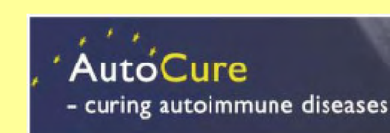
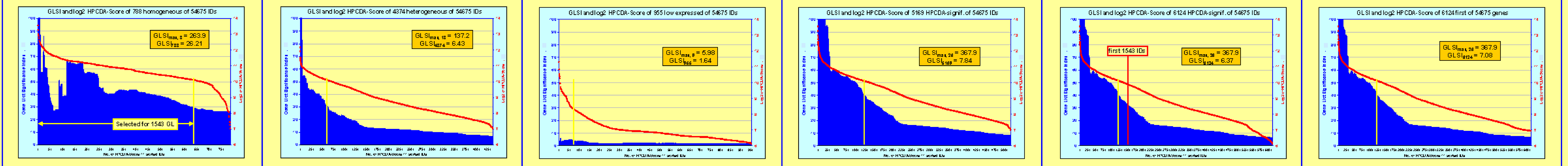
Gene List Significance Index (GLSI) improves our method High Performance Chip Data Analysis (HPCDA) dramatically

Quantifying the quality of different lists of analyzed significant genes

Joachim R. Grün[1], Andreas Grützkau[1], Marta Steinbrich-Zöllner[3], Thomas Häupl[2], Ria Baumgrass[1], Jochen Sieper[3], Gerd-Rüdiger Burmester[2], Andreas Radbruch[1]

[1]) Deutsches Rheumaforschungszentrum (DRFZ) Berlin, [2]) Charité CCM Berlin, [3]) Charité CBF Berlin





1.1: HPCDA significant genes (homogeneous, DC dataset)

1.2: HPCDA significant genes (heterogeneous, DC dataset)

1.3: HPCDA significant genes (low expressed, DC dataset)

1.4: HPCDA significant genes (homo- & heterogeneous, DC dataset)

1.5: All 6124 HPCDA significant genes (DC dataset)

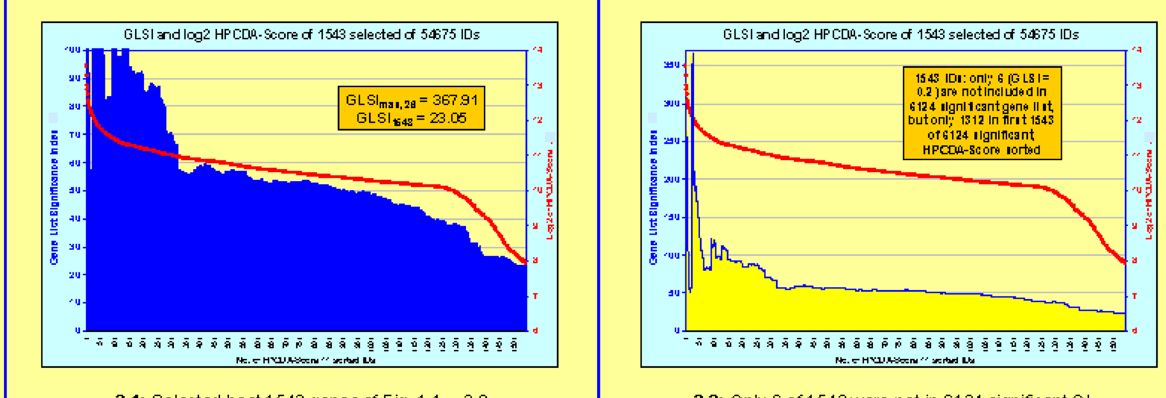
2.1: 6124 first of 54675 genes, HPCDA-Score sorted

Abstract: The Bioinformatics group of DRFZ is mainly involved in gene expression profiling. With the first part of High Performance Chip Data Analysis (HPCDA) a method to detect small numbers of differentially expressed genes (35 – 65) up to more than 10.000 was developed (1 – 3). Both results could be attained without changing any parameters. It is important to use a parameter independent method, because otherwise the numbers of significant genes rely on the users choices. We have validated the first part of HPCDA on data sets with known results and compared our findings to SAM and dChip, both tools dependent on parameters (1). In the meantime, we successfully transferred an HPCDA analogue technique to the red/green miRNA chips of Miltenyi. The second part of HPCDA reduces the list of significant genes to the most relevant ones to classify two or more different groups of chips (the predictive genes). We have shown that it is possible to find signatures of predictive genes with monocytes of RA, AS, OA, and SLE patients in comparison to normal donors. We have transformed this HPCDA analogue technique to FACS data of immune monitoring. With our MS Access database ImmuMon, we extracted significant parameters of 37 individuals (AS, RA, SLE, and ND). Newly analyzed patients and normal donors were correctly classified (PAM and HC) with the reduced list of predictive parameters (4). In total 14237 different FACS parameters were checked for significance. **Results and discussion:** **High Performance Chip Data Analysis (HPCDA)** Our method HPCDA was validated with the Latin Square dataset. We could show that HPCDA outperforms dChip and that SAM has the disadvantage of finding a huge number of false positive genes in some analyses. **Chip data analysis of miRNA chips** An HPCDA analogue method was applied to red/green miRNA chips of Miltenyi. Most of 41 significant miRNAs were already validated. **Database ImmuMon for Immune Monitoring** We have analyzed 37 individuals (AS, RA, SLE, and ND) with database ImmuMon. It is obvious, that immune monitoring leads to new parameters, which could be important either alone or in combination with others for diagnosis, differentiation, or responder detection. **Perspectives:** For utilization of tools, downstream of chip data analysis (e.g. Ingenuity, DEEP, or DAVID, but also systems biology), it becomes more and more relevant to obtain gene lists with high accuracy. There is no method for quantifying the quality of gene lists (GLs) available today. We are trying to change this with the new Gene List Significance Index (GLSI). It is a relative value and makes GL rankings independent of normalization methods. A randomly selected GL w/o significant genes achieves values near 1.0; below this are GLs of normalization IDs or control genes, not at all significant. With increasing fractions of true positive genes in the list, GLSI also increases. Only with a quantifier for GL quality you can objectively rank a list of extracted significant genes in a decreasing order of significance. We can show with GLSI, that neither FC nor t-test, %Change calls nor other data are sufficient to rank genes in an optimal way. Our new empirical HPCDA-Score achieves much better rankings.

References: **1.** Menßen A, Edinger G, Grün JR, Haase U, Baumgrass R, Grützkau A, Radbruch A, Burmester GR, Häupl T. SiPaGene: A new repository for instant online retrieval, sharing and meta-analyses of GeneChip expression data. BMC Genomics 2009 Mar 5;10:98 **2.** Röck J, Schneider E, Grün JR, Grützkau A, Küppers R, Schmitz J, Winkels G. CD303 (BDCA-2) signals in plasmacytoid dendritic cells via a BCR-like signalosome involving Syk, Slp65 and PLCc2. Eur J Immunol 2007; 37: 3564–3575 **3.** Biesen R, Demir C, Barkhudarova F, Grün JR, Steinbrich-Zöllner M, Backhaus M, Häupl T, Rudwaleit M, Riemekasten G, Radbruch A, Hiepe F, Burmester GR, Grützkau A. Sialic acid-binding Ig-like lectin 1 expression in inflammatory and resident monocytes is a potential biomarker for monitoring disease activity and success of therapy in systemic lupus erythematosus. Arthritis Rheum 2008; 58/4: 1136-1145 **4.** Steinbrich-Zöllner M, Grün JR, Kaiser T, Biesen R, Raba K, Wu P, Thiel A, Rudwaleit M, Sieper J, Burmester GR, Radbruch A, Grützkau A. From transcriptome to cytome: Integrating cytometric profiling, multivariate cluster and prediction analyses for a phenotypical classification of inflammatory diseases. Cytometry, Part A 2008; 73A: 333–340.

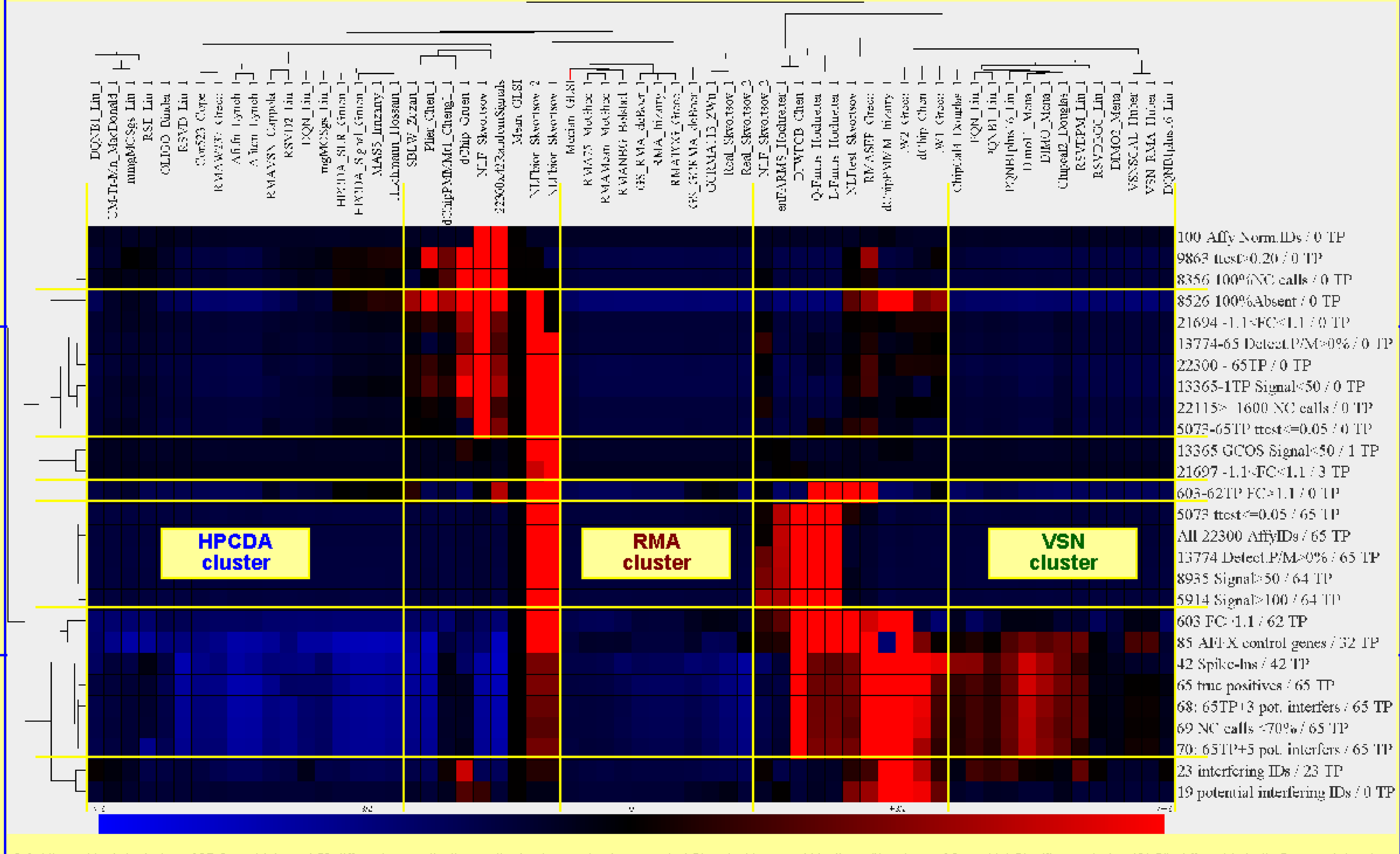
Funding: AutoCure (Curing autoimmune diseases); LSHB-CT-20 13

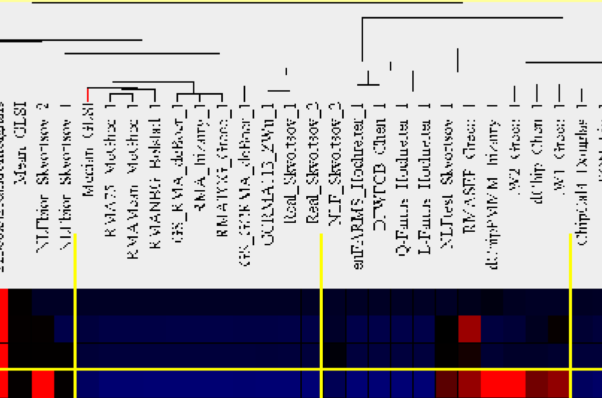
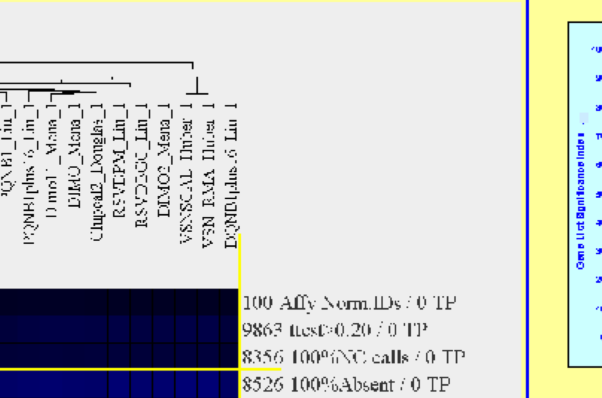
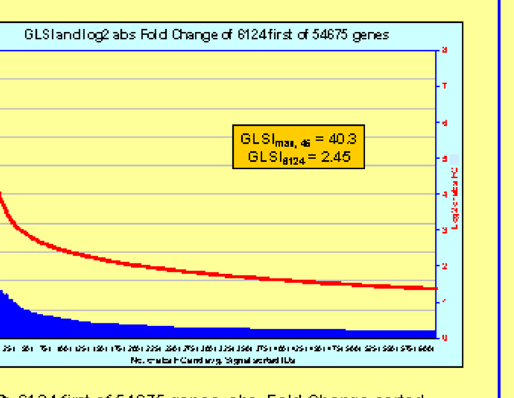
2.2: 6124 first of 54675 genes, abs. Change call sorted



3.1: Selected best 1543 genes of Fig. 1.1 – 2.8

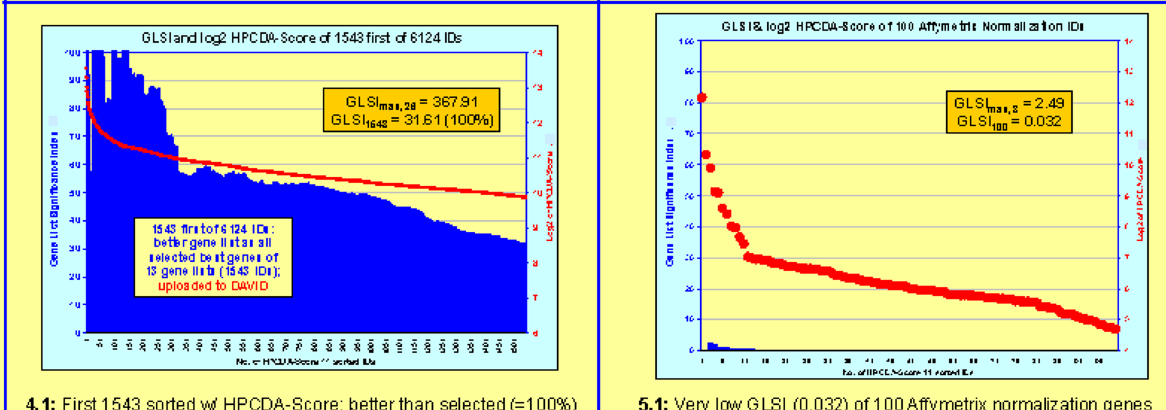
3.2: Only 6 of 1543 were not in 6124 significant GL



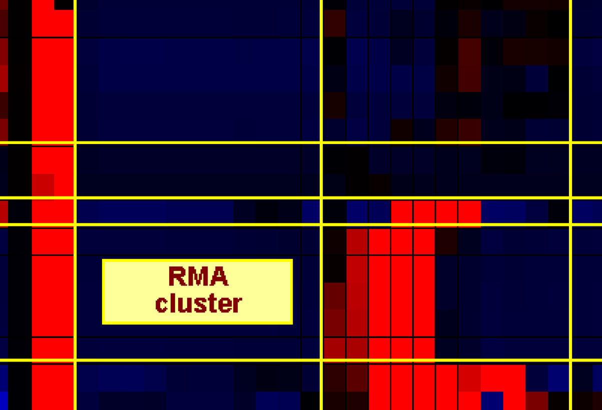
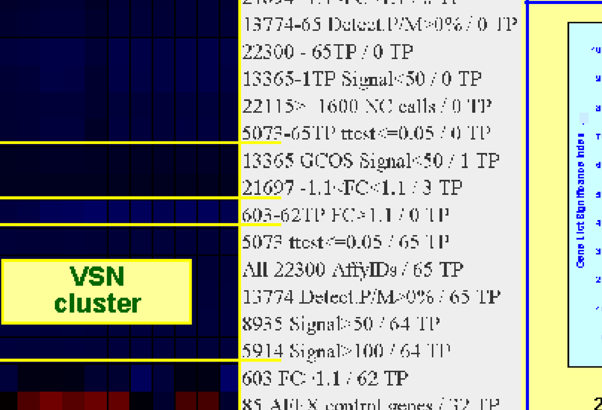
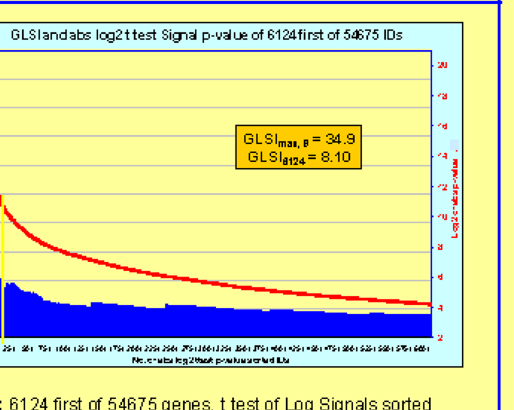
6.1: Hierarchical clustering of 27 Gene Lists and 59 different normalization methods plus randomly generated Signals, Mean and Median with values of Gene List Significance Index (GLSI). Affymetrix Latin Square dataset with 65 True Positive genes or transcripts. Normalized Signals of all most important 56 methods were downloaded from AffyComp Website, those for additional three methods were created with original Affymetrix GCOS Signals. Methods clustered as expected: nearly all RMA methods together with the median values. HPCDA for Signals clustered together with the MAS5 method of Irrizarry (both used nearly identical normalization methods) and HPCDA with Signal Log Ratios (SLRs) is only marginal distinct from those both. A third method cluster includes VSN, this cluster is on the opposite method main branch than HPCDA and RMA cluster. The most interesting finding here is that GLSI of all 59 methods clustered all 27 gene lists according to their fraction of true positive genes within 7 cluster. Clearly separated for all normalization methods were Gene Lists with and those without significant genes. All GLs were created with our HPCDA method; nevertheless all normalization methods came to similar findings while quantifying the quality of these gene lists by GLSI. One exception was the single GL cluster of 603 IDs w/o 62 TPs, found with HPCDA with a FC>=1.1. It seems that the majority of normalization methods puts too much weight on FCs, so this GL without any TP clusters on the wrong side.

2.3: 6124 first of 54675 genes, abs. Fold Change sorted

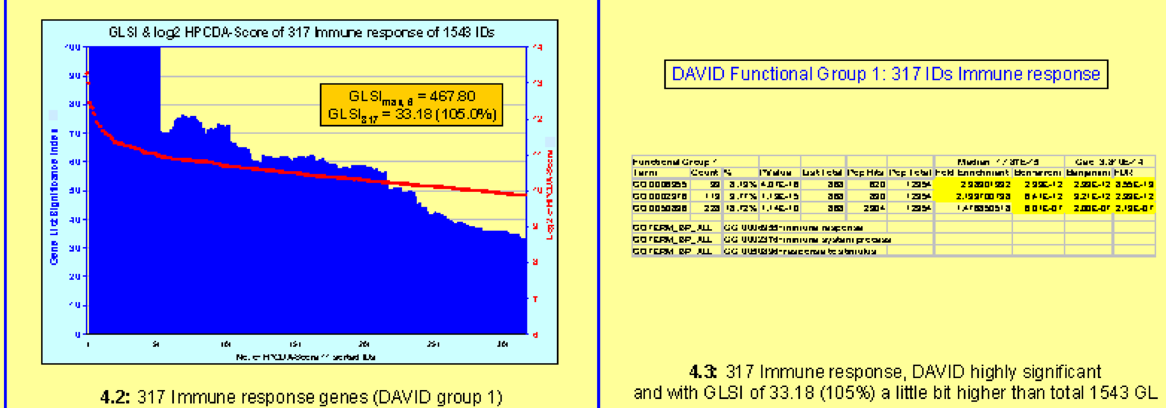


4.1: First 1543 sorted w/ HPCDA-Score: better than selected (=100%)

5.1: Very low GLSI (0.032) of 100 Affymetrix normalization genes

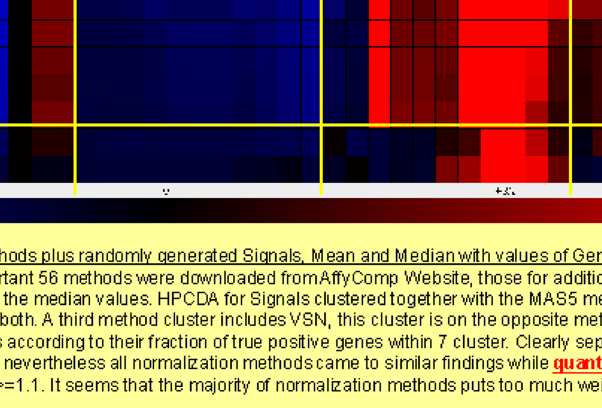
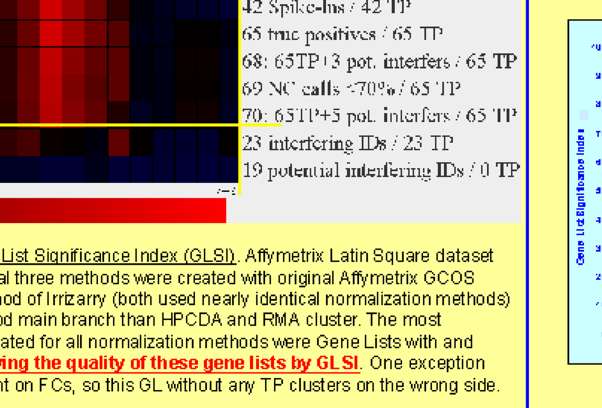
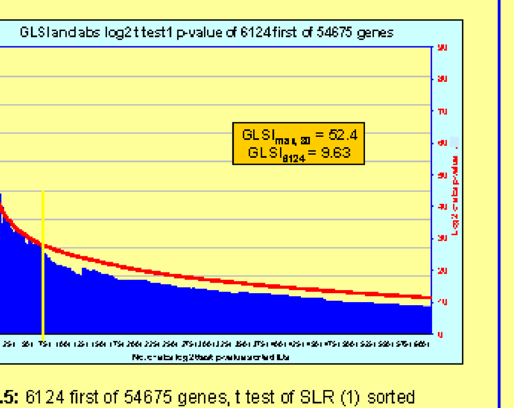




2.4: 6124 first of 54675 genes, t test of Log Signals sorted

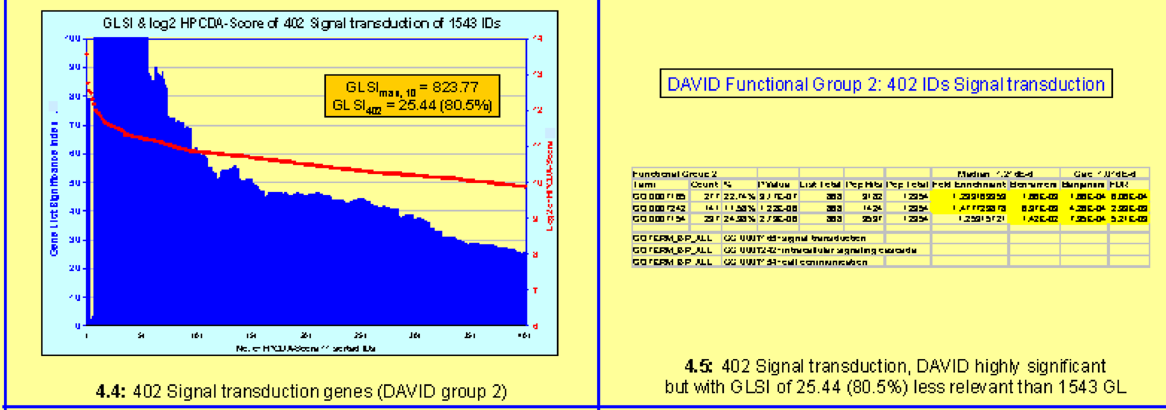


4.2: 317 Immune response genes (DAVID group 1)

4.3: 317 Immune response, DAVID highly significant and with GLSI of 33.18 (105%) a little bit higher than total 1543 GL

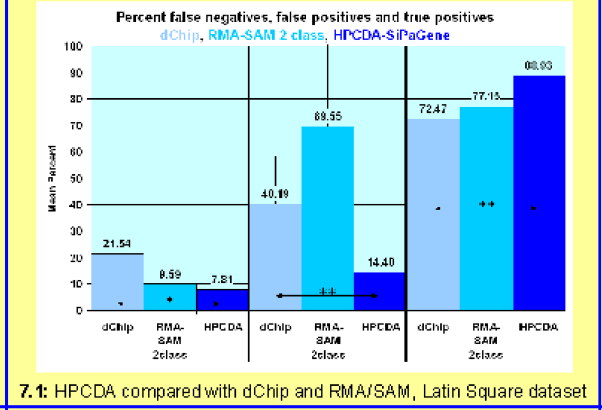




2.5: 6124 first of 54675 genes, t test of SLR (1) sorted

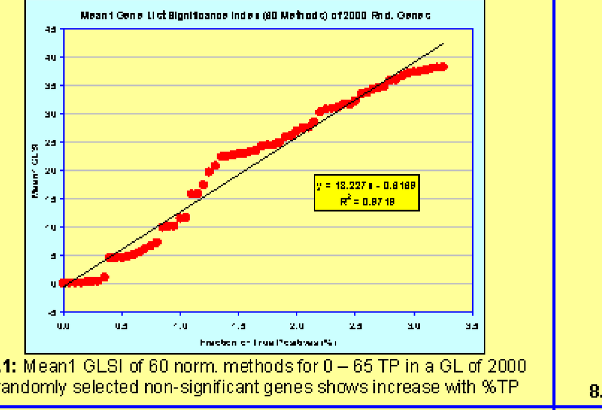


4.4: 402 Signal transduction genes (DAVID group 2)

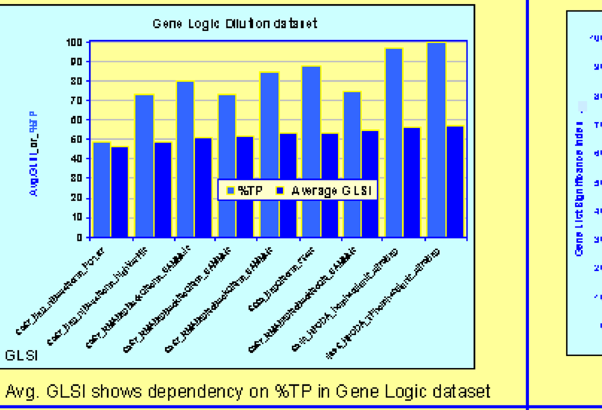
4.5: 402 Signal transduction, DAVID highly significant but with GLSI of 25.44 (80.5%) less relevant than 1543 GL



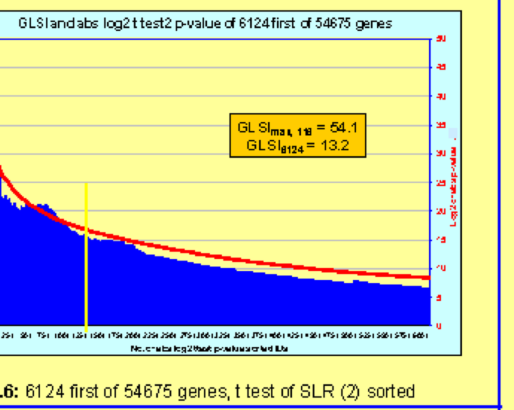
7.1: HPCDA compared with dChip and RMA/SAM, Latin Square dataset



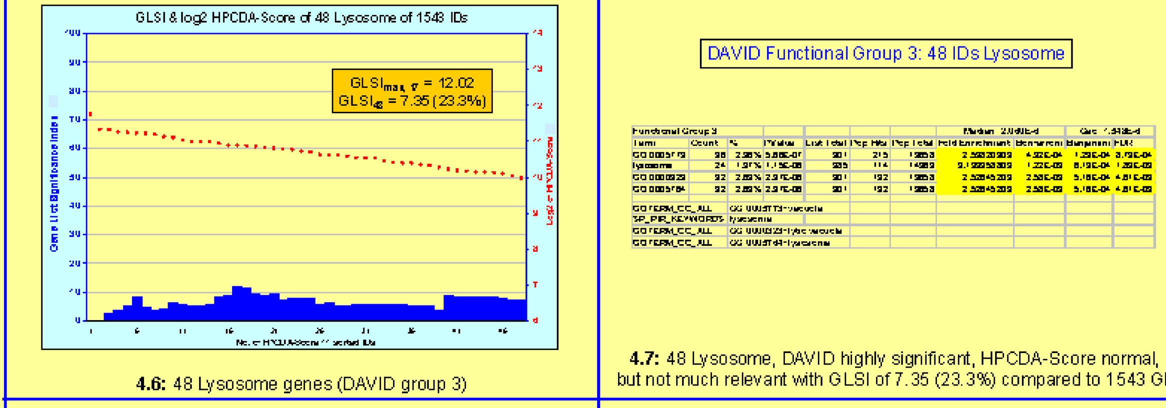
8.1: Mean1 GLSI of 60 norm. methods for 0 – 65 TP in a GL of 2000 randomly selected non-significant genes shows increase with %TP



8.2: Avg. GLSI shows dependency on %TP in Gene Logic dataset

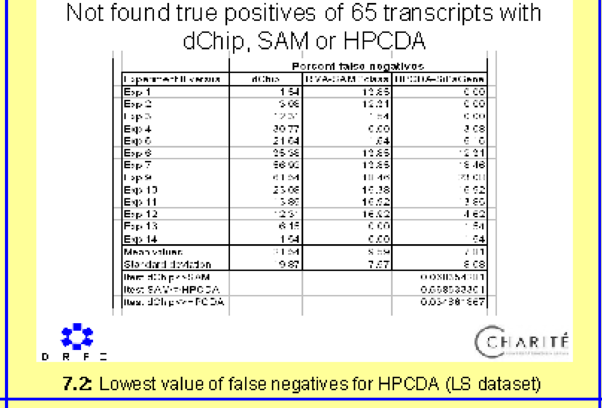


2.6: 6124 first of 54675 genes, t test of SLR (2) sorted

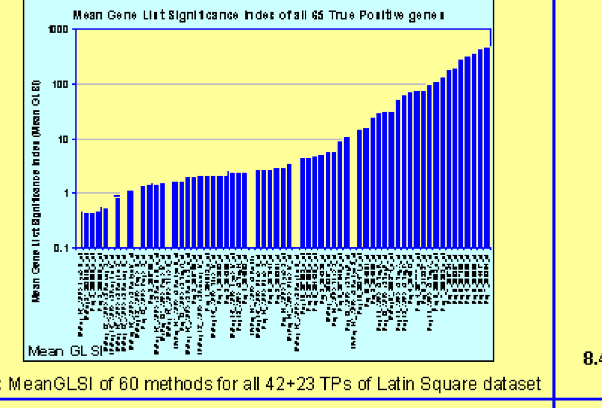


4.6: 48 Lysosome genes (DAVID group 3)

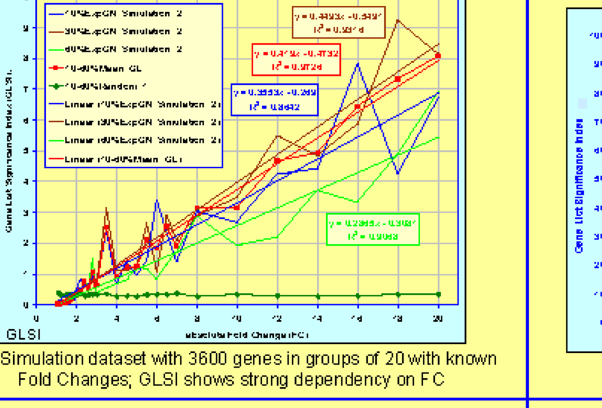
4.7: 48 Lysosome, DAVID highly significant, HPCDA-Score normal, but not much relevant with GLSI of 7.35 (23.3%) compared to 1543 GL



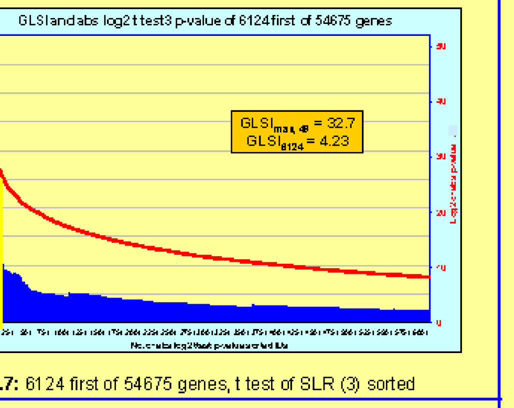
7.2: Lowest value of false negatives for HPCDA (LS dataset)



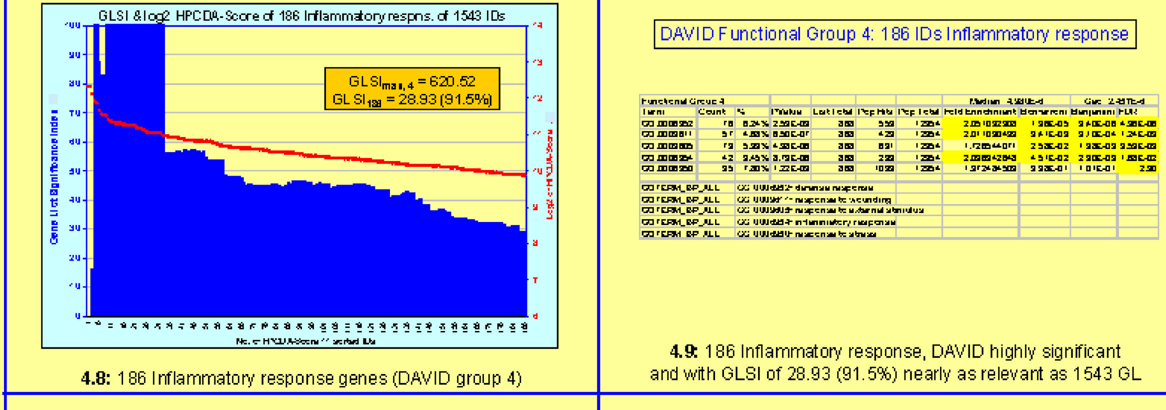
8.3: Mean GLSI of 60 methods for all 42+23 TPs of Latin Square dataset



8.4: Simulation dataset with 3600 genes in groups of 20 with known Fold Changes; GLSI shows strong dependency on FC

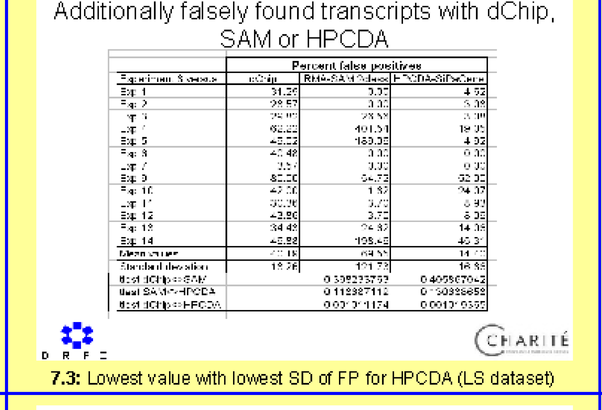


2.7: 6124 first of 54675 genes, t test of SLR (3) sorted

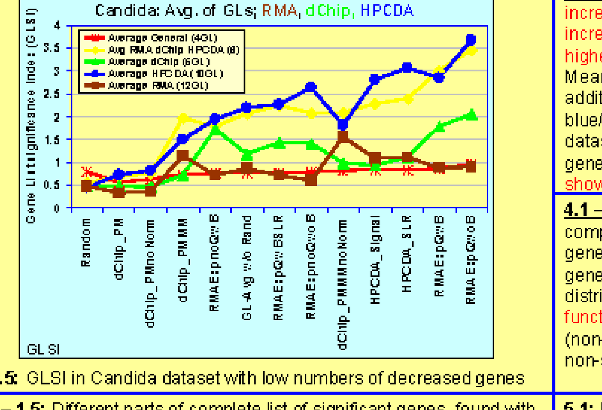


4.8: 186 Inflammatory response genes (DAVID group 4)

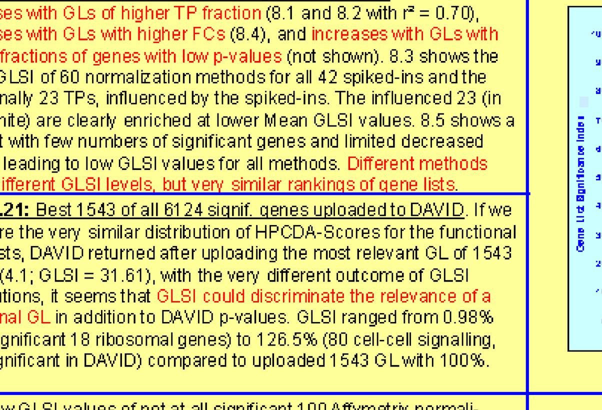
4.9: 186 Inflammatory response, DAVID highly significant and with GLSI of 28.93 (91.5%) nearly as relevant as 1543 GL



7.3: Lowest value with lowest SD of FP for HPCDA (LS dataset)

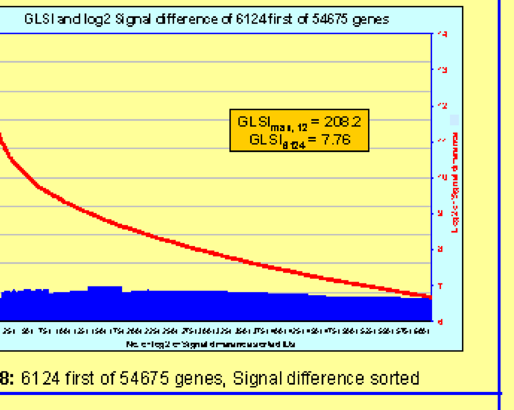


8.5: GLSI in Candida dataset with low numbers of decreased genes

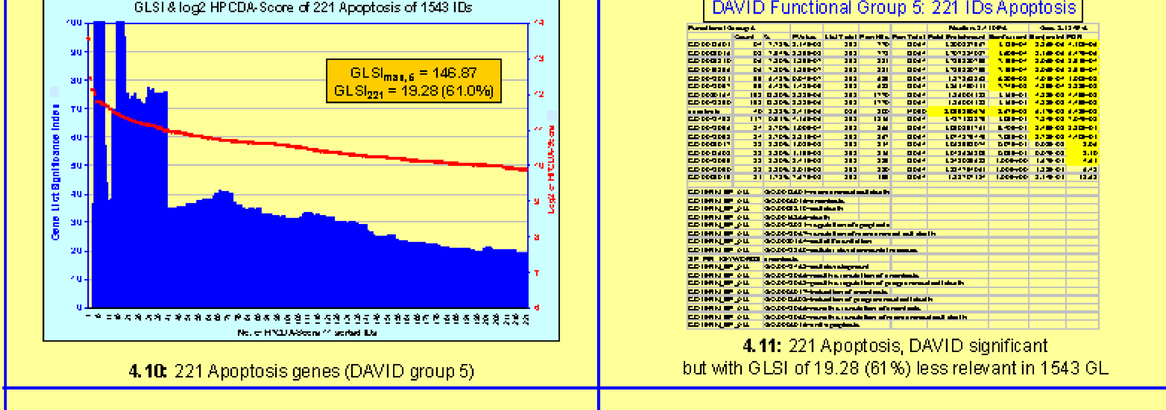


8.1 – 8.5: Dependencies of Gene List Significance Index. GLSI increases with GLs of higher TP fraction (8.1 and 8.2 with r^2 = 0.70), increases with GLs with higher FCs (8.4), and increases with GLs with higher fractions of genes with low p-values (not shown). 8.3 shows the Mean GLSI of 60 normalization methods for all 42 spiked-ins and the additionally 23 TPs, influenced by the spiked-ins. The influenced 23 (in blue/white) are clearly enriched at lower Mean GLSI values. 8.5 shows a dataset with few numbers of significant genes and limited decreased genes, leading to low GLSI values for all methods. Different methods show different GLSI levels, but very similar rankings of gene lists.

4.1 – 4.21: Best 1543 of all 6124 signif. genes uploaded to DAVID. If we compare the very similar distribution of HPCDA-Scores for the functional gene lists, DAVID returned after uploading the most relevant GL of 1543 genes (4.1; GLSI = 31.61), with the very different outcome of GLSI distributions, it seems that GLSI could discriminate the relevance of a functional GL in addition to DAVID p-values. GLSI ranged from 0.98% (non-significant 18 ribosomal genes) to 126.5% (80 cell-cell signalling, non-significant in DAVID) compared to uploaded 1543 GL with 100%.

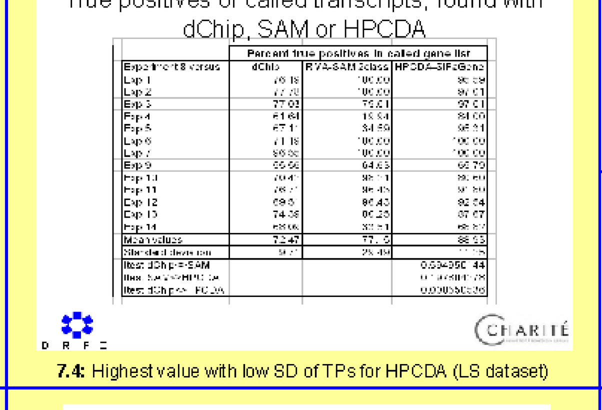


2.8: 6124 first of 54675 genes, Signal difference sorted

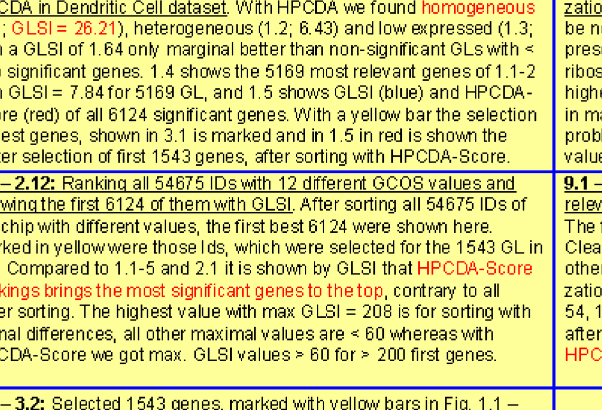


4.10: 221 Apoptosis genes (DAVID group 5)

4.11: 221 Apoptosis, DAVID significant but with GLSI of 19.28 (61%) less relevant in 1543 GL

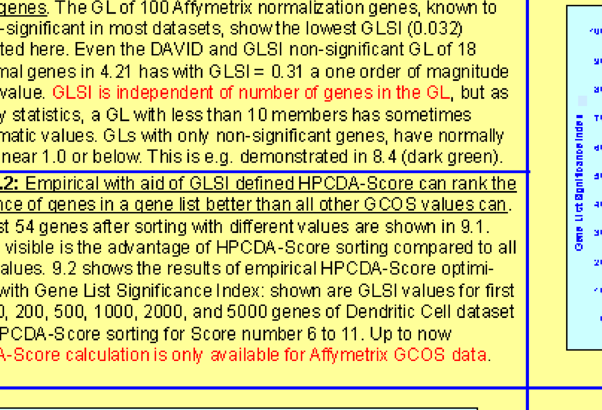


7.4: Highest value with low SD of TPs for HPCDA (LS dataset)



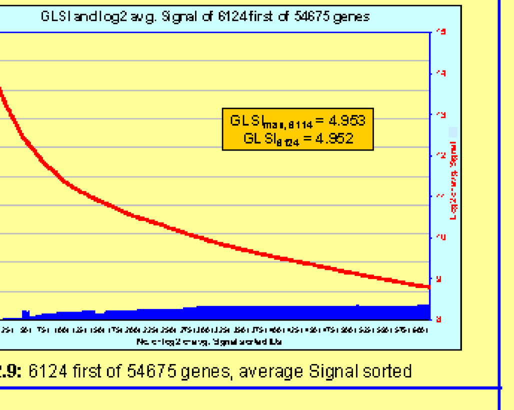
1.1 – 1.5: Different parts of complete list of significant genes, found with HPCDA in Dendritic Cell dataset. With HPCDA we found homogeneous (1.1; GLSI = 26.21), heterogeneous (1.2; 6.43) and low expressed (1.3; with a GLSI of 1.64 only marginal better than non-significant GLs with < 1.0) significant genes. 1.4 shows the 5169 most relevant genes of 1.1-2 with GLSI = 7.84 for 5169 GL, and 1.5 shows GLSI (blue) and HPCDA-Score (red) of all 6124 significant genes. With a yellow bar the selection of best genes, shown in 3.1 is marked and in 1.5 in red is shown the better selection of first 1543 genes, after sorting with HPCDA-Score.

2.1 – 2.12: Ranking all 54675 IDs with 12 different GCOS values and showing the first 6124 of them with GLSI. After sorting all 54675 IDs of the chip with different values, the first best 6124 were shown here. Marked in yellow were those Ids, which were selected for the 1543 GL in 3.1. Compared to 1.1-5 and 2.1 it is shown by GLSI that HPCDA-Score rankings brings the most significant genes to the top, contrary to all other sorting. The highest value with max GLSI = 208 is for sorting with Signal differences, all other maximal values are < 60 whereas with HPCDA-Score we got max. GLSI values > 60 for > 200 first genes.

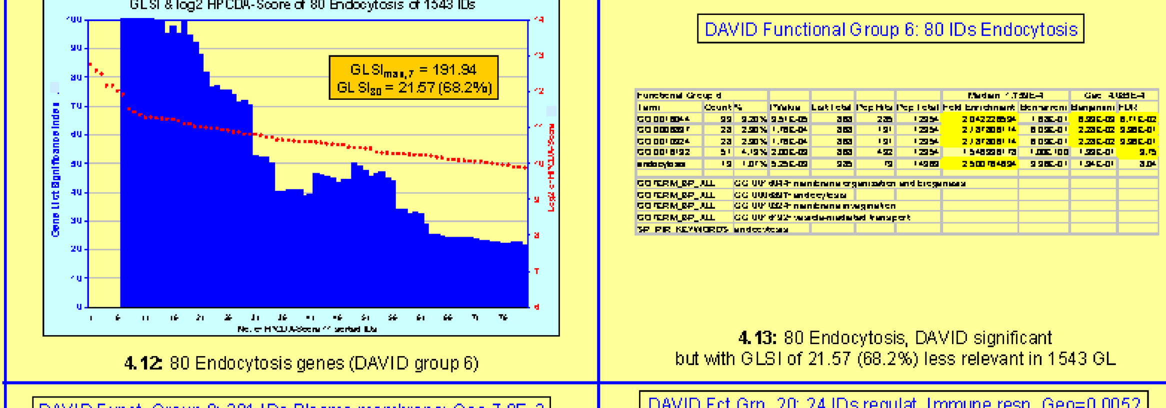


5.1: Low GLSI values of not at all significant 100 Affymetrix normalization genes. The GL of 100 Affymetrix normalization genes, known to be non-significant in most datasets, show the lowest GLSI (0.032) presented here. Even the DAVID and GLSI non-significant GL of 18 ribosomal genes in 4.21 has with GLSI = 0.31 a one order of magnitude higher value. GLSI is independent of number of genes in the GL, but as in many statistics, a GL with less than 10 members has sometimes problematic values. GLs with only non-significant genes, have normally values near 1.0 or below. This is e.g. demonstrated in 8.4 (dark green).

9.1 – 9.2: Empirical with aid of GLSI defined HPCDA-Score can rank the relevance of genes in a gene list better than all other GCOS values can. The first 54 genes after sorting with different values are shown in 9.1. Clearly visible is the advantage of HPCDA-Score sorting compared to all other values. 9.2 shows the results of empirical HPCDA-Score optimization with Gene List Significance Index: shown are GLSI values for first 54, 100, 200, 500, 1000, 2000, and 5000 genes of Dendritic Cell dataset after HPCDA-Score sorting for Score number 6 to 11. Up to now HPCDA-Score calculation is only available for Affymetrix GCOS data.

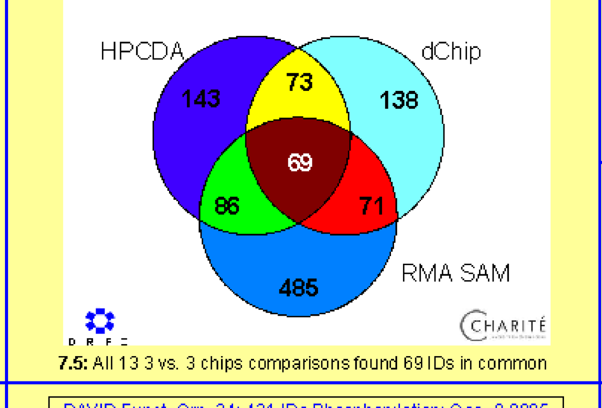


2.9: 6124 first of 54675 genes, average Signal sorted

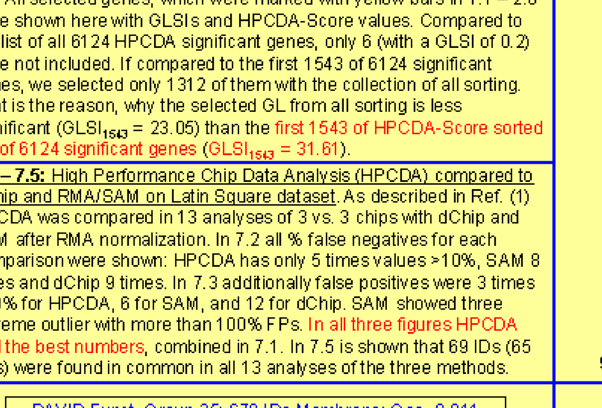


4.12: 80 Endocytosis genes (DAVID group 6)

4.13: 80 Endocytosis, DAVID significant but with GLSI of 21.57 (68.2%) less relevant in 1543 GL

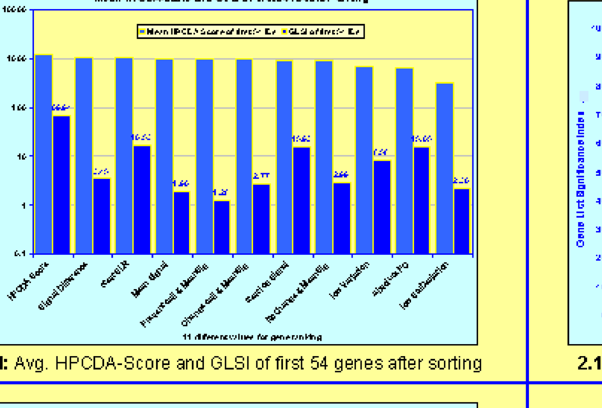


7.5: All 13 3 vs. 3 chips comparisons found 69 IDs in common

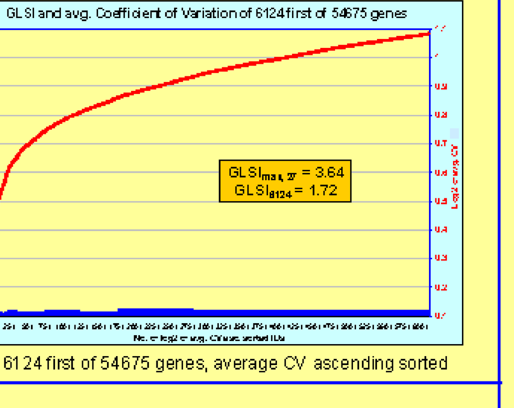


3.1 – 3.2: Selected 1543 genes, marked with yellow bars in Fig. 1.1 – 2.8. All selected genes, which were marked with yellow bars in 1.1 – 2.8 were shown here with GLSIs and HPCDA-Score values. Compared to the list of all 6124 HPCDA significant genes, only 6 (with a GLSI of 0.2) were not included. If compared to the first 1543 of 6124 significant genes, we selected only 1312 of them with the collection of all sorting. That is the reason, why the selected GL from all sorting is less significant ($GLSI_{1543}$ = 23.05) than the first 1543 of HPCDA-Score sorted GL of 6124 significant genes ($GLSI_{1543}$ = 31.61).

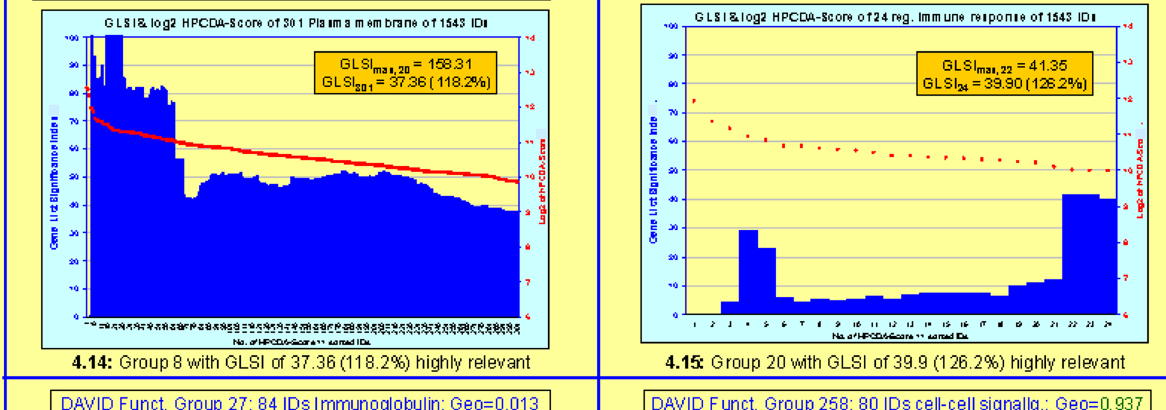
7.1 – 7.5: High Performance Chip Data Analysis (HPCDA) compared to dChip and RMA/SAM on Latin Square dataset. As described in Ref. (1) HPCDA was compared in 13 analyses of 3 vs. 3 chips with dChip and SAM after RMA normalization. In 7.2 all % false negatives for each comparison were shown: HPCDA has only 5 times values >10%, SAM 8 times and dChip 9 times. In 7.3 additionally false positives were 3 times >20% for HPCDA, 6 for SAM, and 12 for dChip. SAM showed three extreme outlier with more than 100% FPs. In all three figures HPCDA had the best numbers, combined in 7.1. In 7.5 is shown that 69 IDs (65 TPs) were found in common in all 13 analyses of the three methods.



9.1: Avg. HPCDA-Score and GLSI of first 54 genes after sorting

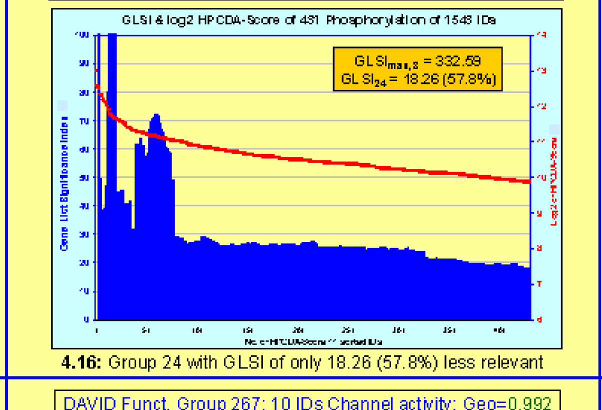


2.10: 6124 first of 54675 genes, average CV ascending sorted

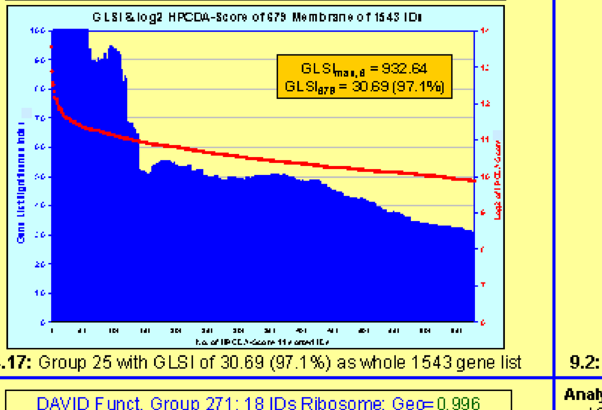


4.14: Group 8 with GLSI of 37.36 (118.2%) highly relevant

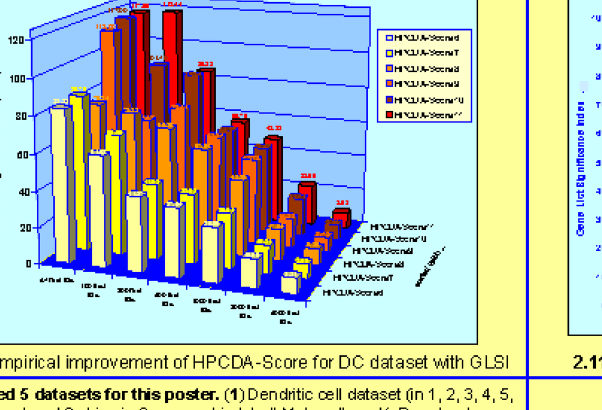
4.15: Group 20 with GLSI of 39.9 (126.2%) highly relevant



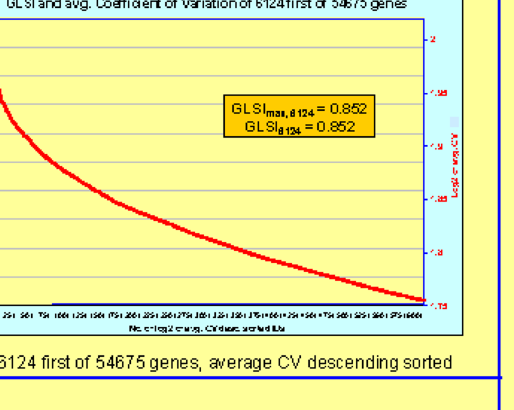
4.16: Group 24 with GLSI of only 18.26 (57.8%) less relevant



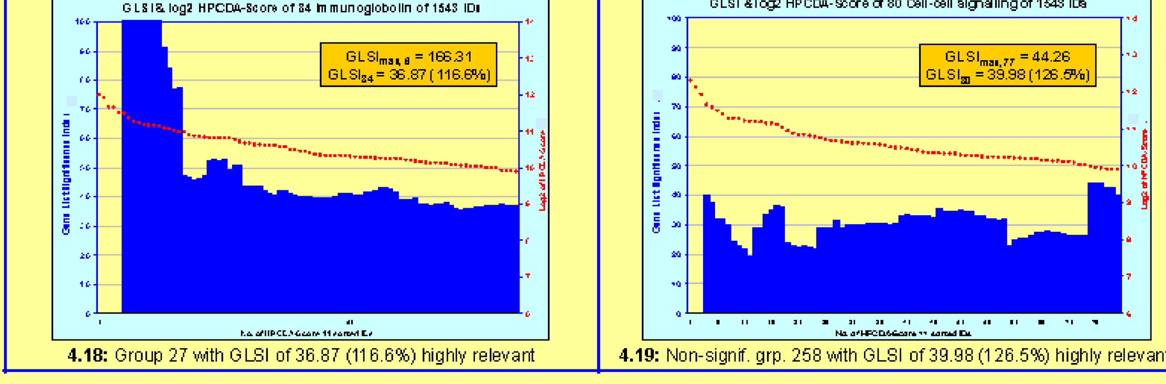
4.17: Group 25 with GLSI of 30.69 (97.1%) as whole 1543 gene list



9.2: Empirical improvement of HPCDA-Score for DC dataset with GLSI

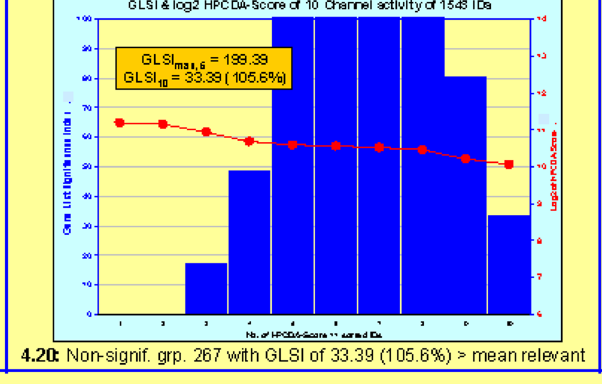


2.11: 6124 first of 54675 genes, average CV descending sorted

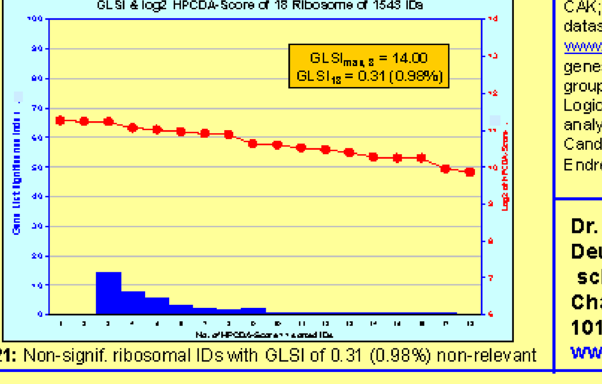


4.18: Group 27 with GLSI of 36.87 (116.6%) highly relevant

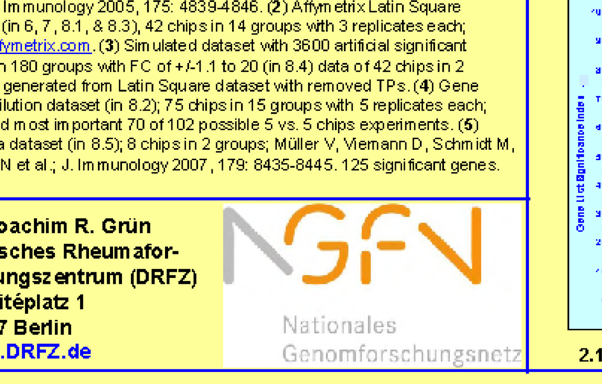
4.19: Non-signif. grp. 258 with GLSI of 39.98 (126.5%) highly relevant



4.20: Non-signif. grp. 267 with GLSI of 33.39 (105.6%) > mean relevant



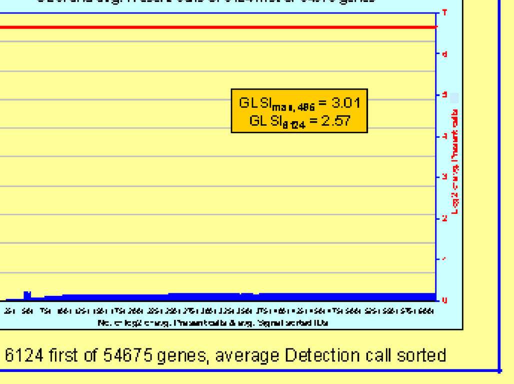
4.21: Non-signif. ribosomal IDs with GLSI of 0.31 (0.98%) non-relevant



Analyzed 5 datasets for this poster. (**1**) Dendritic cell dataset (in 1, 2, 3, 4, 5, and 9) analyzed 6 chips in 2 groups; Lindstedt M, Lundberg K, Borrebaek CAK; J. Immunology 2005, 175: 4839-4846. (**2**) Affymetrix Latin Square dataset (in 6, 7, 8.1, & 8.3), 42 chips in 14 groups with 3 replicates each; www.affymetrix.com. (**3**) Simulated dataset with 3600 artificial significant genes in 180 groups with FC of +/-1.1 to 20 (in 8.4) data of 42 chips in 2 groups; generated from Latin Square dataset with removed TPs. (**4**) Gene Logic Dilution dataset (in 8.2); 75 chips in 15 groups with 5 replicates each; analyzed most important 70 of 102 possible 5 vs. 5 chips experiments. (**5**) Candida dataset (in 8.5); 8 chips in 2 groups; Müller V, Viemann D, Schmidt M, Endres N et al.; J. Immunology 2007, 179: 8435-8445. 125 significant genes.

Dr. Joachim R. Grün
Deutsches Rheumaforschungszentrum (DRFZ)
Charitéplatz 1
10117 Berlin
www.DRFZ.de

Nationales Genomforschungsnetz



2.12: 6124 first of 54675 genes, average Detection call sorted