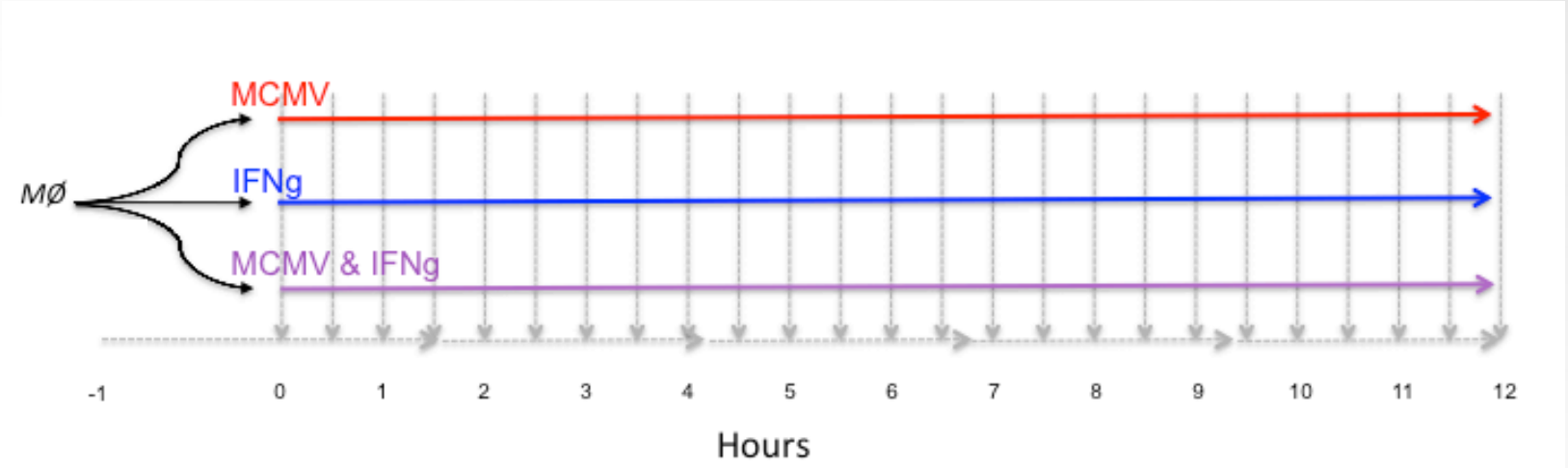
Case Study – Time-shifted correlations

Case study – Source data

A microarray study consisting of 3 time courses, each one with 25 half-hourly time points (0 to 12 hours)

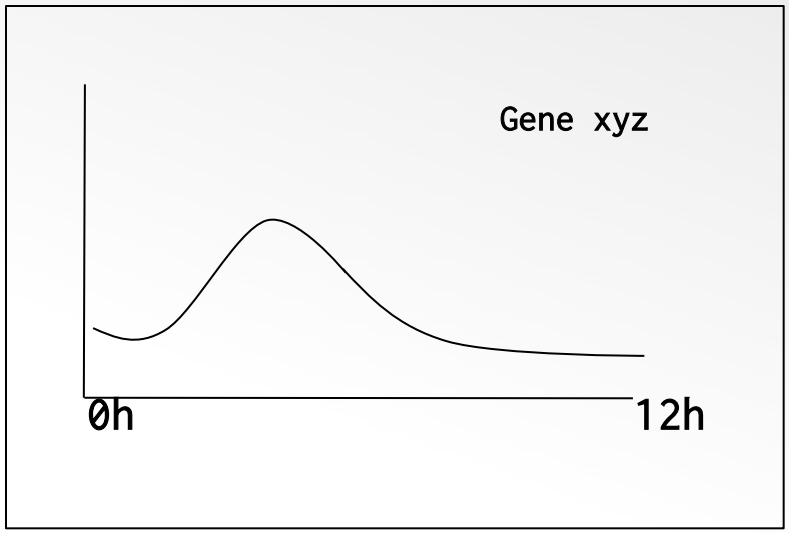


-> 3 conditions, 25 half-hourly time points each (=75 microarrays)

-> 22939 gene probes

-> All 75 samples co-hybridised with a common pooled reference sample

Gene expression profile across 12h



For **each** of the three biological conditions, there are **15K** gene expression profiles (after removal of “flat” genes)

1. Are there any genes which differ between the three biological conditions (in their 12h expression profile)

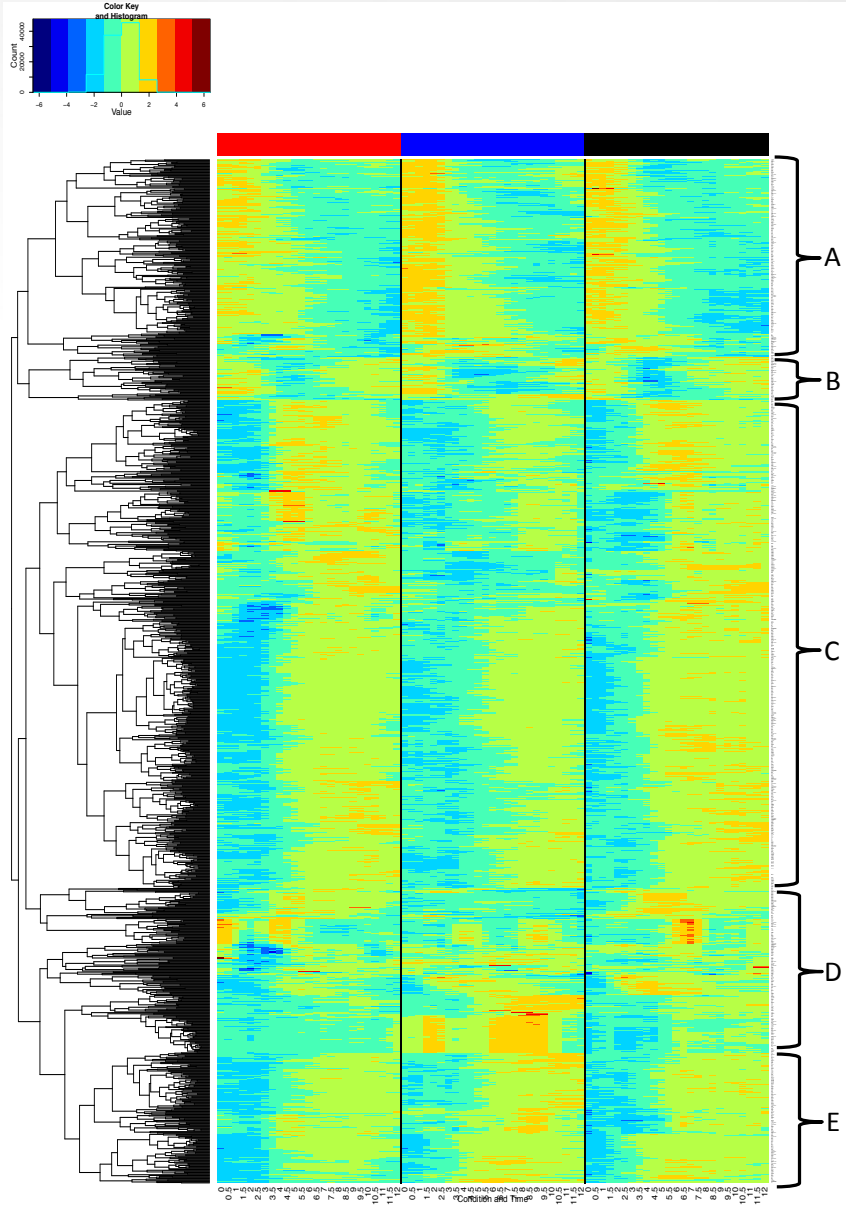
- This data set does not have replicated time point samples, i.e. no stats applicable apart from identifying non-flat profiles

- Instead, use explorative analyses

- SPRINT pcor() to compute all gene-gene distances (1-cor)

- heatmap.2() function to generate clustered heatmap

Case study – Global overview



Genes clustered across all 75 arrays

SPRINT function **pcor()** to compute distance between genes as 1-pcor(t(data))

Distance matrix then supplied to heatmap.2()

There would of course be no need for parallelisation, if a very stringent expression filter is applied before (not after) calculating gene-gene similarities.

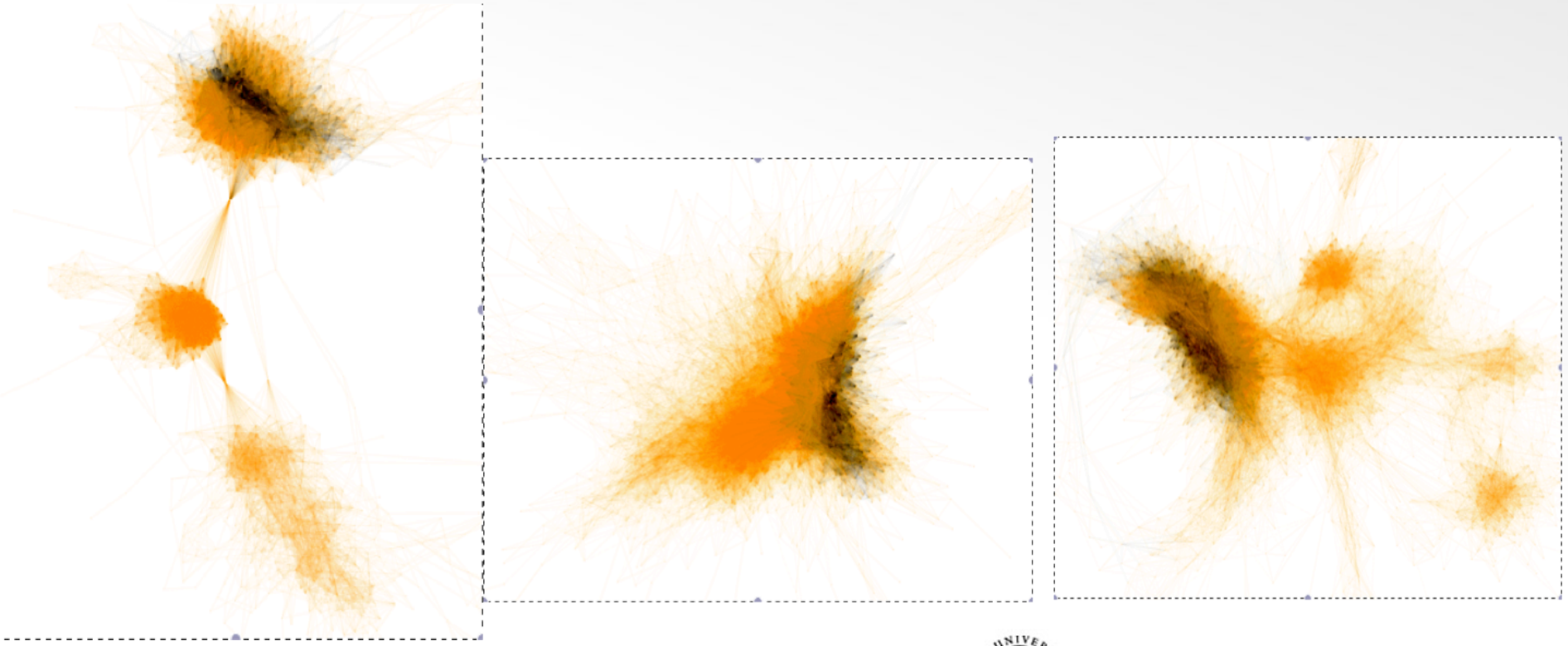
2. Do the biological conditions differ in their topology?

- Use SPRINT pcor() to compute all gene-gene correlations
- Filter data *after* all correlations have been computed
- Generate network graph (Cytoscape)

Case study – Gene co-expression network graphs

1. All pairwise gene-gene correlations for each biological condition
2. Reduce correlation matrix to significant independent correlation (**PCIT**)
3. Reduce further based on how much graphing package can handle
4. Plot networks (Cytoscape)

SPRINT function **pcor()** to compute correlation between genes: pcor(t(data))

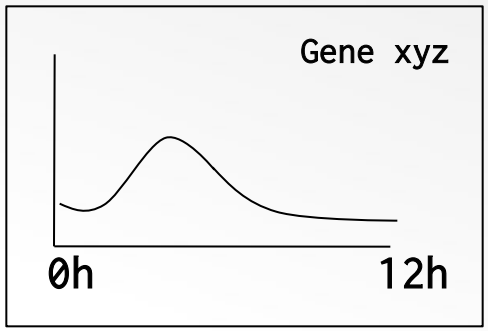


3. We now know how genes are expressed similarly or differently across the entire 12h of observations.

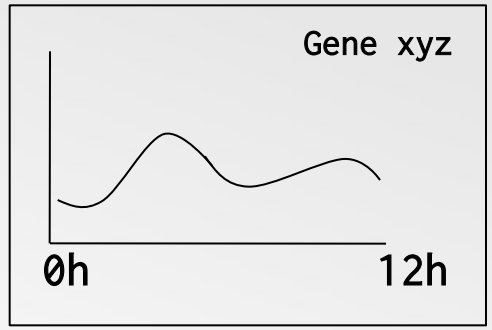
- But can we say anything about genes matching other genes *in part*?
- And matching in part to genes in another biological condition?

Case study – Time-shifted correlations

Hypothesis: genes are correlated **between** conditions, shifted in time

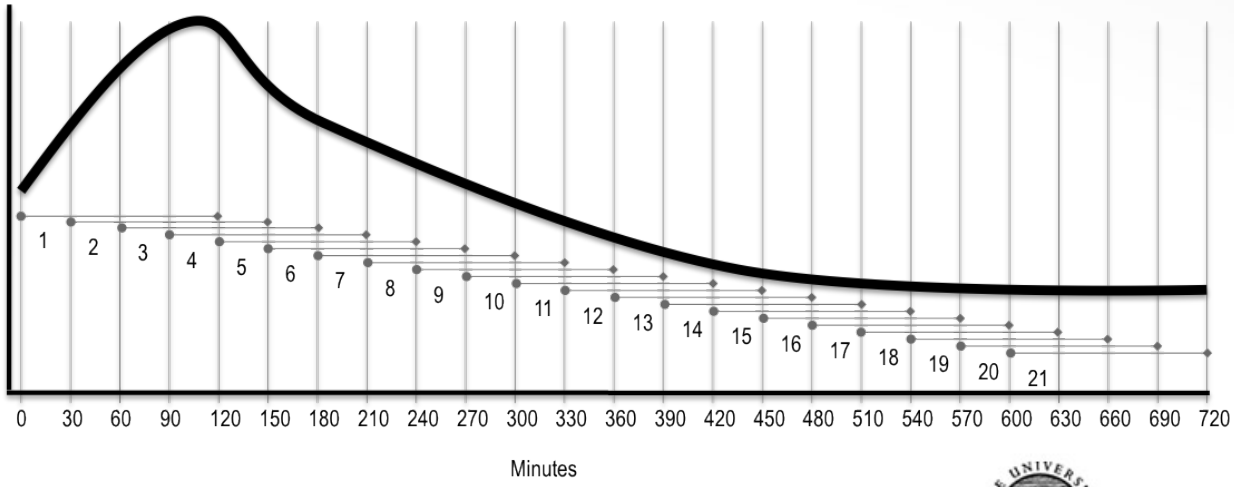


IFNg treated macrophages



Virus infected macrophages

Split 12h profile into 21 time windows of 2h each (overlapping)



THE UNIVERSITY of EDINBURGH

But if we split the 12h time course into multiple time windows for each of the 15K genes and in each of the biological conditions, the required number of correlations to be measured increases…a lot.

Case study – Computational demands

Number of Pearson correlation coefficients calculated when comparing macrophage IFNg time course to macrophage virus time course

We wish to compute all correlations of short time windows **between 2** conditions

(14819 genes * 21 time windows)^2 = **96.84 billion** calculations

Time: **NA**

The above is too much for a Mac*, use a much smaller set of genes instead

(5561genes * 21 time windows)^2 = **13.64** billion calculations

Time: **~3h**

Same computations With **SPRINT pcor()** on 256 cores
(Note: …forgot if this time was for 5561 or 14819 genes…)

Time: **~10min**

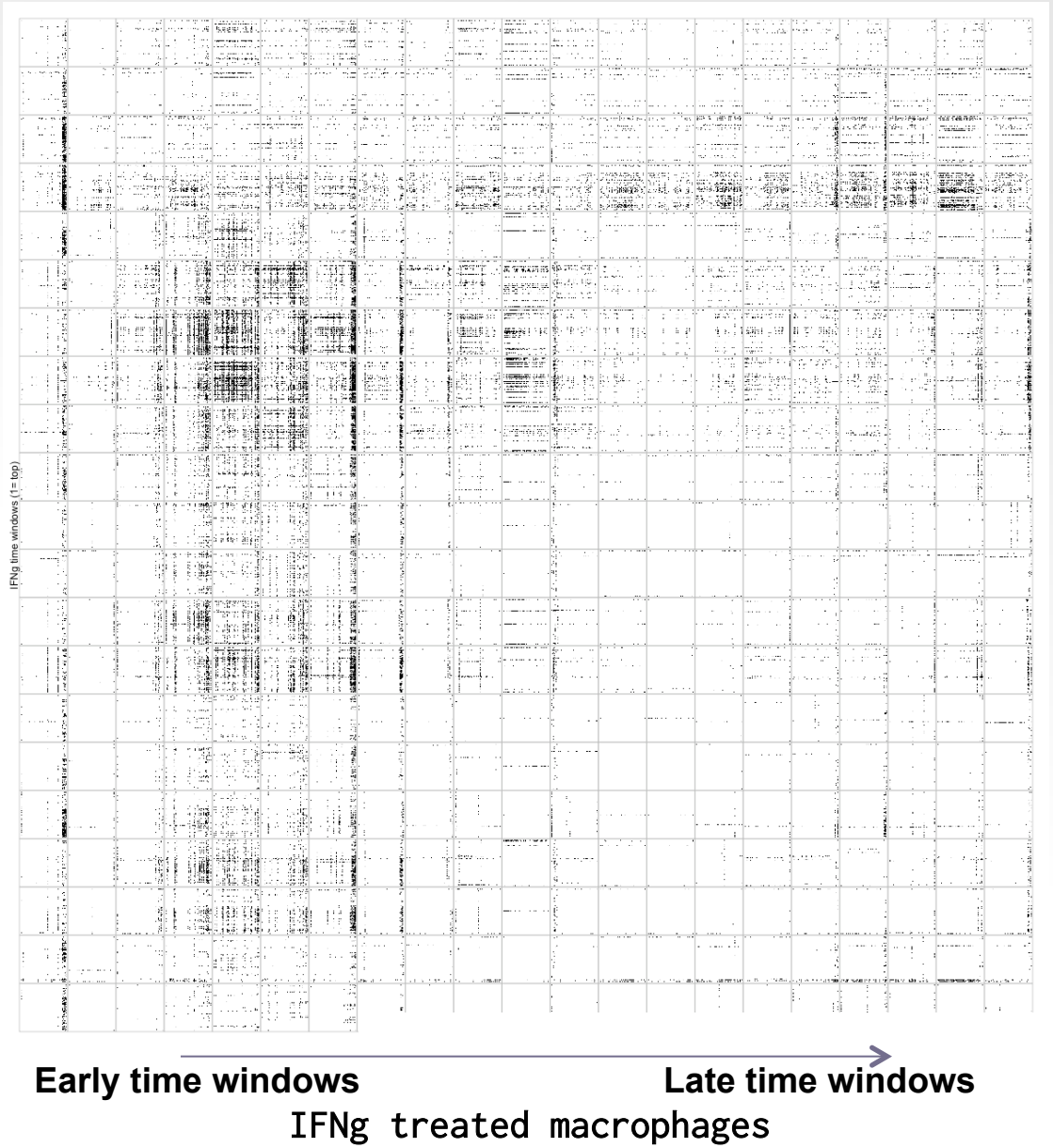
*At the time this was run, R was only using a single CPU even if the machine was multicore

Case study – Visualised results

How many genes correlate in which time windows

Data point = gene expression profile match between two biological conditions.

Each panel = one combination of time windows.

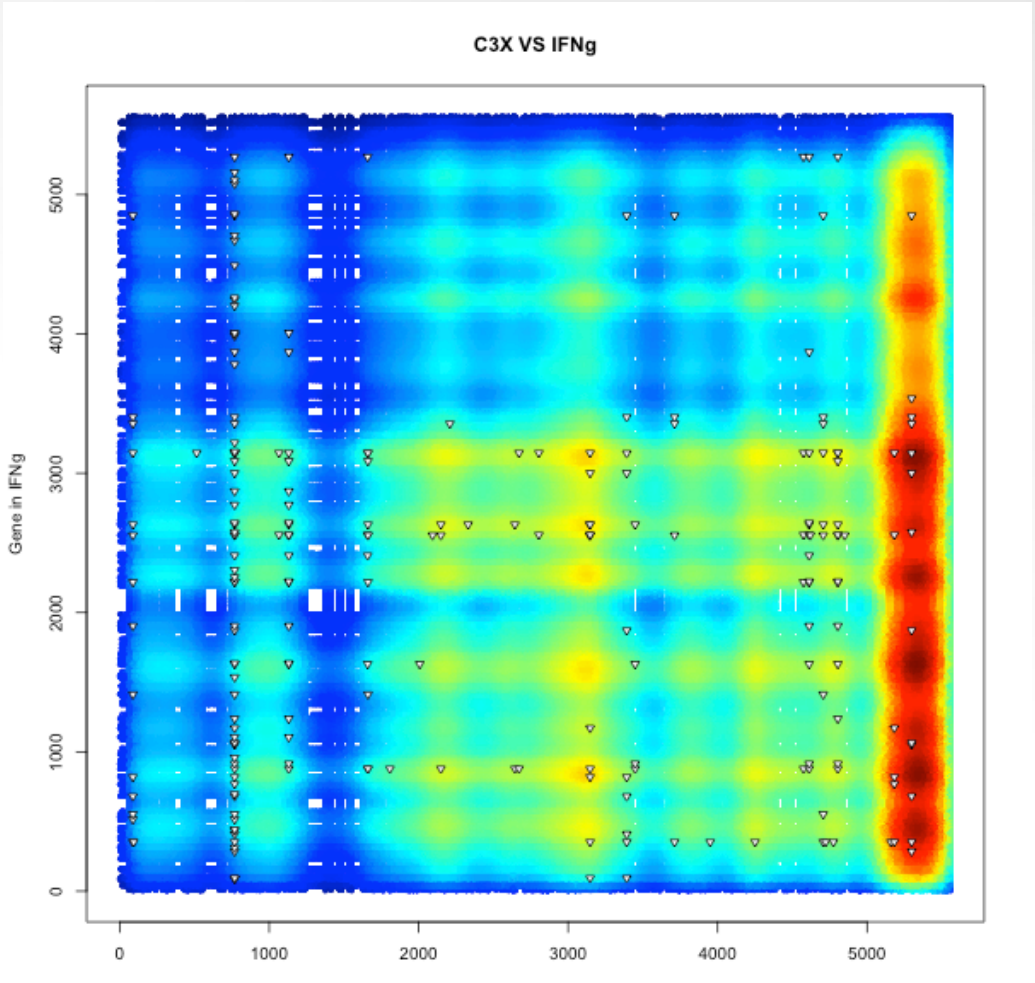


Per-gene, per time-window correlations can be aggregated to gene level.

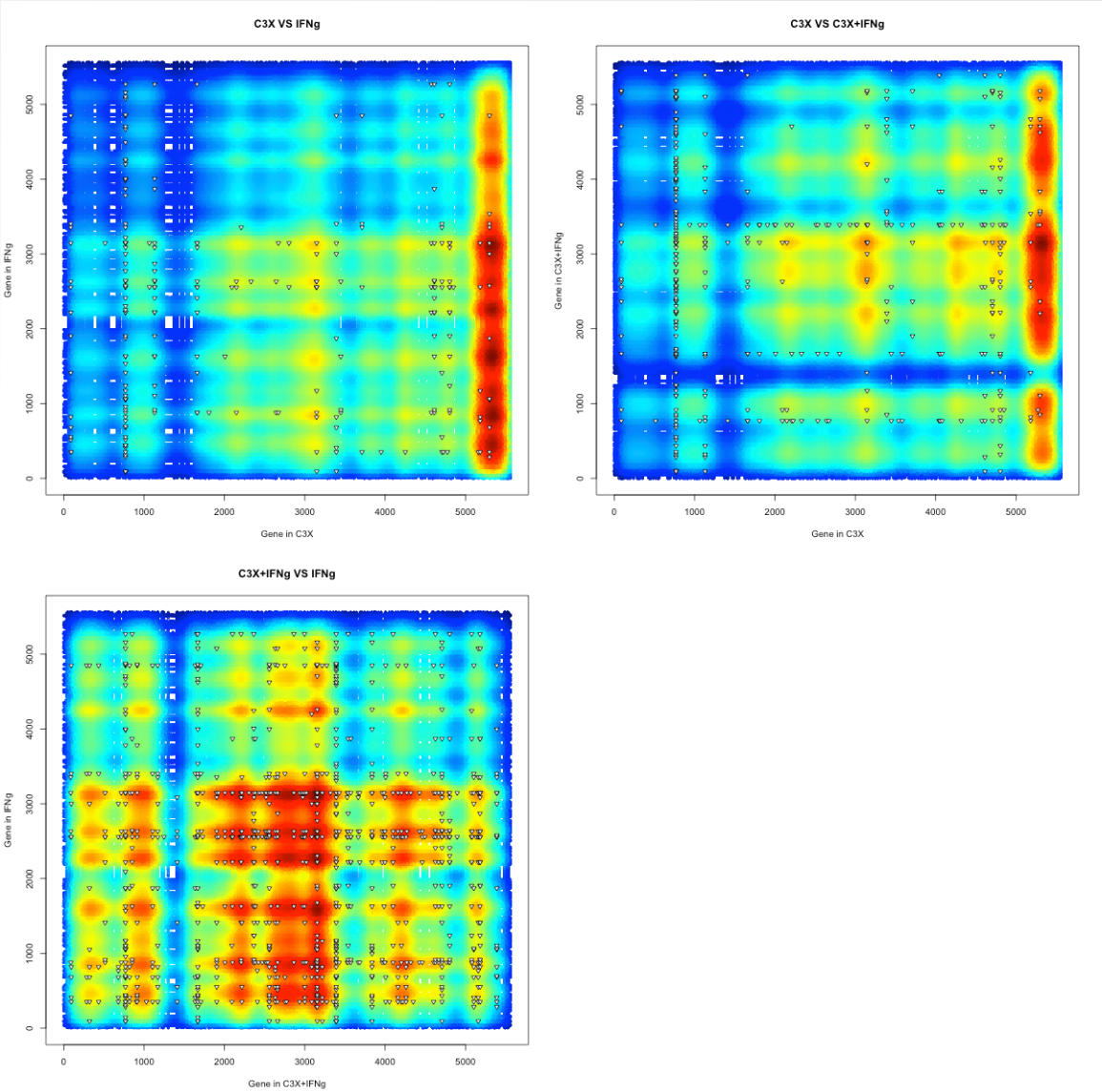
How frequently does the part-expression-profile of one gene match part-expression-profiles of other genes in another biological condition?

Case study – Visualised results

Red represents gene-to-gene "links" that occur frequently over time.



Case study – Visualised results



We are now investigating if genes that *in part* match a high number of other genes in another biological condition (also *in part*):

1. Differ numerically or biologically from the set of full 12h matches

2. Provide any new hypotheses on cause-and-effect transcription networks in macrophage activation.

And we are keeping in mind that correlation-based explorative analyses are not as focused as statistical hypothesis testing…

Example application for papply() and pboot()

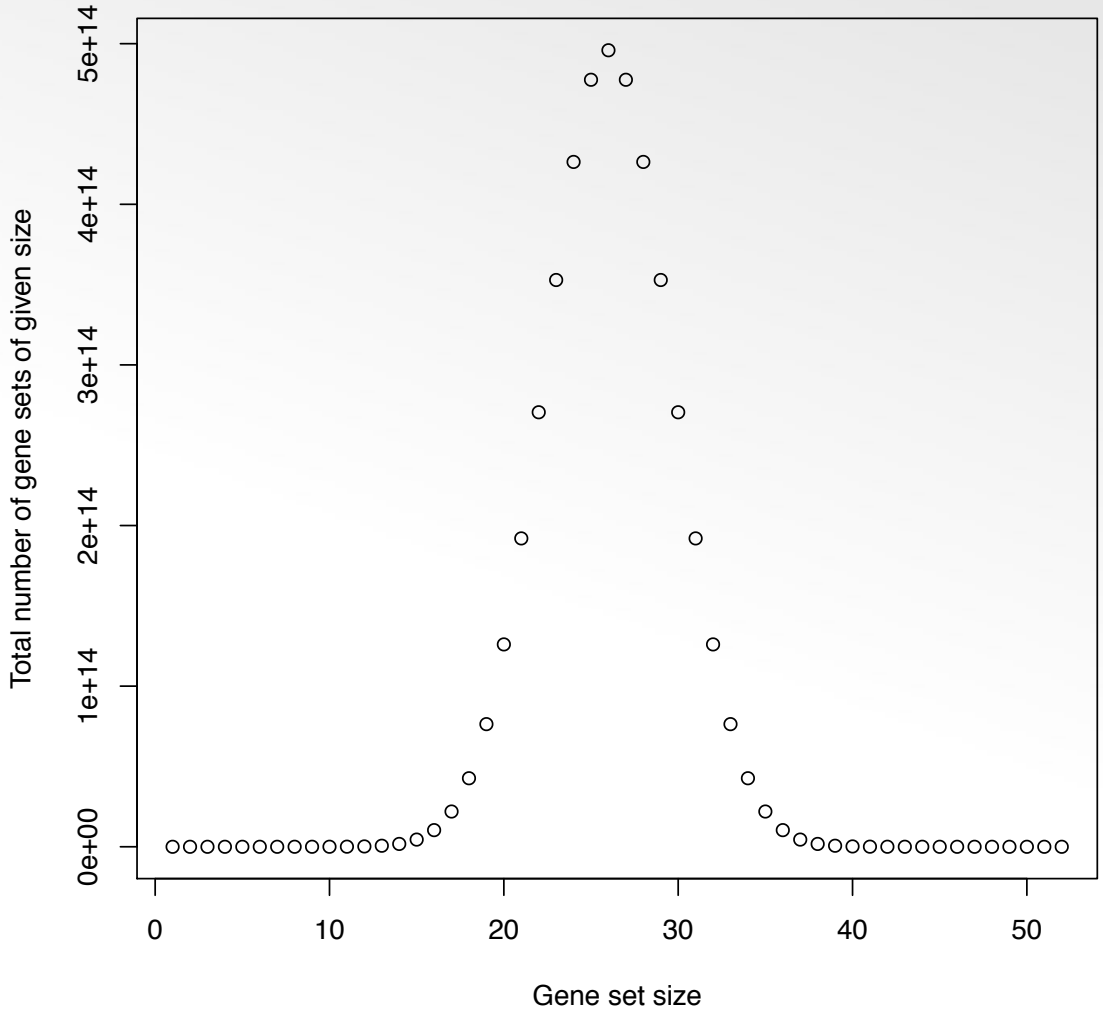
Assume you have identified (biologically relevant and statistically significant) 52 genes that can predict if a microarray-hybridised blood-RNA sample is a sepsis case or a healthy control.

Question: do you need all 52 genes to be part of this classifier? Would smaller gene subsets work as well and reduce cost for a clinical assay?

Methodology 1: Rank genes by some criteria and test top N=1,…,52 as classifier

Methodology 2: Test all possible subsets of N genes between size 1 and 52. That's each gene individually, all possible sets of 2 genes, all possible sets of 3 genes etc.

…that's 4.5 quadrillion possible gene sets to test, which at 1 sec per run takes ~140 million years



That's too big for parallelisation.

a) Use MCMC sampling scheme? We haven't parallelised this in SPRINT…

b) Test only lower sizes of gene sets? If you reduce the problems to smaller gene set sizes, you may get away with a few million computations.

-> Use papply() and/or boot() to distribute across as many cores as possible to get results in a reasonable time frame of days.