

# Meta-analysis of heterogeneous data

MAX-PLANCK-INSTITUT  
FÜR MOLEKULARE GENETIK

Ralf Herwig

Max Planck Institute for Molecular Genetics

Innestr. 73, 14195 Berlin

[herwig@molgen.mpg.de](mailto:herwig@molgen.mpg.de)

WERKSTÄTTEN  
NACHTPFORTE

# Structure

**Meta-analysis approach**

**Type-2 diabetes mellitus**

**Trisomy 21**

# Goals of meta-analysis

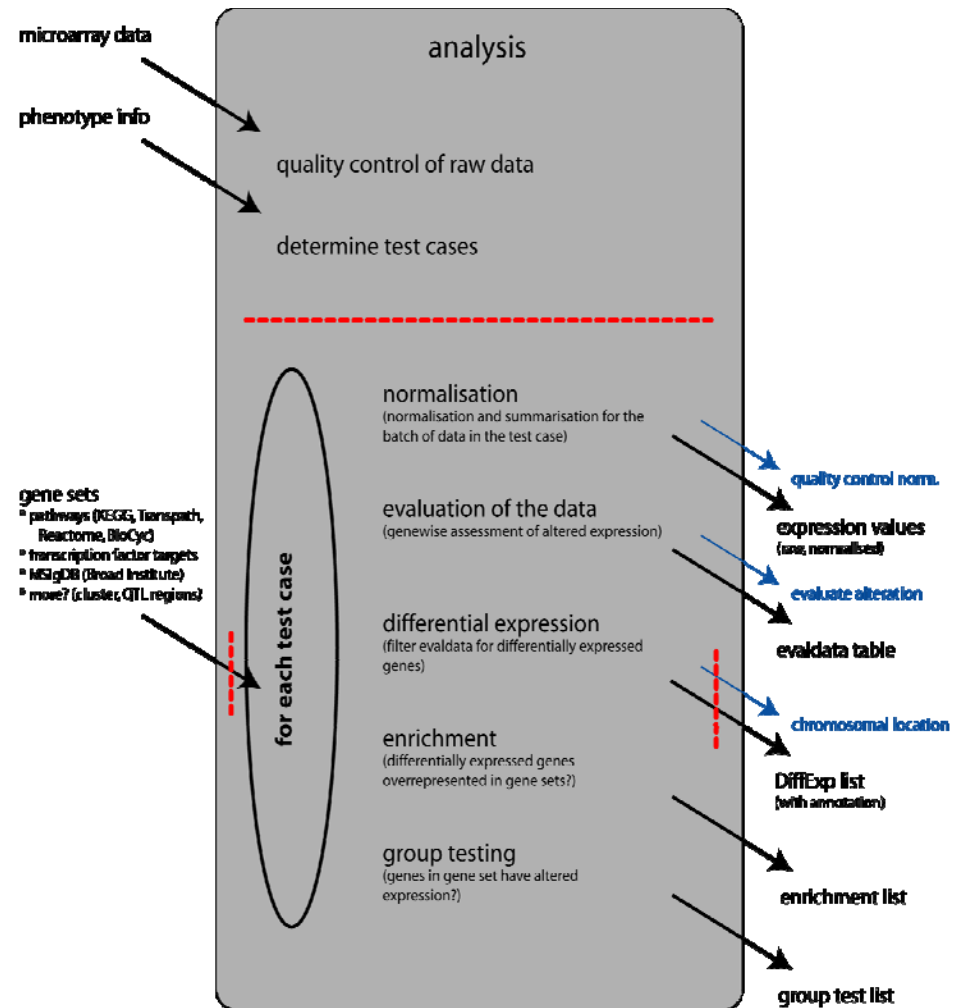
- **to identify new (better?) markers**
  
- **to analyse gene networks**



# Data processing

## Issues to consider:

- re-mapping of probes (alternative cdf files)
- normalisation
- Differential expression
- Functional characterisation of candidate genes



# Scoring

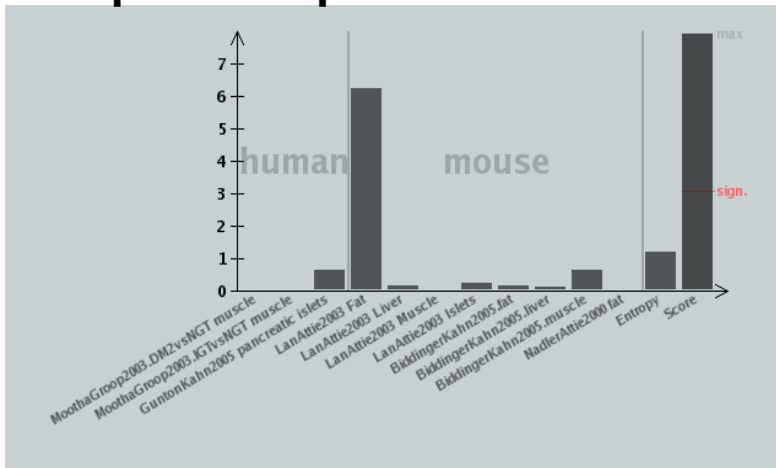
## Quantitative microarray studies

$$s_{ij} = \begin{cases} \left| \log_2(r_{ij}) \right| \left( 1 - \frac{e_{ij}}{r_{ij}} \right) (1 - p_{ij}), & p_{ij} \leq 0.1 \text{ and } e_{ij} / r_{ij} \leq 1 \\ 0, & \text{else} \end{cases} .$$

***r*** : fold change,  
***p*** : average detection P-value  
***e*** : standard error of the ratio

# Meta-analysis – generality vs specificity

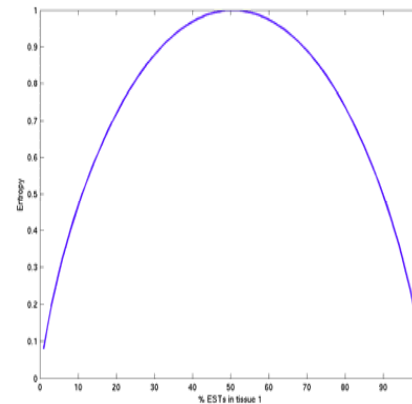
## Serpina1 – specific marker



## PdK4 – general marker



## Entropy criterion:



Number of studies: 2

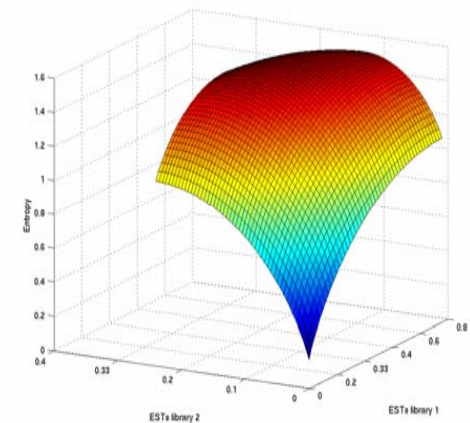
$p$  proportion of study 1  
 $1-p$  proportion of study 2

$$H = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$$

Number of studies: 3

$p$  proportion of study 1  
 $q$  proportion of study 2  
 $1-p-q$  proportion of study 3

$$H = -p \cdot \log_2(p) - q \cdot \log_2(q) - (1-p-q) \cdot \log_2(1-p-q)$$



# Robustifying marker detection

- identified 213 potential marker genes
- reasonable marker set on the pathway level
- higher accordance to OMIM genes
- only limited overlap to previous studies
- 95 new uncharacterised genes
- no significant chromosomal spots

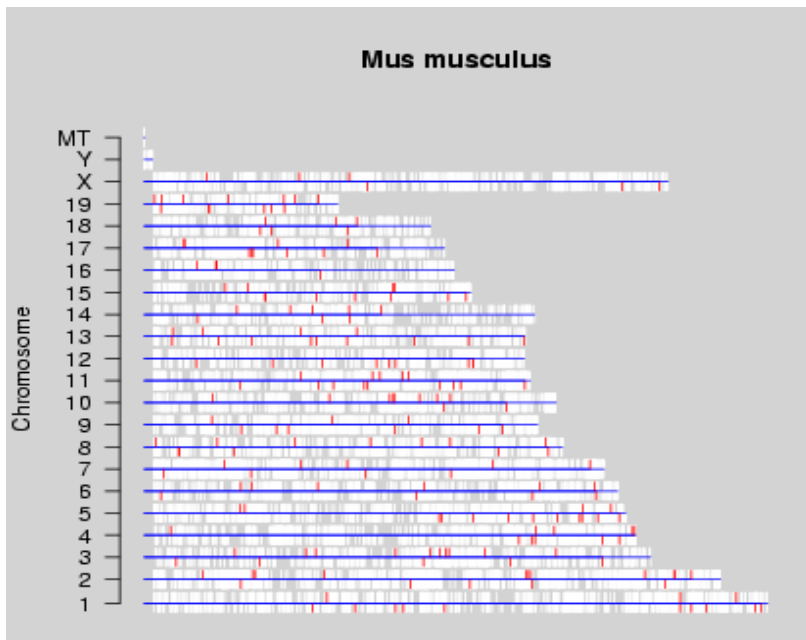


Table 3 Results for T2DM OMIM genes.

SourceName	mgf_symbol	StumvollGoldstein2005	DeanMcEntyre2004	OMIM	PubMedGeneRIF	KOmitceJax	NandiAccili2004	score	entropy	significant gene	rank (out of 15,277)
ENSMUSG00000012705	Retn			*	*			4.597	2.106	*	31
ENSMUSG00000026827	Gpd2			*				4.452	2.88	*	39
ENSMUSG00000023951	Vegfa			*	*			4.273	2.724	*	52
ENSMUSG00000038894	Irs2	*		*	*	*		3.907	2.112	*	82
ENSMUSG00000020679	Tcf2		*	*	*			3.175	1.605	*	176
ENSMUSG00000041798	Gck	*	*	*	*	*		3	1.585		234
ENSMUSG00000029644	Ipf1	*	*	*	*	*		3	1.585		234
ENSMUSG00000029556	Tcf1		*	*	*	*		3	1.585		234
ENSMUSG00000040136	Abcc8	*	*	*	*			2.795	1.848		325
ENSMUSG00000034701	Neurod5 ; Neurod1		*	*	*			2.393	1.48		608
ENSMUSG00000017950	Hnf4a	*	*	*	*			2.36	1.614		642
ENSMUSG00000024985	Tcf7l2			*	*			2.192	1.371		811
ENSMUSG00000037370	Enpp1			*	*			2.106	1.237		918
ENSMUSG00000027223	Mapk8ip1			*				1	0		3013

Supplementary Table 3 Pairwise overlap of different T2DM candidate approaches.

	study	study(lower)	Z200 8	F200 7	OMI M	TH200 6	DGC G	PG200 4	KM200 4	LK200 7
this_study	***	213	0	1	5	6	2	5	12	7
this_study (lower cut-off)	213	***	1	5	12	20	4	13	12	25
Zeggini2008	0	1	***	1	0	0	0	1	1	0
Frayling2007	1	5	1	***	3	0	0	2	2	0
OMIM	5	12	0	3	***	1	0	6	6	2
TiffinHide2006	9	20	0	0	1	***	0	1	4	3
DiabetesGenomeCG	2	4	0	0	0	0	***	1	1	0
ParikhGroop2004	5	13	1	2	6	1	1	***	7	2
Kitano2004	12	38	1	2	6	4	1	7	***	20
LiuKasif2007	7	25	0	0	2	3	0	2	20	***

This\_study: 0.001 significance level; this\_study (lower cut-off): 0.01 significance level; Zeggini2008, Frayling2007: two meta-analyses of GWA studies; OMIM: T2DM genes according to OMIM; TiffinHide2006, DiabetesGenomeCG, ParikhGroop2004, LiuKasif2007: candidate studies using different data sets; Kitano2004: genes used for a physiological model of T2DM.

# Measure enrichment of functional information

- assume that you identify a set of  $n$  genes
- under these  $n$  genes  $k$  have a certain functional category (e.g. *catalytic activity*)

Question:

Are genes with *catalytic activity* enriched in the selected gene set ?

- $N$  number of all genes under consideration  
 $n$  number of differentially expressed genes  
 $k$  number of genes with the specific category  
 $K$  number of all genes with that category

$$P(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

P-value:

$$p = \sum_{j \geq k} P(j)$$

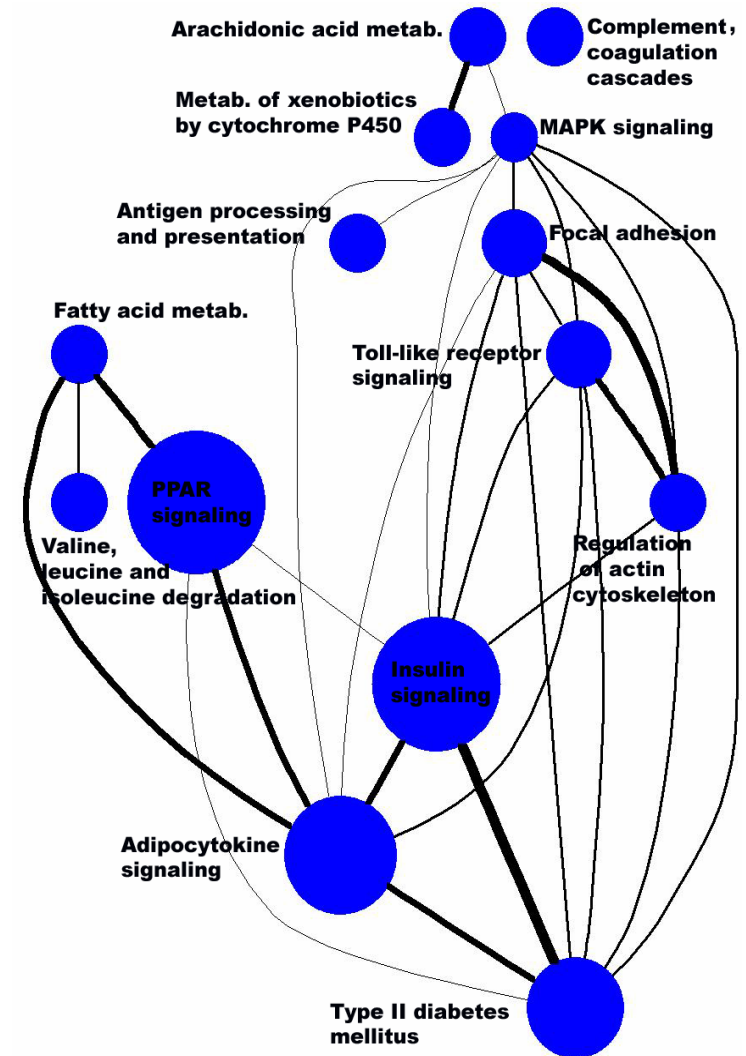
# Over-representation and gene enrichment

Table 5 Gene set enrichment of the most significant KEGG pathways.

Pathway ID	SigSet	Set	Sig	All	P-value	Q-value	Pathway description
path:mmu03320	13	69	213	15274	1.02E-11	1.37E-09	PPAR signaling pathway
path:mmu04920	12	73	213	15274	3.46E-10	1.66E-08	Adipocytokine signaling pathway
path:mmu04930	10	44	213	15274	3.69E-10	1.66E-08	Type II diabetes mellitus
path:mmu04910	13	128	213	15274	2.70E-08	9.09E-07	Insulin signaling pathway
path:mmu04612	6	38	213	15274	1.30E-05	0.000351	Antigen processing and presentation
path:mmu00280	6	44	213	15274	3.11E-05	0.000697	Valine, leucine and isoleucine deg.
path:mmu04610	7	67	213	15274	3.98E-05	0.000764	Complement and coagulation casc.

All are the genes under consideration, Sig the number of candidate genes, Set is the number of genes in the pathway under study and SigSet the overlap of genes in the pathway and the candidate genes. P-values were computed with the upper tail of the hypergeometric distribution indicating the probability of observing this overlap by chance. Q-values are the multiple testing corrected P-values [60, 61].

- simple count statistics based on hypergeometric distribution
- each gene is weighted equally
- no crosstalk between pathways is taken into account

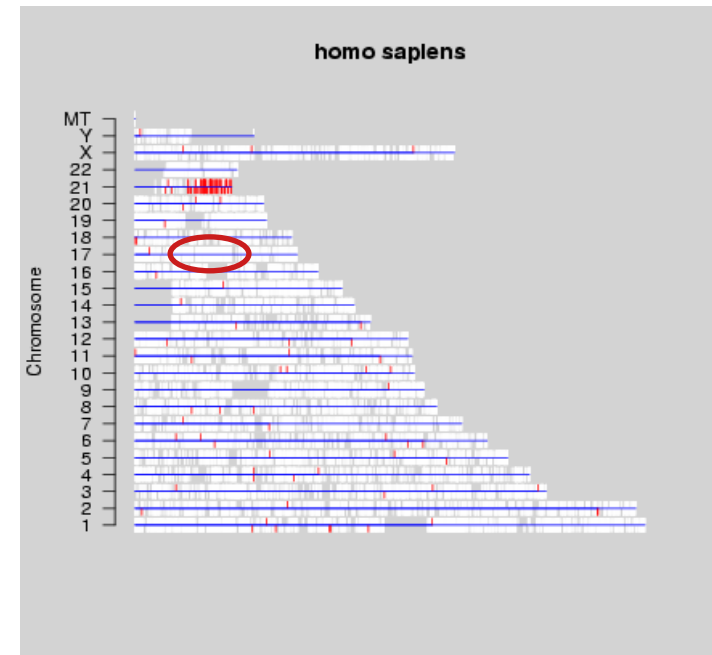
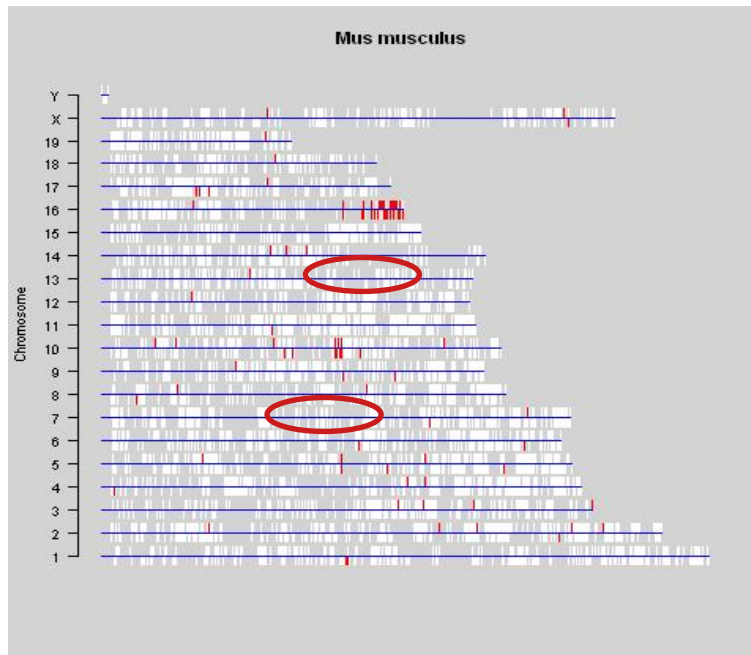


# Marker identification in Down's Syndrome

HUGO_NAME	Score
ADAMTS1	16,462
CBR1	15,094
IFNGR2	14,559
PIGP	14,552
APP	14,372
SON	14,266
TMEM50B	14,024
ITSN1	13,896
SOD1	13,696
TTC3	12,96
DOPEY2	12,947
GART	12,529
USP16	12,291
IL10RB	11,965
BACE2	11,9
MX1	11,62
ZNF294	11,275
RIPK4	10,94
CRYZL1	10,886
ETS2	10,719
RCAN1	10,694
DYRK1A	10,496
ATP5O	10,411
TIAM1	10,365
MORC3	10,307
RUNX1	10,098
IFNAR2	10,053
HLCS	10,048
MX2	9,878
CBR3	9,853
CCT8	9,827
KCNE2	9,654
SETD4	9,55
GABPA	9,478
ICNAR1	9,397

...

- identified 146 potential markers
- identified several new marker genes (29)
- strong enrichment for human chromosome 21 ( $P < 10^{-20}$ ) and mouse chromosomes 16 ( $P < 10^{-20}$ ) and 10 ( $P = 0.002$ )



# Meta analysis – KEGG overrepresentation analysis

path:hsa05010	3	27	146	18273	0,00127	Alzheimers disease - Homo sapiens (human)
path:hsa04670	5	112	146	18273	0,00206	Leukocyte transendothelial migration - Homo sapiens (human)
path:hsa04514	4	103	146	18273	0,00939	Cell adhesion molecules (CAMs) - Homo sapiens (human)
path:hsa04530	4	112	146	18273	0,0125	Tight junction - Homo sapiens (human)
path:hsa01430	4	114	146	18273	0,0133	Cell Communication - Homo sapiens (human)
path:hsa00030	2	25	146	18273	0,0169	Pentose phosphate pathway - Homo sapiens (human)
path:hsa00450	2	29	146	18273	0,0224	Selenoamino acid metabolism - Homo sapiens (human)
path:hsa05040	2	30	146	18273	0,0238	Huntingtons disease - Homo sapiens (human)
path:hsa00052	2	31	146	18273	0,0253	Galactose metabolism - Homo sapiens (human)
path:hsa04512	3	82	146	18273	0,028	ECM-receptor interaction - Homo sapiens (human)
path:hsa04630	4	146	146	18273	0,0297	Jak-STAT signaling pathway - Homo sapiens (human)
path:hsa01510	2	38	146	18273	0,037	Neurodegenerative Disorders - Homo sapiens (human)
path:hsa00750	1	5	146	18273	0,0393	Vitamin B6 metabolism - Homo sapiens (human)
path:hsa04330	2	43	146	18273	0,0463	Notch signaling pathway - Homo sapiens (human)

# Meta analysis – pathway crosstalk

