A model high level analysis, and correlation analysis

Harshita Gaur
Hossameldin Ali

A model high level analysis, and correlation analysis

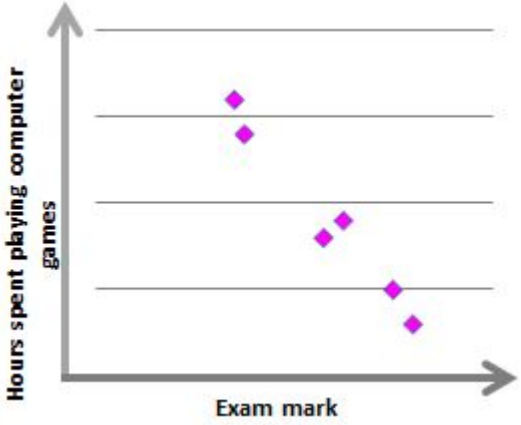
- What is model analysis?
- What is correlation analysis?
- What is a correlation network?

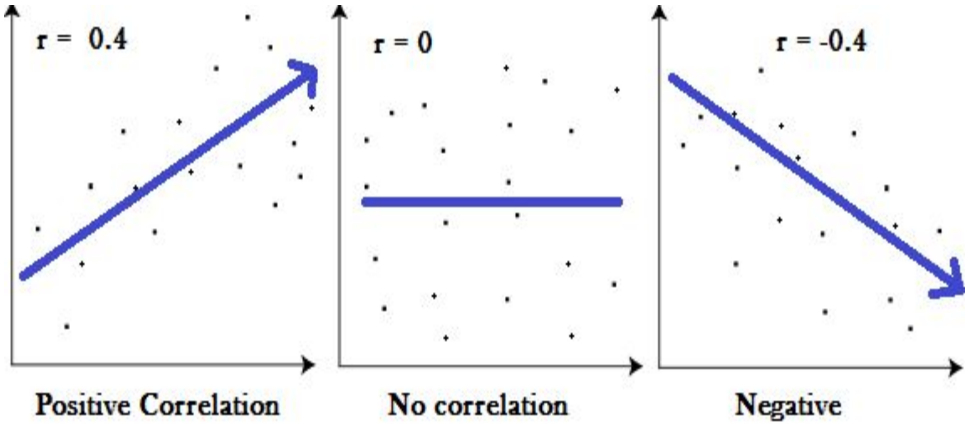
Model high level analysis

- Ex: Jaguar doors testing
- We are more into computer models

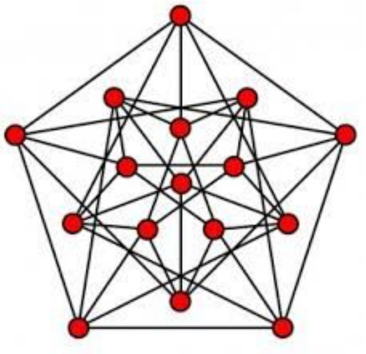
Correlation analysis

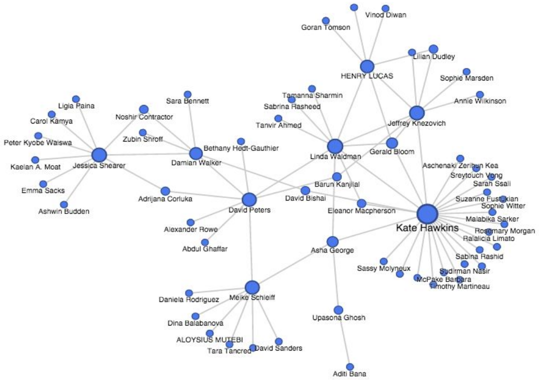
- Range -1 to 1
- Stronger correlation >0.5 or > -0.5





What is a network

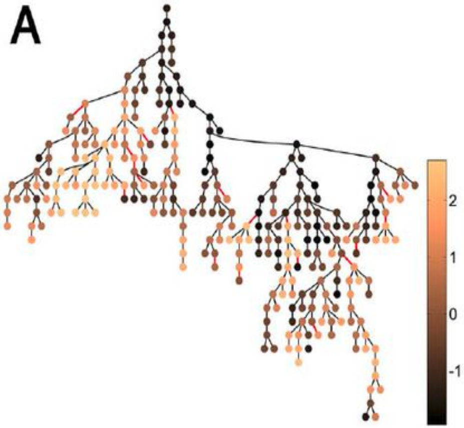


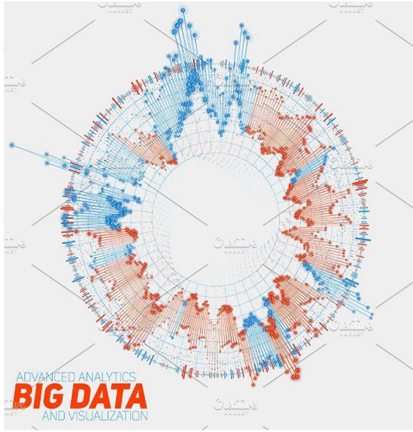


CONNECTIONS
3.200 AIRPORTS
martingrandjean.ch

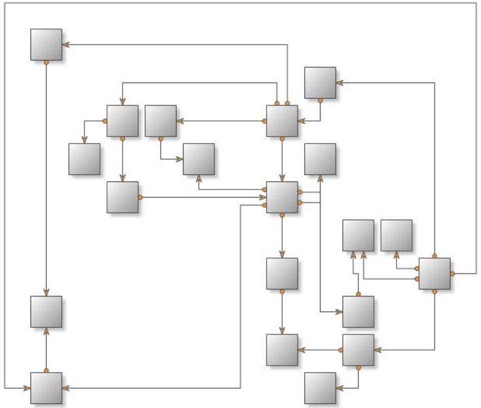
Graphs: mathematical structures used to model pairwise relations between objects

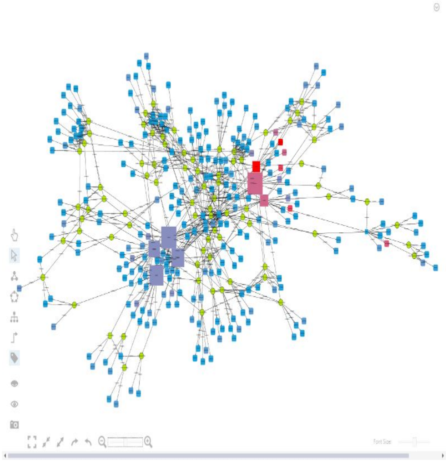
Biological network





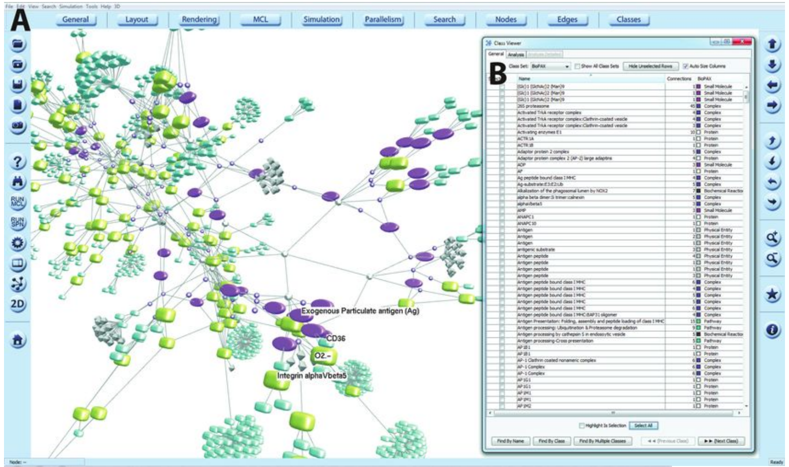
- Nodes are usually genes, transcripts or proteins.
- Nodes: shape, size and color
- Edges: directed or undirected
- Edges: can be weighted
- Aesthetically pleasing: number of edges crossing is minimized

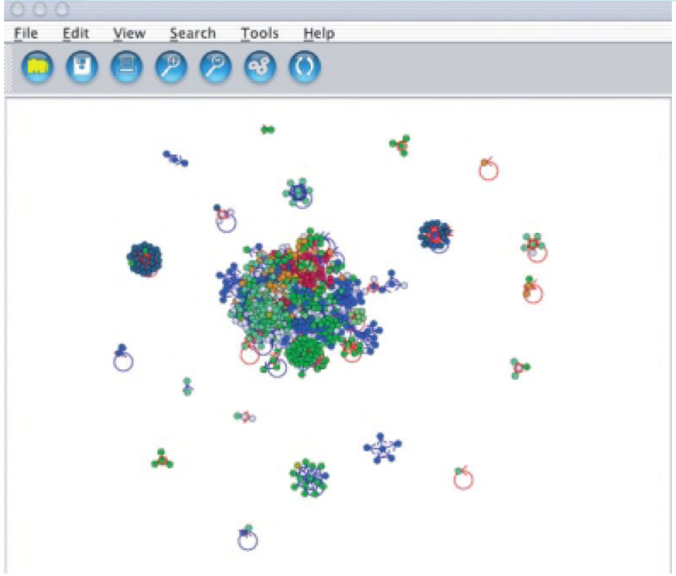




BioLayout Express 3D

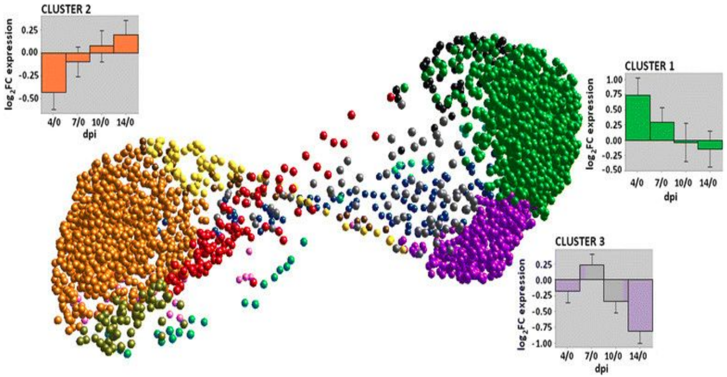
OpenGL®





Biolayout Express 3D

- 30,000 nodes and 2 – 3 million edges
- Appropriate data from any source, biological or otherwise
- Analysis gene expression data
- Optimized layout algorithm and graph clustering



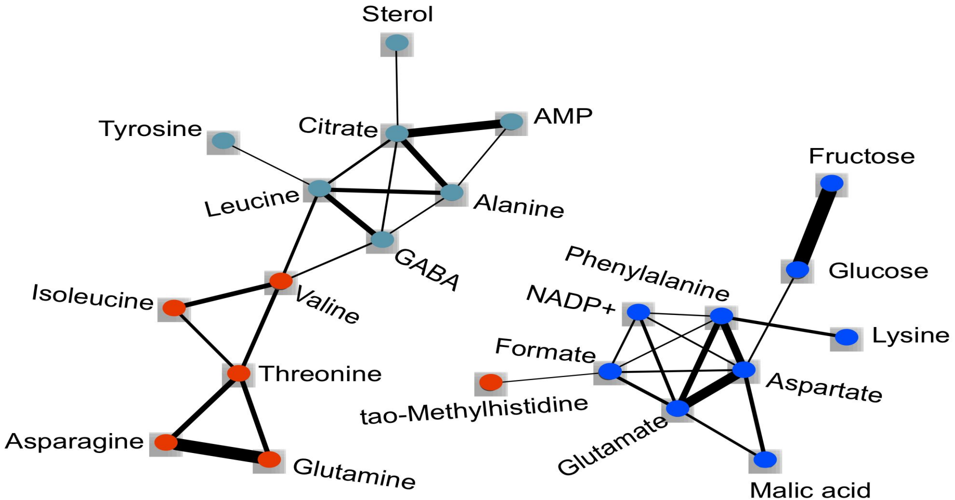
The graph paradigm for microarray expression data

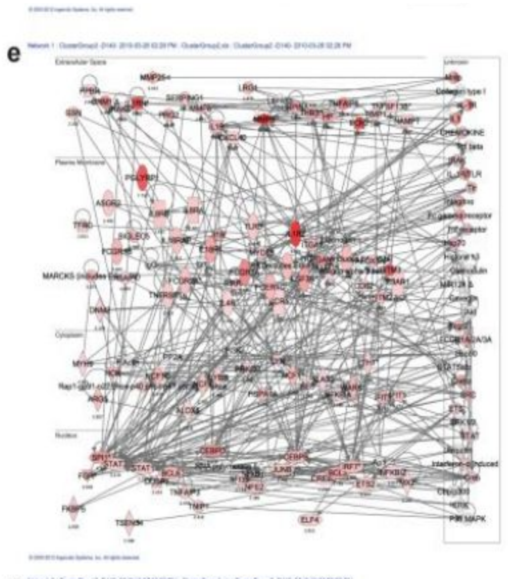
- Makes no earlier assumptions regarding the experimental design, normalization method, microarray platform or questions being addressed by the study

- However, all of these factors do influence the properties of a network, and therefore it is worth spending some time to discuss them.

Experimental design

- The size of a data set is a function of the number of probes on the array and the number of samples analyzed
    - number of samples is small (< 10): higher correlation between data derived from different probes
    - ~ 20 samples or more: no need to filter data





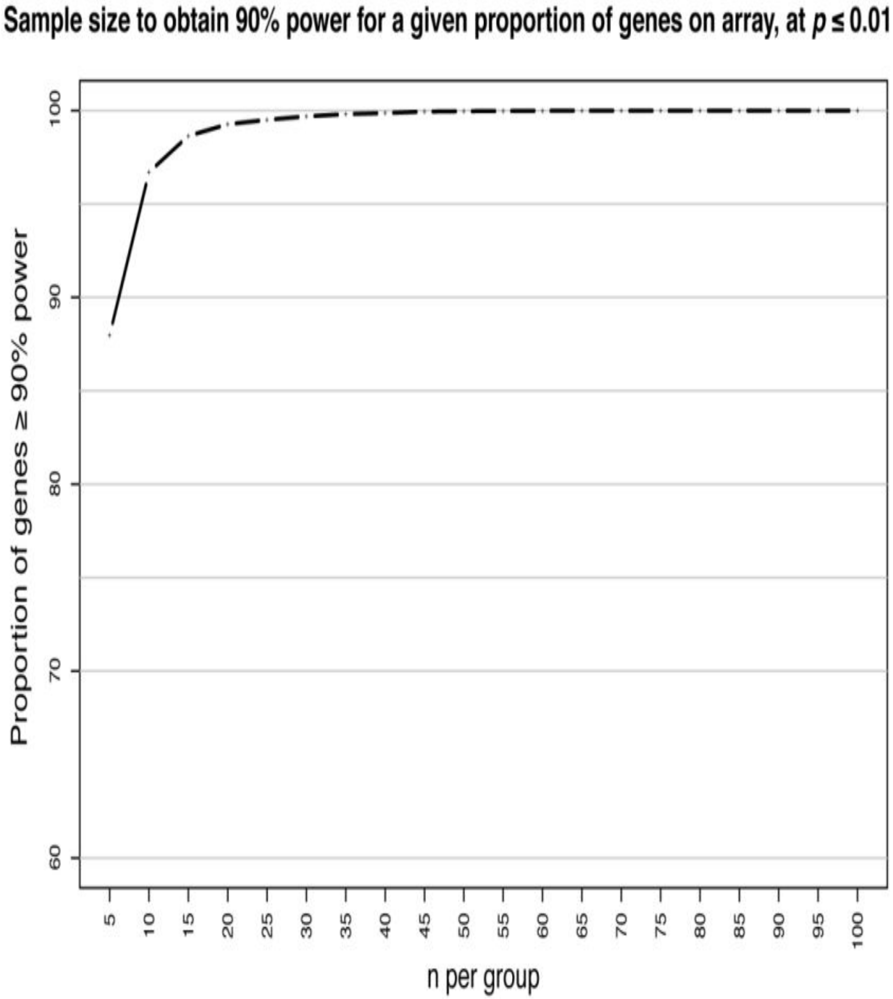
Correlation network (WGCNA)

- Weighted Gene Co-expression Network Analysis
- A method used to discover patterns among variables based on correlation
- Modules (Clusters)
- WGCNA is linear:
  - Faster computations
  - No overfitting problems
  - Suitable for small datasets

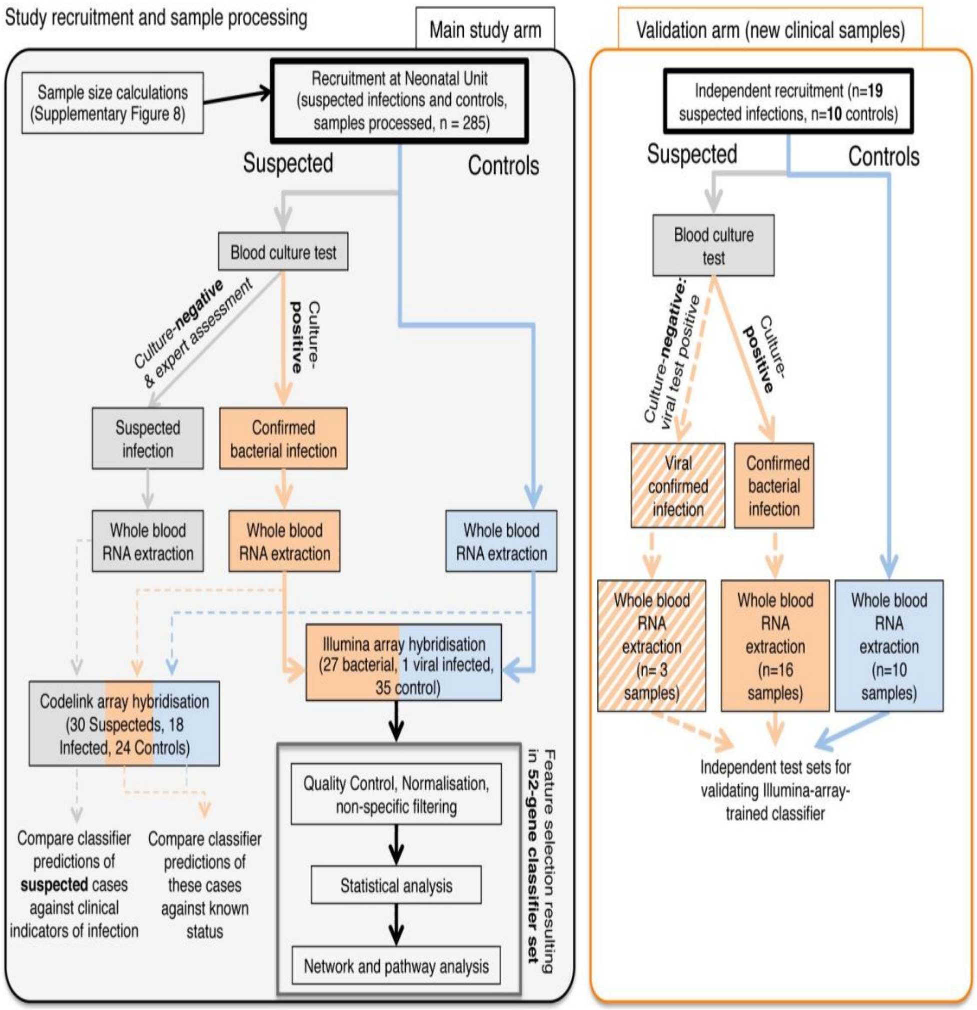
Identification of a human neonatal immune-metabolic network associated with bacterial infection

Power Calculations

- 30 control pre-vaccinated 9 month old Gambian infant samples was used to perform the power calculations
- A high-performance computing model for host RNA demonstrated that as few as 24 whole blood RNA markers would be sufficient for predicting infection.



what is the data



Study recruitment and sample processing

Main study arm

Sample size calculations (Supplementary Figure 8)

Recruitment at Neonatal Unit (suspected infections and controls, samples processed, n = 285)

Suspected

Controls

Blood culture test

*Culture-**negative** & expert assessment*

Culture-**positive**

Suspected infection

Confirmed bacterial infection

Whole blood RNA extraction

Whole blood RNA extraction

Whole blood RNA extraction

Codelink array hybridisation (30 Suspecteds, 18 Infected, 24 Controls)

Illumina array hybridisation (27 bacterial, 1 viral infected, 35 control)

Quality Control, Normalisation, non-specific filtering

Statistical analysis

Network and pathway analysis

Feature selection resulting in **52-gene classifier set**

Compare classifier predictions of **suspected** cases against clinical indicators of infection

Compare classifier predictions of these cases against known status

Validation arm (new clinical samples)

Independent recruitment (n=**19** suspected infections, n=**10** controls)

Suspected

Controls

Blood culture test

Culture-**negative**: viral test positive

Culture-**positive**

Viral confirmed infection

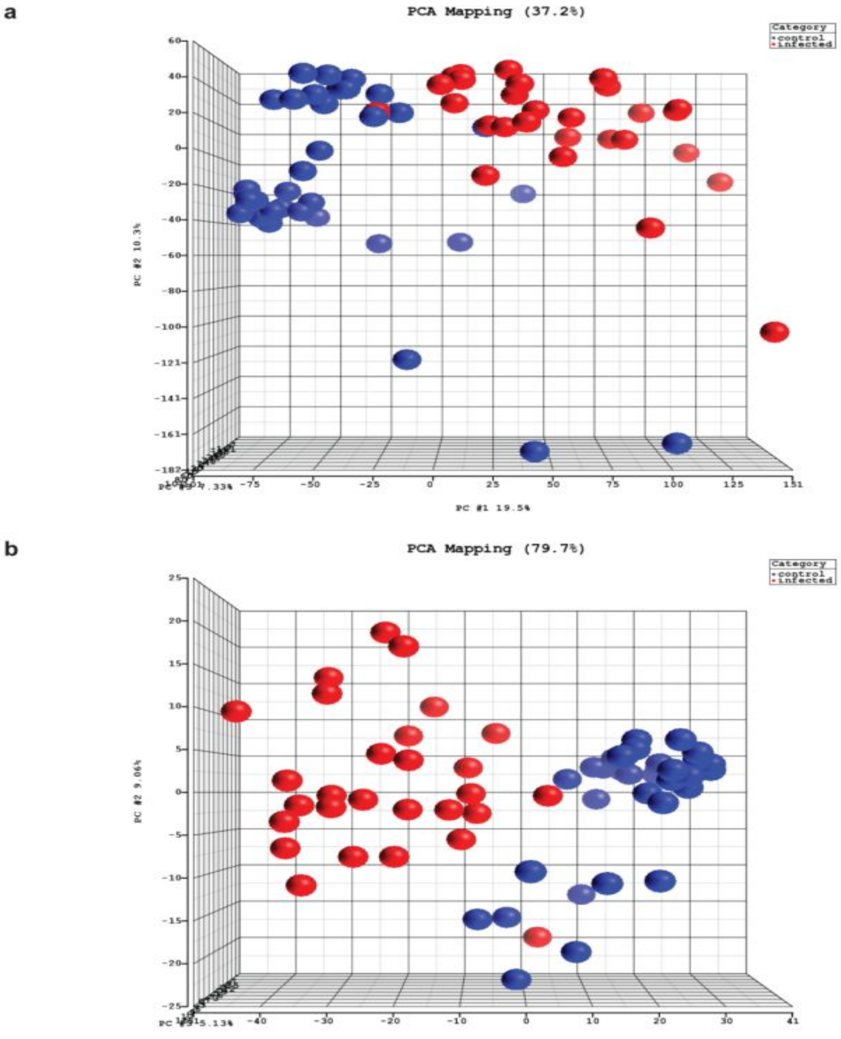
Confirmed bacterial infection

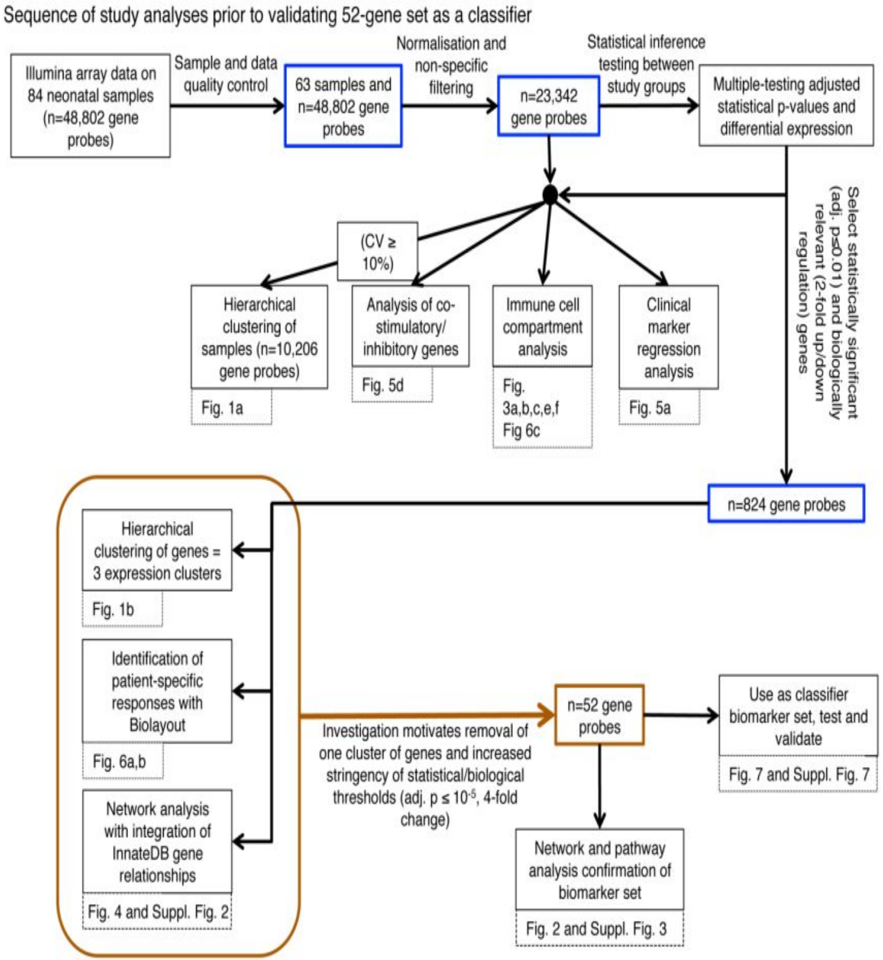
Whole blood RNA extraction (n= 3 samples)

Whole blood RNA extraction (n=16 samples)

Whole blood RNA extraction (n=10 samples)

Independent test sets for validating Illumina-array-trained classifier

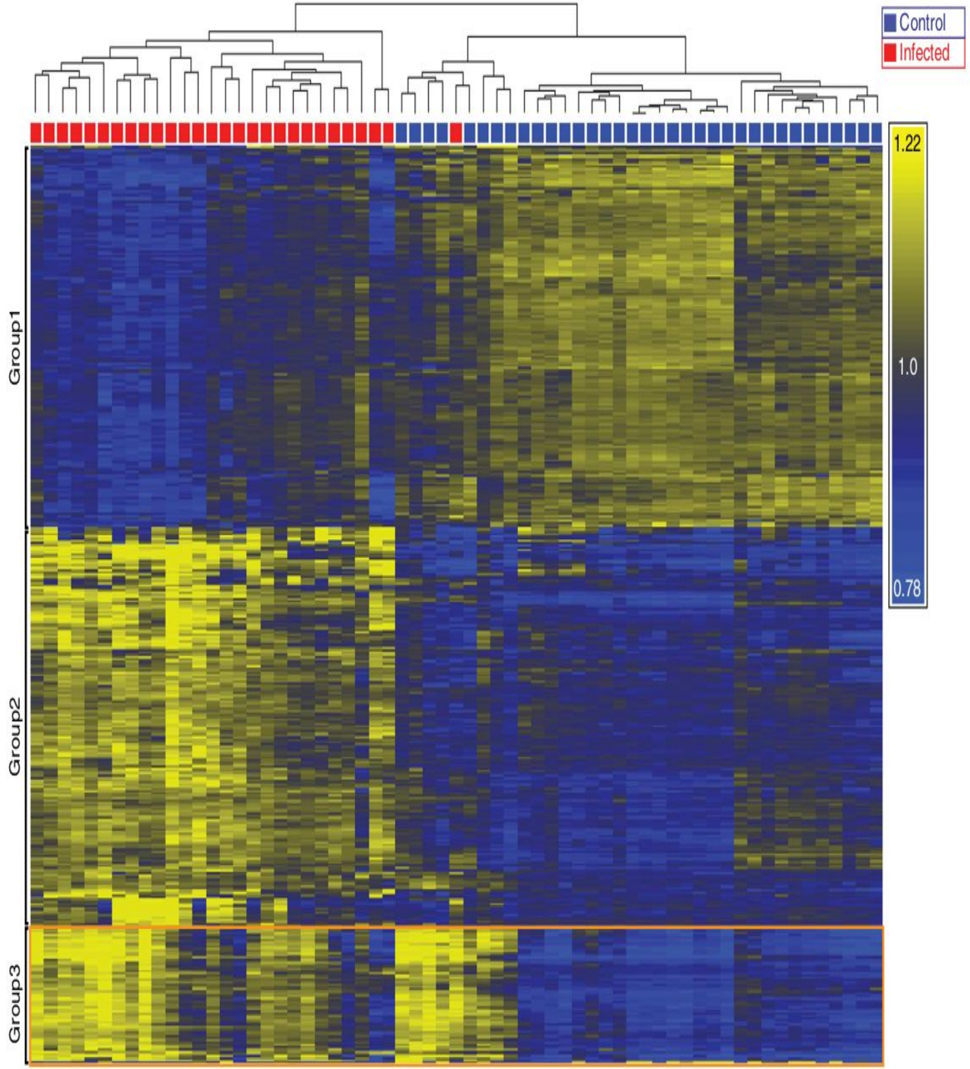

a
PCA Mapping (37.2%)
Category
control
infected
60
40
20
0
-20
-40
-60
-80
-100
-121
-141
-161
-182
PC #2 10.3%
-75
-50
-25
0
25
50
75
100
125
151
PC #1 19.5%
PC #3 7.33%
b
PCA Mapping (79.7%)
Category
control
infected
25
20
15
10
5
0
-5
-10
-15
-20
-25
PC #2 9.06%
-40
-30
-20
-10
0
10
20
30
41
PC #1 65.5%
PC #3 5.13%

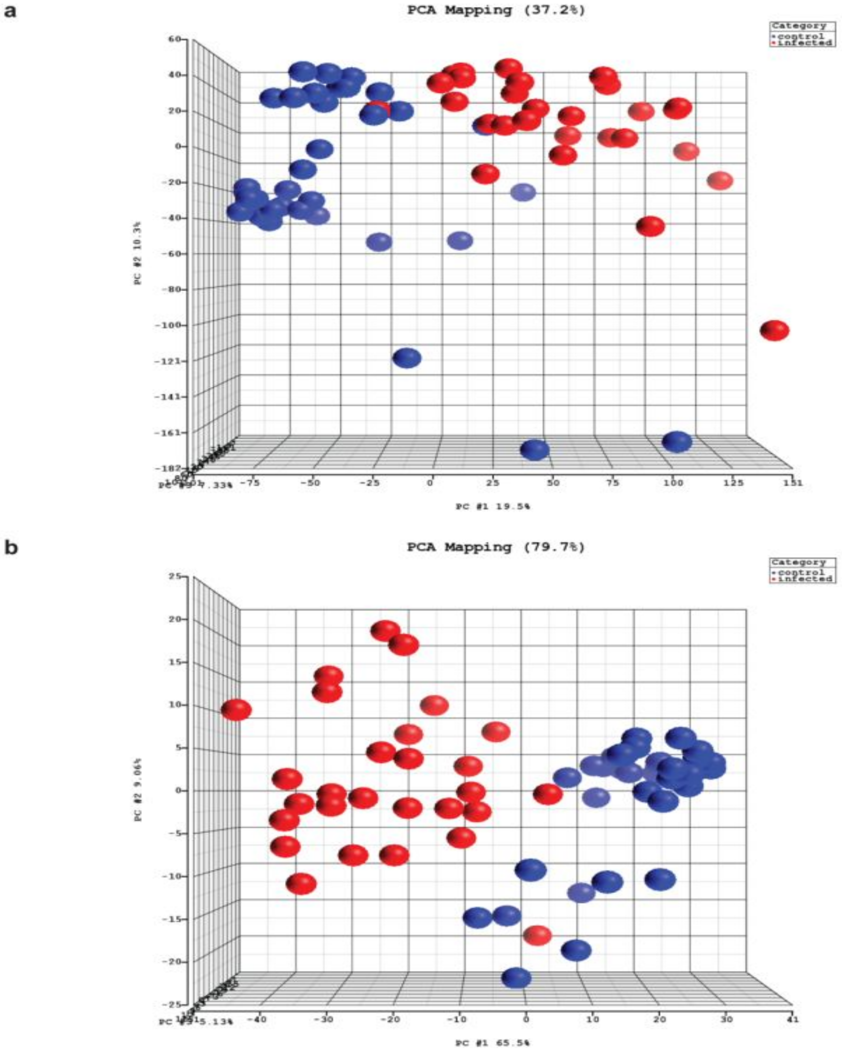

Sequence of study analyses prior to validating 52-gene set as a classifier
Illumina array data on 84 neonatal samples (n=48,802 gene probes)
Sample and data quality control
63 samples and n=48,802 gene probes
Normalisation and non-specific filtering
n=23,342 gene probes
Statistical inference testing between study groups
Multiple-testing adjusted statistical p-values and differential expression
Select statistically significant (adj. p≤0.01) and biologically relevant (2-fold up/down regulation) genes
(CV ≥ 10%)
Hierarchical clustering of samples (n=10,206 gene probes)
Fig. 1a
Analysis of co-stimulatory/ inhibitory genes
Fig. 5d
Immune cell compartment analysis
Fig. 3a,b,c,e,f Fig 6c
Clinical marker regression analysis
Fig. 5a
n=824 gene probes
Hierarchical clustering of genes = 3 expression clusters
Fig. 1b
Identification of patient-specific responses with Biolayout
Fig. 6a,b
Network analysis with integration of InnateDB gene relationships
Fig. 4 and Suppl. Fig. 2
Investigation motivates removal of one cluster of genes and increased stringency of statistical/biological thresholds (adj. p ≤ 10-5, 4-fold change)
n=52 gene probes
Use as classifier biomarker set, test and validate
Fig. 7 and Suppl. Fig. 7
Network and pathway analysis confirmation of biomarker set
Fig. 2 and Suppl. Fig. 3

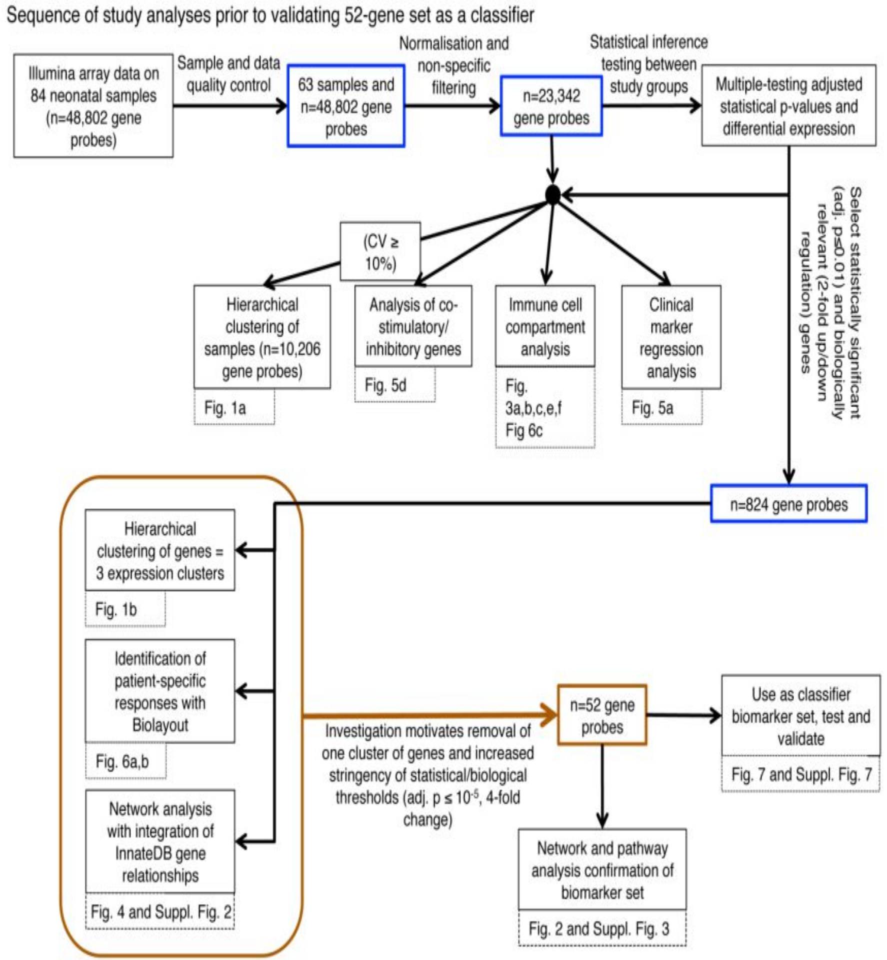
Heat map showing hierarchical clustering

- using eBayes with Benjamini-Hochberg correction adj.P<= 0.01, absolute fold change >=2

- 824 propes






Sequence of study analyses prior to validating 52-gene set as a classifier

Illumina array data on 84 neonatal samples (n=48,802 gene probes)

Sample and data quality control

63 samples and n=48,802 gene probes

Normalisation and non-specific filtering

n=23,342 gene probes

Statistical inference testing between study groups

Multiple-testing adjusted statistical p-values and differential expression

(CV ≥ 10%)

Hierarchical clustering of samples (n=10,206 gene probes)

Fig. 1a

Analysis of co-stimulatory/ inhibitory genes

Fig. 5d

Immune cell compartment analysis

Fig. 3a,b,c,e,f Fig 6c

Clinical marker regression analysis

Fig. 5a

Select statistically significant (adj. p≤0.01) and biologically relevant (2-fold up/down regulation) genes

n=824 gene probes

Hierarchical clustering of genes = 3 expression clusters

Fig. 1b

Identification of patient-specific responses with Biolayout

Fig. 6a,b

Network analysis with integration of InnateDB gene relationships

Fig. 4 and Suppl. Fig. 2

Investigation motivates removal of one cluster of genes and increased stringency of statistical/biological thresholds (adj. p ≤ 10^{-5}, 4-fold change)

n=52 gene probes

Use as classifier biomarker set, test and validate

Fig. 7 and Suppl. Fig. 7

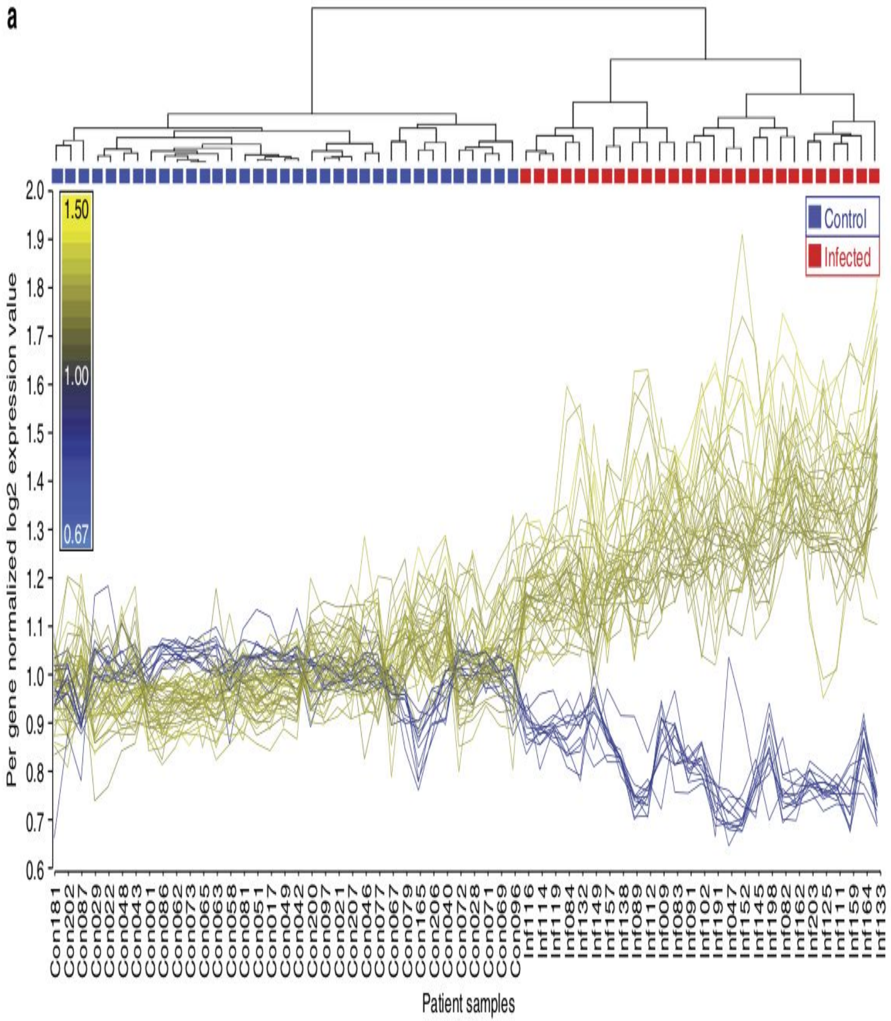
Network and pathway analysis confirmation of biomarker set

Fig. 2 and Suppl. Fig. 3

Pathway biology responses to infection

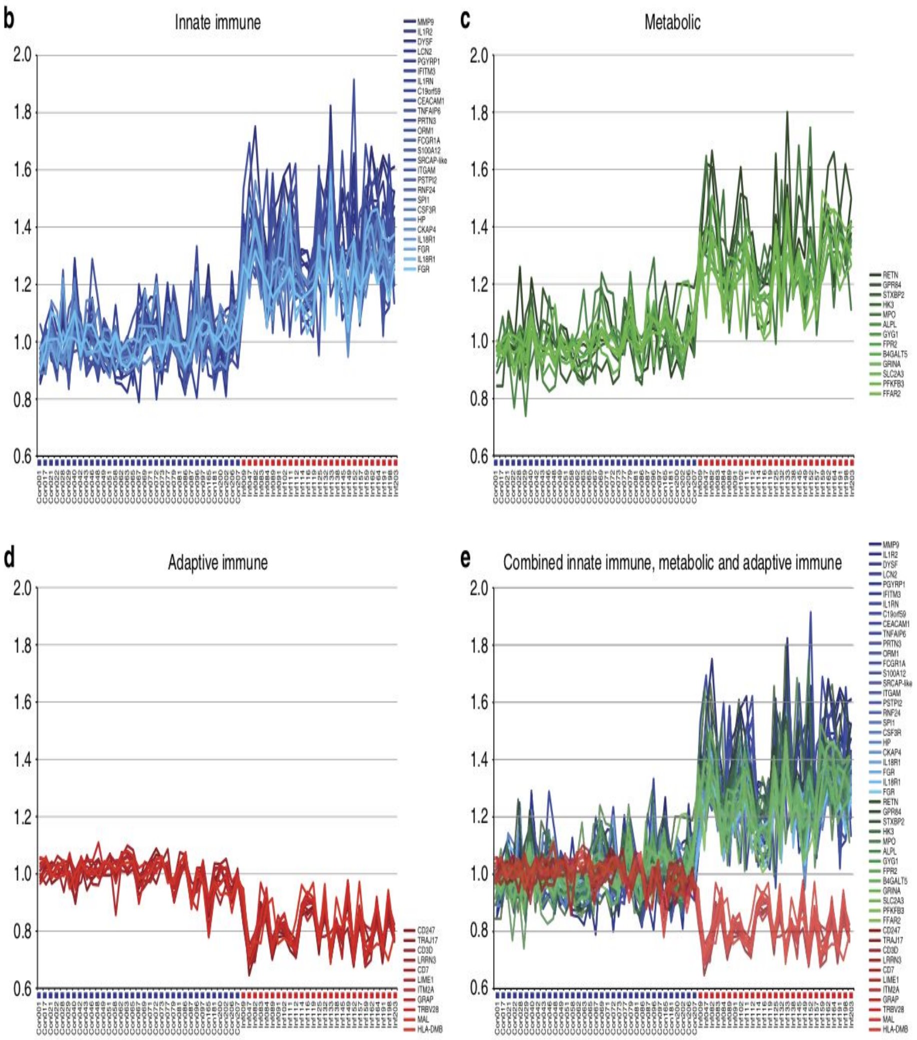
additional filtering to groups 1 and 2 using more stringent cutoffs

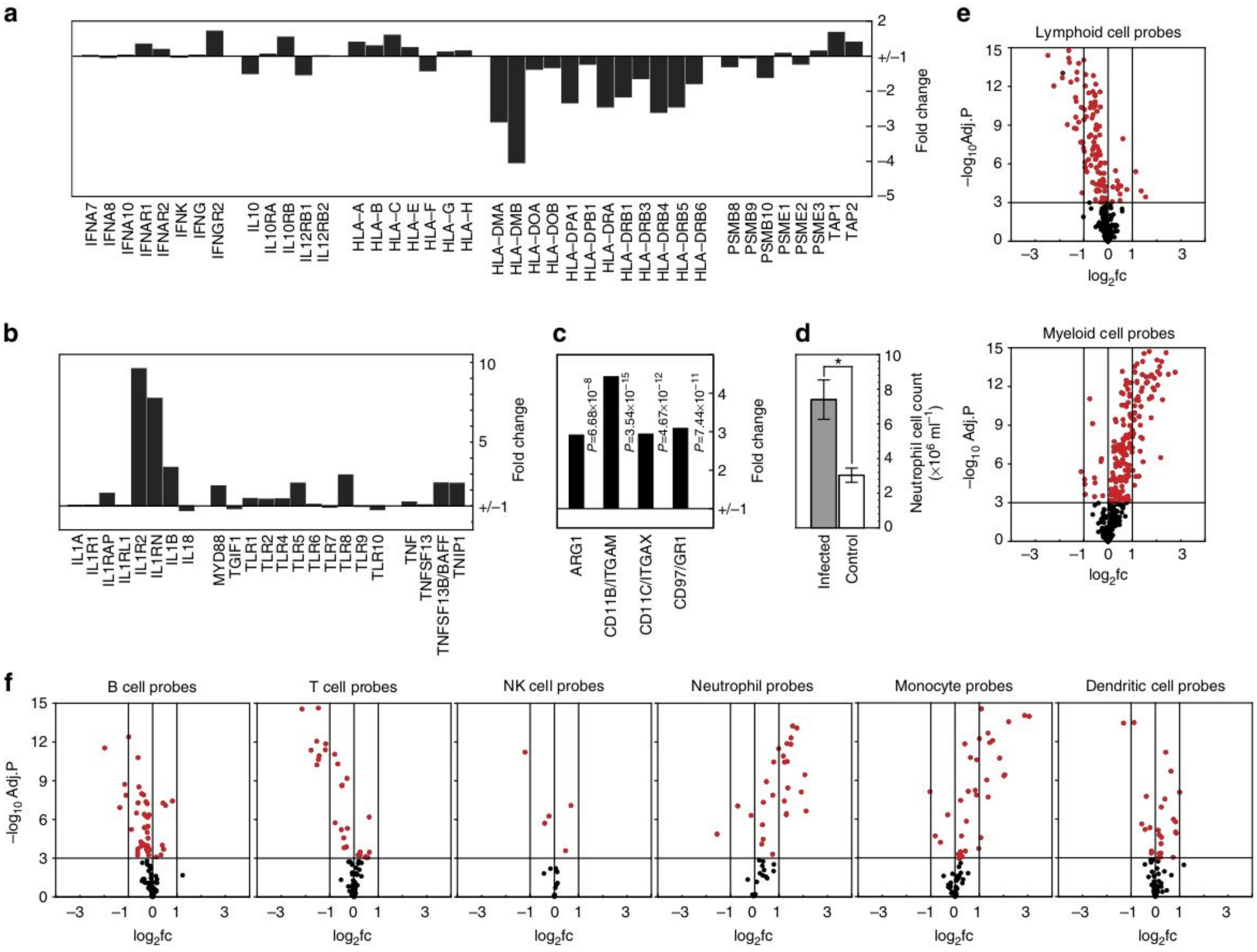
Analysis of the resulting 52 genes revealed sub-networks



Three functional pathway classes

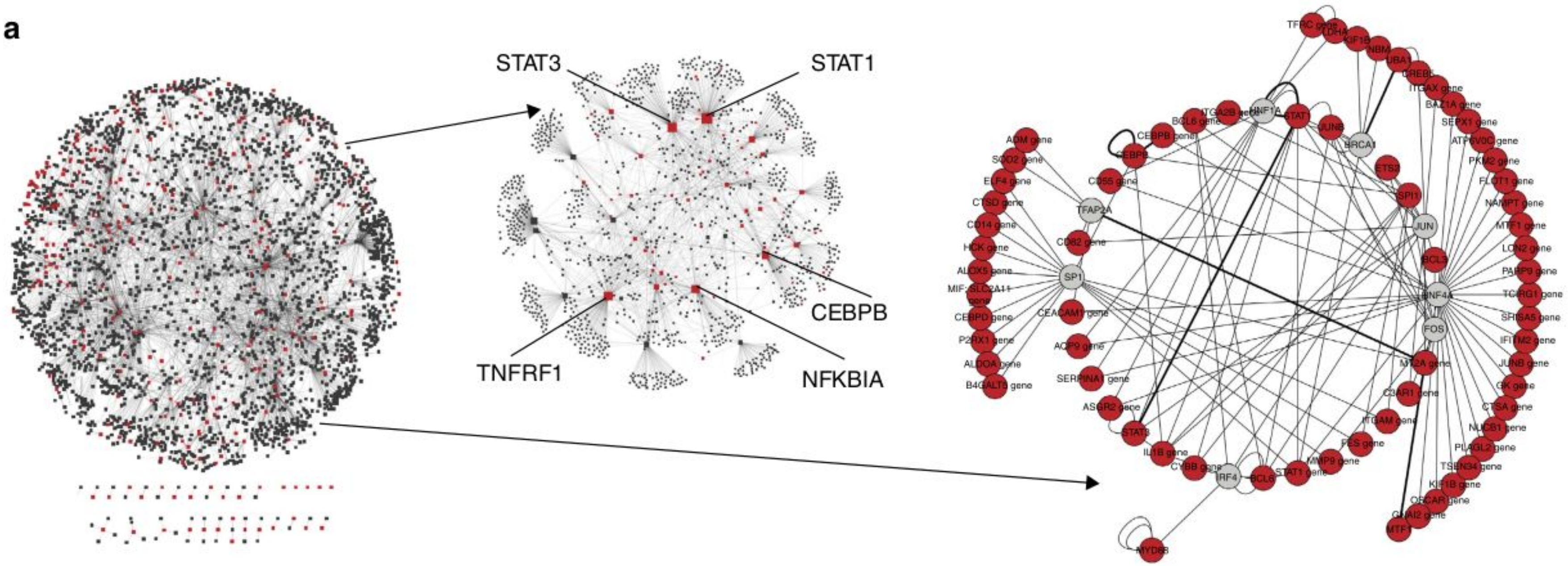
- Innate immunity
- Adaptive immunity
- Unexpectedly metabolic pathways for sugar and lipid metabolism




a
2
+/–1
–2
–3
–4
–5
Fold change
IFNA7
IFNA8
IFNA10
IFNAR1
IFNAR2
IFNK
IFNG
IFNGR2
IL10
IL10RA
IL10RB
IL12RB1
IL12RB2
HLA–A
HLA–B
HLA–C
HLA–E
HLA–F
HLA–G
HLA–H
HLA–DMA
HLA–DMB
HLA–DOA
HLA–DOB
HLA–DPA1
HLA–DPB1
HLA–DRA
HLA–DRB1
HLA–DRB3
HLA–DRB4
HLA–DRB5
HLA–DRB6
PSMB8
PSMB9
PSMB10
PSME1
PSME2
PSME3
TAP1
TAP2
b
10
5
+/–1
Fold change
IL1A
IL1R1
IL1RAP
IL1RL1
IL1R2
IL1RN
IL1B
IL18
MYD88
TGIF1
TLR1
TLR2
TLR4
TLR5
TLR6
TLR7
TLR8
TLR9
TLR10
TNF
TNFSF13
TNFSF13B/BAFF
TNIP1
c
P=6.68×10−8
P=3.54×10−15
P=4.67×10−12
P=7.44×10−11
4
3
2
+/–1
Fold change
ARG1
CD11B/ITGAM
CD11C/ITGAX
CD97/GR1
d
*
10
8
6
4
2
0
Neutrophil cell count (×10⁶ ml⁻¹)
Infected
Control
e
Lymphoid cell probes
15
12
9
6
3
0
–log10Adj.P
–3
–1
0
1
3
log2fc
Myeloid cell probes
15
12
9
6
3
0
–log10 Adj.P
–3
–1
0
1
3
log2fc
f
B cell probes
T cell probes
NK cell probes
Neutrophil probes
Monocyte probes
Dendritic cell probes
15
12
9
6
3
0
–log10 Adj.P
–3
–1
0
1
3
log2fc

Changes in immune cell compartment

“We conclude that in sepsis the heightened innate immune cellular response is driven by monocytes/macrophages and neutrophils and is counter-balanced by inhibitory pathways resulting in a net suppression of the adaptive immune arm, especially associated with the T-cell compartment.”



InnateDB resource filtered for a high confidence manually annotated data set of approximately 2,500 human molecular interactions is used

Regulatory nodes

we conclude that the overriding pathophysiological signal associated with neonatal infection is one of the increased innate immune metabolic responses with an unbalanced homeostatic regulation of adaptive immunity. The specific and intense activation of innate immune signalling, moderated via inhibitory pathways is consistent with the notion of an elevated set point in neonates in comparison with adults for guiding a suppressed adaptive immune response

Metabolic pathway biology and immune homeostasis

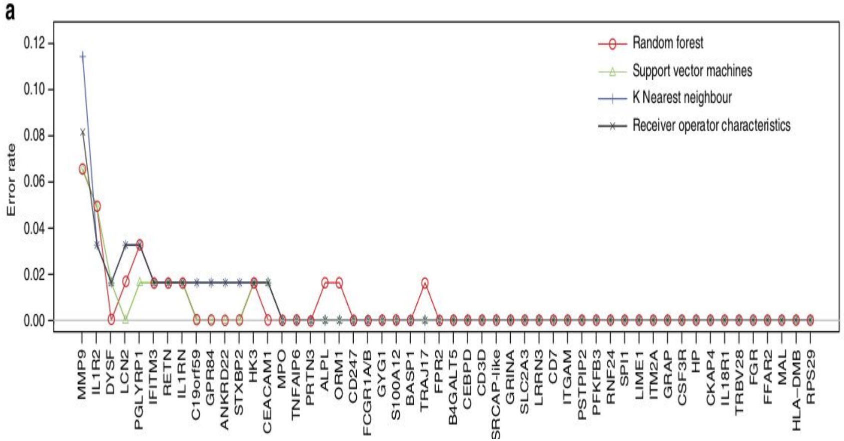
Observed changes in specific metabolic pathways including glucose, energy and cholesterol metabolism.

Including change in B4GALT5 and lipocalin 2, involved in innate response to bacterial infection is also involved in regulating lipid and glucose metabolism.

Indicating theses metabolic processes are likely to be linked with innate immune response

52-Gene dual-network accurately predicts neonatal infection

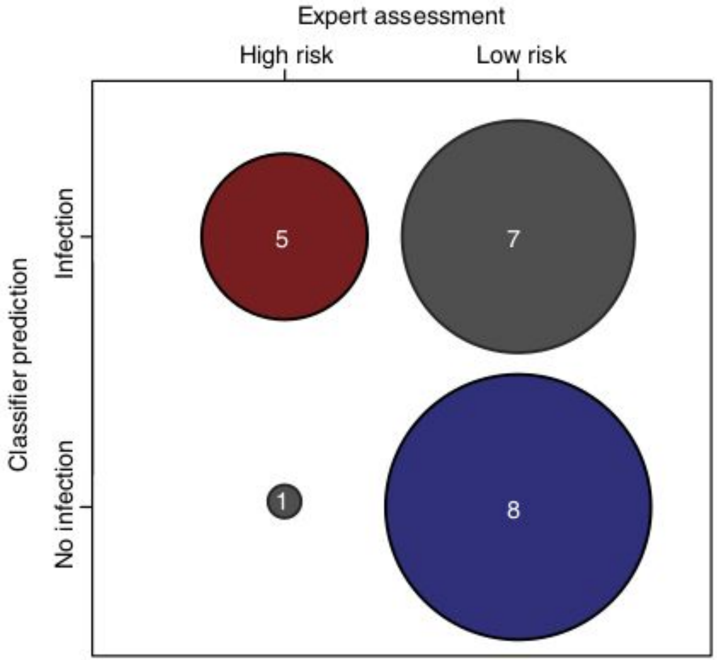
- Four distinct machine-learning algorithms
  - Random forest
  - support vector machine
  - k-nearest neighbour
  - Receiver operator characteristics (ROC)
- When the number of genes included was 19 or more the error rate was consistently 0%*



*consistent with the in-silico model

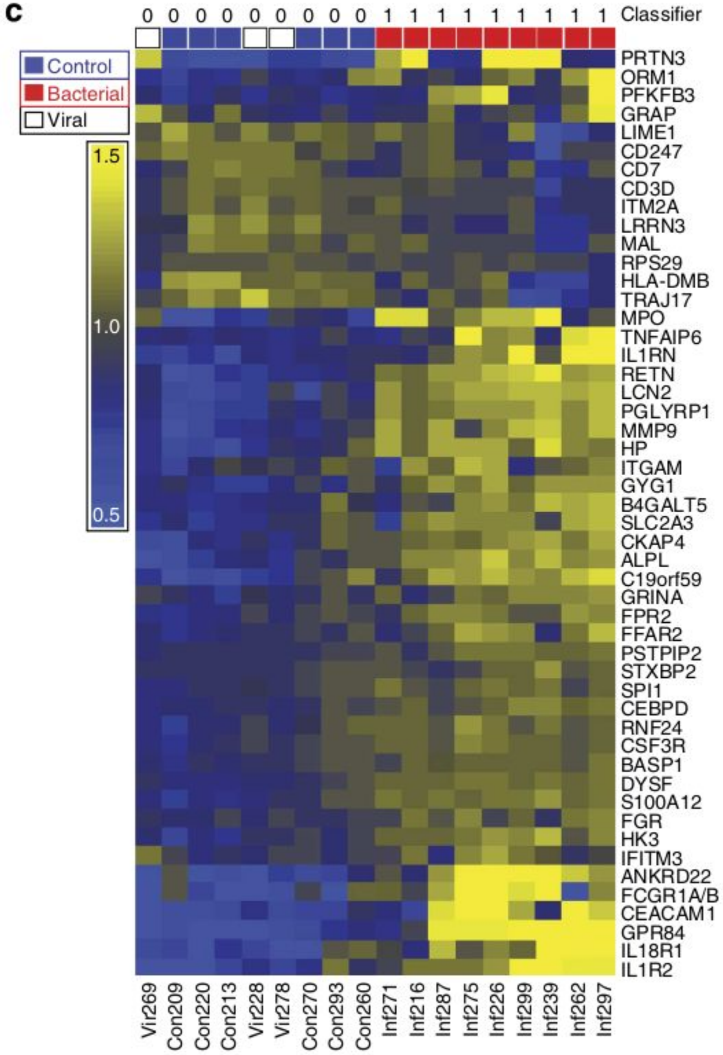
52-gene classifier vs Expert assessment

- These findings highlight the difficulty faced by clinicians in determining cases of blood culture-negative sepsis and strongly support the future clinical utility of the 52-gene classifier



Using other platforms

A ROC classifier when applied to genes with a completely different microarray platform correctly assigned 100% of samples to control or bacterially infected groups (sensitivity 100%, specificity 100%).


c
Classifier
Control
Bacterial
Viral
PRTN3
ORM1
PFKFB3
GRAP
LIME1
CD247
CD7
CD3D
ITM2A
LRRN3
MAL
RPS29
HLA-DMB
TRAJ17
MPO
TNFAIP6
IL1RN
RETN
LCN2
PGLYRP1
MMP9
HP
ITGAM
GYG1
B4GALT5
SLC2A3
CKAP4
ALPL
C19orf59
GRINA
FPR2
FFAR2
PSTPIP2
STXBP2
SPI1
CEBPD
RNF24
CSF3R
BASP1
DYSF
S100A12
FGR
HK3
IFITM3
ANKRD22
FCGR1A/B
CEACAM1
GPR84
IL18R1
IL1R2
Vir269
Con209
Con220
Con213
Vir228
Vir278
Con270
Con293
Con260
Inf271
Inf216
Inf287
Inf275
Inf226
Inf299
Inf239
Inf262
Inf297