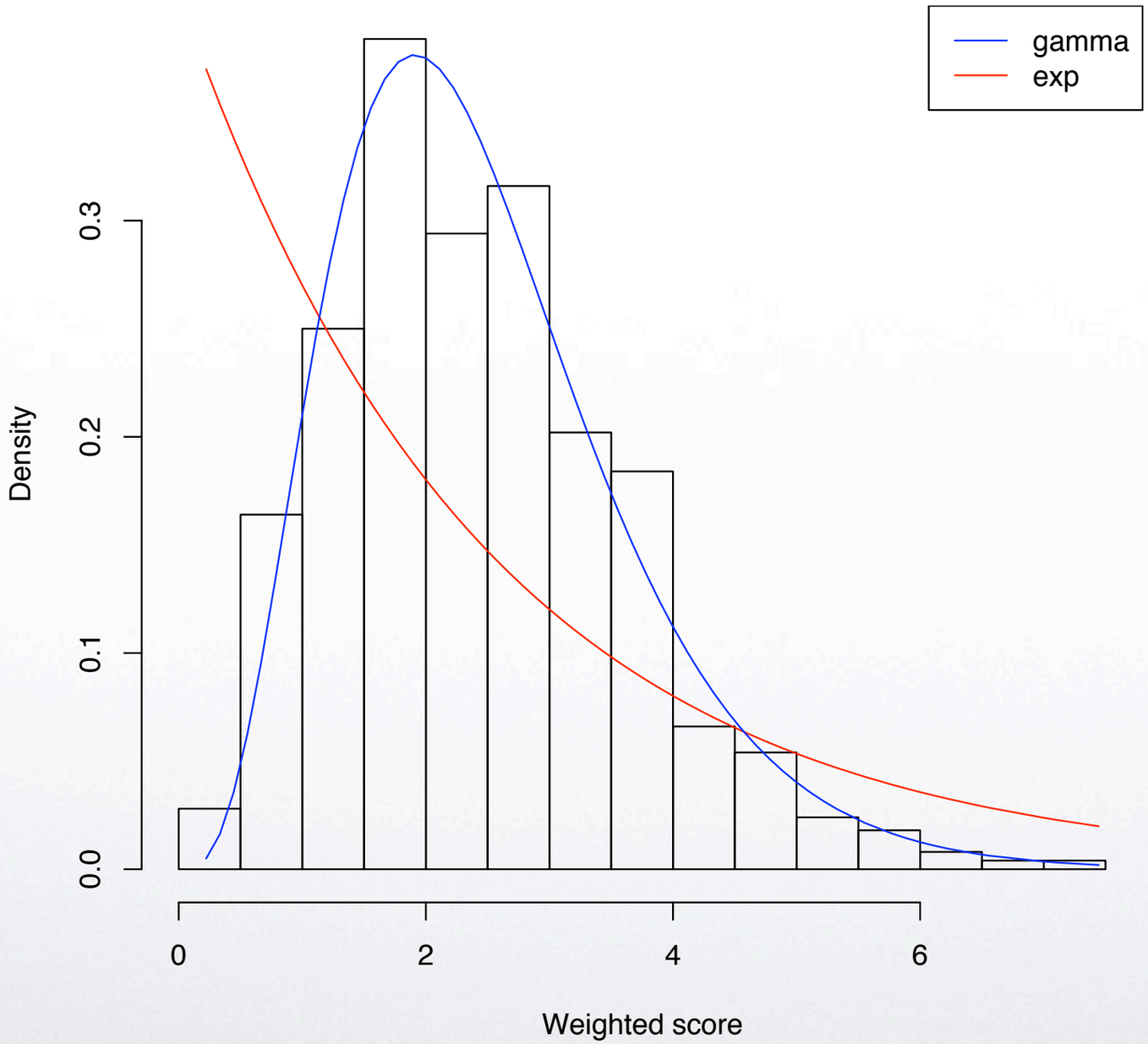
$$Score(g) = \sum_{i=1}^{n} \frac{(\sum_{j=1}^{9} x_{ij})^2}{9}$$

n = the number of genes in group g

$$x_{ij} = \begin{cases} 1 & \text{if gene } x_i \text{ is upregulated in list j} \\ 0 & \text{otherwise} \end{cases}$$

Assessment of significance



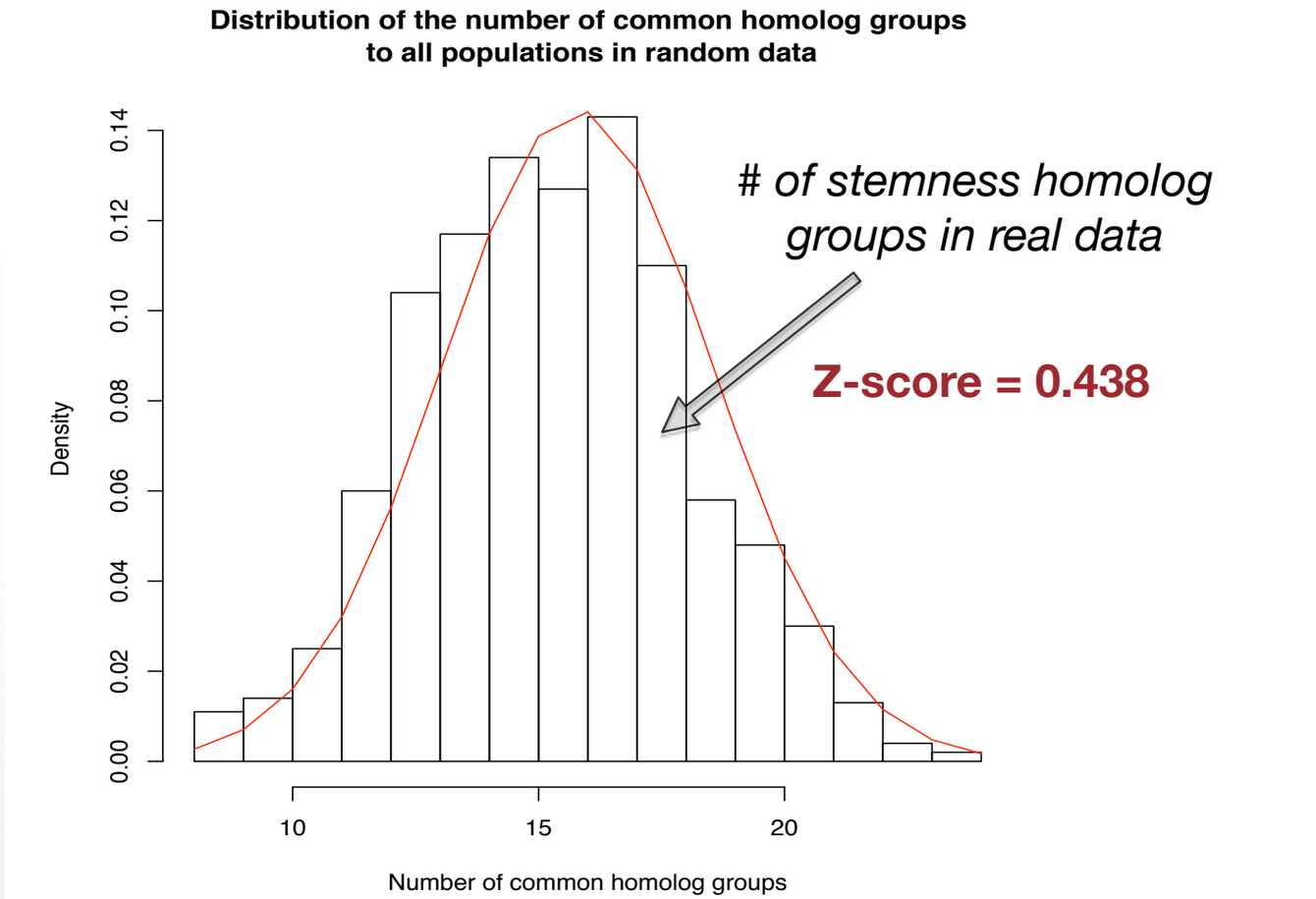

Distribution of randomized weighted scores for a group of size 7
gamma
exp
Density
0.0
0.1
0.2
0.3
0
2
4
6
Weighted score

Group	Group	Size	Average score p-value	Max score p-value	Gene-weight score p-value
104	**integrin alpha**	7	0.0858	1.23E-06	**3.62E-05**
1128	**myc**	5	0.0044	0.0003	**2.72E-05**
1154	**map kinase**	10	0.0897	0.0029	**0.0088**
128	**kinesin/spectrin**	26	0.1752	0.0714	**0.0263**
156	**keratin**	41	0.9999	0.0998	0.9500
195	**collagen/notch/delta-like**	77	0.7165	0.0003	**0.0009**
198	**laminin**	16	0.7211	0.0047	0.1837
2632	**melanoma antigen**	6	0.0089	0.0185	**0.0029**
281	**protein kinase C**	19	0.7224	0.2832	0.6016
286	**myosin**	31	0.7421	0.0040	0.1105
360	**oncogenes/tyrosine kinases**	21	0.9724	0.0513	0.4930
396	**protein tyrosine phosphatases**	21	0.3740	0.3084	0.1774
4117	**coagulation factor receptors**	3	0.0082	0.0024	**0.0011**
437	**serpins**	10	0.6048	0.0029	0.0688
701	**ldlrp**	9	0.1017	0.1512	**0.0409**
8	**zinc fingers**	54	2.46E-06	0.0115	**3.30E-07**
92	**tubulin**	12	0.1509	0.0328	**0.0392**

Other studies

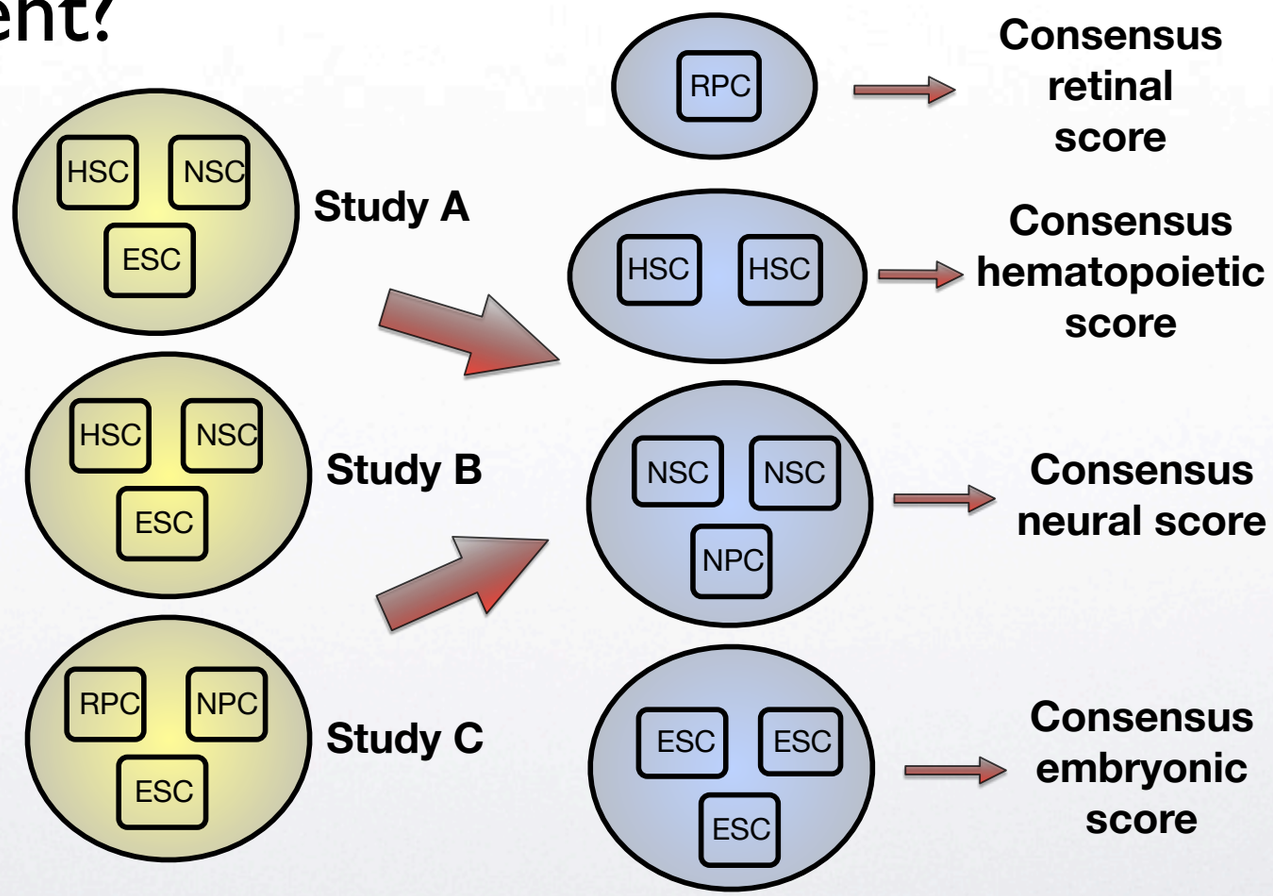
GroupName	Group	Combined	Forsberg	Akashi
Cluster104	integrin alpha	√	√	
Cluster1128	myc	√	√	
Cluster1154	map kinase	√		
Cluster128	kinesin/spectrin	√	√	√
Cluster156	keratin	√	√	√
Cluster195	collagen/notch/others	√	√	√
Cluster198	laminin	√	√	√
Cluster2632	melanoma antigen	√	√	√
Cluster281	protein kinase	√		
Cluster286	myosin	√	√	√
Cluster360	oncogenes/tyrosine kinases	√		
Cluster396	protein tyrosine phosphatases	√	√	√
Cluster4117	coagulation factor receptors	√	√	
Cluster437	Serpins	√	√	√
Cluster701	low-density lipoprotein receptor-related proteins	√		√
Cluster8	zinc fingers	√	√	√
Cluster92	tubulins	√	√	

Significance of the number of groups



Transformation of data

- Are data points (expression in same population) independent?

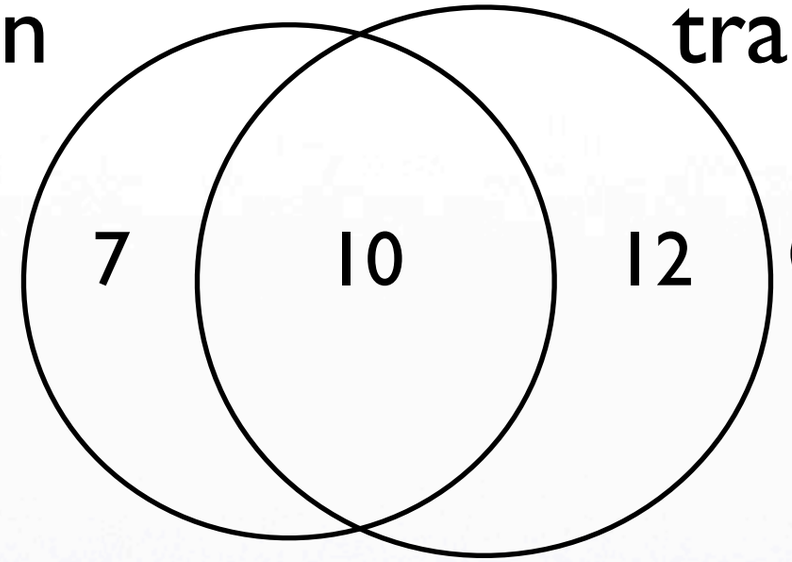


Stemness homolog groups

before transformation

17

(Z-score = 0.438)



after transformation

22

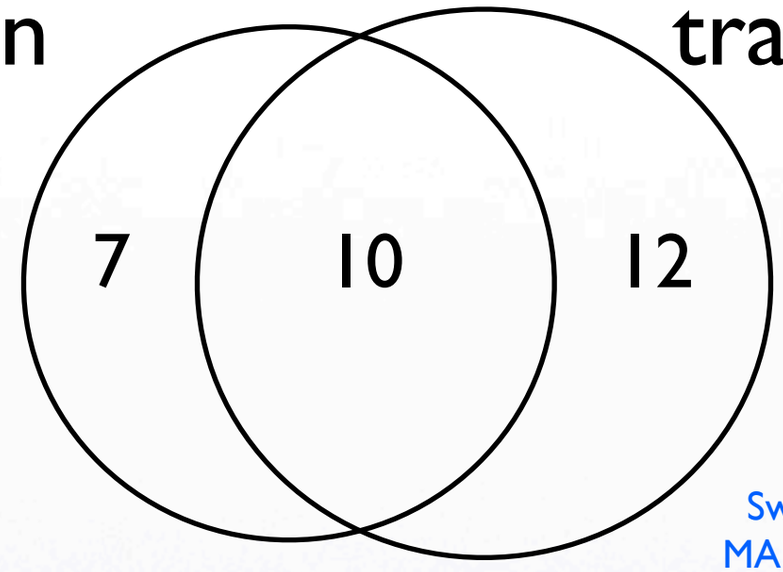
(Z-score = 0.830)

Stemness homolog groups

before transformation

17

(Z-score = 0.438)



after transformation

22

(Z-score = 0.830)

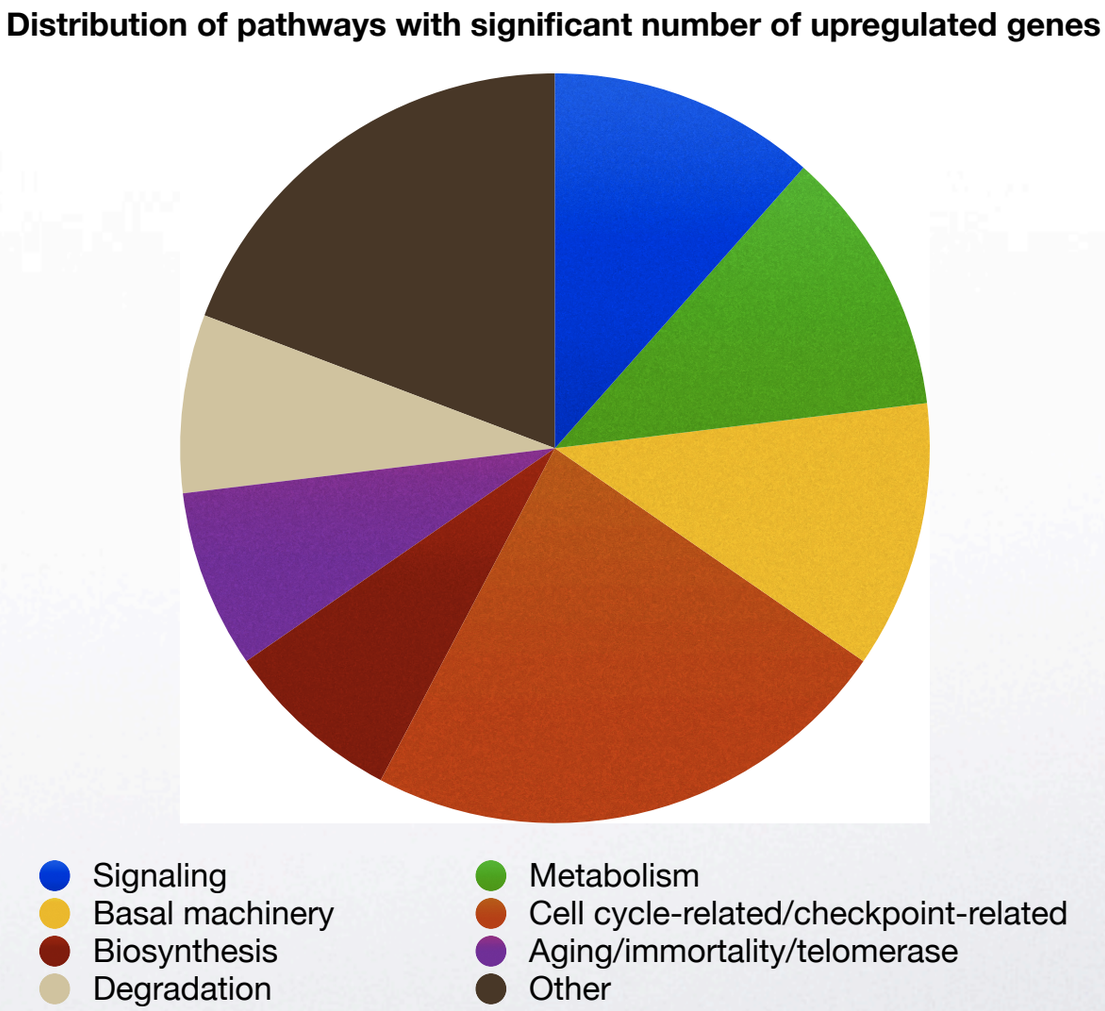
Mapk family
kinesin/spectrin
myosin
oncogene/tyrosine kinase family (Lck)
thrombin family
Serpin family
tubulin

integrin alpha
Myc
keratin
Collagen/Notch/Delta-like
laminin
melanoma antigen
protein kinase C
Ptpr
Lrp
zinc finger family

Riken cDNA 4930507D05 gene (1 gene)
Swi/Snf related/chromodomain helicase family (8 genes)
MAP/microtubule affinity-regulating kinase family (6 genes)
nucleolar protein family (3 genes)
ring finger protein 138 (1 gene)
tripartite motif containing family (2 genes)
phosphoribosyl pyrophosphate synthetase (2 genes)
dystroglycan 1 (1 gene)
spermine synthases (2 genes)
cyclin-dependent kinase family (8 genes)
polyhomeotic-like family (2 genes)
Hey1 (1 gene)

Initial pathway analysis

- 354 pathways from Biocarta and Kegg
- 1551 genes represented
- 73 common patways - at least one gene overexpressed in each experiment
- Initial assessment: is there a significant number of genes from the pathway among the upregulated genes in each experiment?



Future work

- Short-term
  - ✓ Homolog analysis - look for TFBS, PFAM domains, or other features that distinguish upregulated from non-upregulated genes
  - ✓ Homolog analysis - cluster groups after reduction of data and distinguish between “single gene-strong expression” group pattern and “many genes-weak expression” group pattern
  - ✓ Pathway analysis - Assess significance of scores for individual pathways
- Long-term
  - ✓ Look for ES-specific, neural-specific, hematopoietic-specific groups (already done within homolog analysis)
  - ✓ Incorporate other stem cell data (2 more studies already compared)
  - ✓ Incorporate downregulated data