Comparison of Li's WGCNA approach for paired data with other biostatistical methods on two different paired tumour datasets

Francine Diethelm[1] **and Florian Hellwig**[2]

[1]**Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland**

Abstract

WGCNA has become a standard biostatistical method for independent data. In 2018, Li et al. extended the procedure to paired data. Neither the paper nor any follow-up has done an evaluation of the effectiveness of WGCNA on paired data. This paper compares it with seven further biostatistical methods, some standard, some deviations from other methods, assessing advantages and disadvantages, by adapting the pipeline of WGCNA at different places, as well as using different approaches (Limma, DiffCoExp and Glasso). The comparison was done using two different paired tumour datasets. While the first one was the same as that Li et al. used in their paper, with 40 patients that each contributed two samples, the second one uses 100 pairs of samples. At the end, the results are compared and checked with what is known from other sources. The study finds a method that works more smoothly in some way than Li et al.'s paired WGCNA method on the two datasets considered.

Keywords

WGCNA, paired data, comparison, Glasso, Limma, DiffCoExp

1 Introduction

In human transcriptomic studies, inter-individual variability often explains a substantial fraction of the observed expression differences. Factors such as genetic background, baseline physiology, and unmeasured environmental influences can therefore mask condition-related effects in purely cross-sectional analyses. A paired design (e.g., pre/post measurements or matched tumour/normal samples) mitigates this problem by using each individual as their own control, such that stable subject-specific differences largely cancel out and the analysis can focus on within-subject changes.

In addition, transcriptomic alterations associated with disease or treatment typically do not manifest as single genes with large effect sizes, but rather as modest, coordinated shifts across groups of genes. Network-based approaches such as WGCNA are well-suited to capture such patterns by clustering genes into co-expression modules and highlighting highly connected hub genes that may be mechanistically informative. When WGCNA is implemented in a way that respects the paired structure, it becomes possible to identify modules whose overall activity and/or connectivity changes systematically between the two states within the same individual, which can yield signals that are more directly attributable to the underlying biological process than those obtained from analyses that ignore pairing.

Weighted gene co-expression network analysis (WGCNA) is a statistical method which builds gene modules based on strong co-expression values across samples. These modules can then be associated with external traits by identifying gene-phenotype relations. The standard pipeline for WGCNA is built for dealing with independent samples. But when normal and tumour samples are taken from the same person, also called paired data, the underlying dependency calls for an adapted strategy. In [1], Li et al. suggest a modified pipeline for WGCNA to paired data. They apply their pipeline to a dataset (GSE45238), and they discuss their results. When proposing a new statistical method, it is very important to show that it delivers robust results. One way to achieve this is by comparing it to other methods which are known to be reliable. This comparison was, however not part of [1], nor has it been addressed by any other biostatistical paper in the meantime and will therefore be the focus of this paper.

First, we reproduce the original WGCNA pipeline from Li et al. as described in [1]. The results of that analysis will then be compared to seven other methods. Four of those methods are based on WGCNA with modifications to the workflow: The first method, we compare it to is so-called "naive" WGCNA (ii). While the building of the modules is the same, the significance of the individual modules and genes is calculated directly by using the Pearson correlation, not taking the road of linear-mixed models. To further explore the path of the original, independent WGCNA, next, a term very similar to the log ratios of the expression values of every patient are calculated. This is then again used as a basis of a (signed) independent WGCNA analysis (iii). Next, the so-called "pair-aware" WGCNA (iv) is implemented, accounting for the common within-subject association by first subtracting the baseline expression for each patient from the expression values. Finally, WGCNA networks are built separately for the tumour data and for the normal data, respectively. It is then compared which modules and genes are included differently in the networks (v) (in the appendix). The remaining three methods are non-WGCNA-based: (vi) Limma for differential expression with a paired design, (vii) DiffCoExp for differential co-expression, and (viii) Graphical Lasso for sparse conditional-dependence network inference. Further details can be found in the methods section.

Each method was then compared to Li et al.'s paired WGCNA, by comparing the obtained modules and hub genes. For this, for WGCNA-similar methods the cartesian product of the modules returned by the two methods was computed, and additional scores like the module assignment agreement ARI and Jaccard index as well as Spearman's correlation test were used. Limma was compared to paired WGCNA by contrasting their ranked lists of phenotype-associated miRNAs using top-N overlap with a one-sided Fisher test, and by checking how limma hits distribute across pairedWGCNA modules. DiffCoExp was compared by analysis hub genes from both methods and how they overlap. Further, the links and genes highlighted from DiffCoExp were examined closer, which paired WGCNA modules they fall into and whether they fell more within or between modules. Graphical lasso was compared by testing whether glasso edges concentrate within paired WGCNA modules versus a size-based expectation and by comparing glasso hubness metrics to paired-WGCNA hub rankings.

Finally, a limited sample of hub genes returned by each method was examined how well known it is empirically to be associated with the corresponding tumour. A discussion concludes the paper.

Note: This paper does not primarily assess the biological advantages and disadvantages of the method proposed by Li et al., but rather whether it returns better and more methodology-sound results than other methods in a standard benchmark comparison. An important biological limitation, however, is that each gene (or miRNA) is assigned to exactly one module, implying that modules are assumed to be disjoint (with unassigned features typically placed in the grey module).

2 Methods

In the following, methods (iii) log ratio WGCNA and (iv) pair-aware WGCNA are explained. The explanation of the further methods can be found in the appendix, however, they follow the standard procedure as known to the field and thus, can be used as a check-up reference only.

2.1 (iii) WGCNA on the log ratios

While the Naive WGCNA approach on paired data is statistically misspecified for paired designs, as it makes assumptions that are not given, WGCNA on the log ratios tries to stick with the original, independent WGCNA method, but correcting for the wrong assumption. This is done not by asking whether genes co-vary across all tumour + normal samples, but whether genes have coordinated tumour-vs-normal changes across patients?
In the preprocessing of the smaller, original dataset, the expression values are transformed by

$$Y_{ij} = \log_2(X_{ij} + 1),$$

while in the larger dataset, the offset of $c = 50$ instead of $c = 1$ is added before taking a log2, i.e.

$$Y_{ij} = \log_2(X_{ij} + 50).$$

Here, Y_{ij} are the transformed expression values, while X_{ij} are the raw expression values. Note, however, that in the second case, things are a bit more complicated as the samples are reconstructed from channel intensities. Both datasets are later also normalised.
Before starting with WGCNA at all, the delta values of the expression values for all genes of each patient i are calculated, i.e.

$$\Delta_i = \mathrm{Y}_{\mathrm{tumour},i} - \mathrm{Y}_{\mathrm{normal},i}.$$

Because of the transformation done before, this corresponds to

$$\Delta_i = \log_2(\mathrm{X}_{\mathrm{tumour},i} + c) - \log_2(\mathrm{X}_{\mathrm{normal},i} + c) = \log_2\left(\frac{\mathrm{X}_{\mathrm{tumour},i} + c}{\mathrm{X}_{\mathrm{normal},i} + c}\right).$$

Δ_i is the vector of gene-wise differences, with the individual elements being Δ_{ig} for a gene g.
So, actually, the delta values are the transformations of the tumour-normal raw expression value (plus some offset) ratios by the logarithm. Hence, they will be called log ratios. Note that this is not true in detail: After the log2-transformation, a quantile normalisation is applied. Therefore, strictly speaking, they should be interpreted as a normalised log-scale difference rather than an exact log-ratio of raw intensities. Nevertheless, for simplicity, the term log-ratio will be used.
The delta values of different patients are independent. Therefore, we can now use those as the values in the sample x genes matrix: The delta values of each patient are one row of the new matrix. On this matrix, we can apply the original, independent WGCNA as explained in "Naive WGCNA".
The only difference to the method described in (ii) is that instead of using the unsigned method for the adjacency scores, the signed method is the way to go. The reason for this is that it should be kept track on whether the ratio of tumour vs. normal is 2 or $\frac{1}{2}$, otherwise a bias is introduced. The signed method divides the correlation values by 2, adds $\frac{1}{2}$ and then again takes the power to the soft threshold, i.e.

$$a_{ij} = \left(\frac{1 + \mathrm{cor}(x_i, x_j)}{2}\right)^{\beta}.$$

As $-1 \leq \mathrm{cor}(x_i, x_j) \leq 1$, the adjacency scores are different for 2 and $\frac{1}{2}$ to use the example from before. For consistency, this signed approach is also used by the Topological Overlap Measure.
The downstream analysis follows mostly the original, independent WGCNA pipeline, outlined when explaining the Naive approach. However, one further problem arises: Because the samples are not split into tumour/normal anymore, the correlation $\mathrm{cor}(x_i, \mathrm{trait})$ could be used only on the variable stages of the tumour sample (similar for the standard gene significance further below). As this is not the wanted module-trait association measure nor always possible (the second dataset does not have a variable indicating the stage of the cancer), a different solution must be considered. Indeed, the wanted tumour effect is the mean of all the tumour/normal log ratio expression values of the genes in the module m. Note that this mean is still a vector with one value for each patient i, i.e. the elements of the vector are

$$S_{im} = \frac{1}{\mathrm{card}(m)} \sum_{g \in m} \Delta_{ig}$$

If the mean is statistically unequal to 0, there is an association between the genes and the trait. To evaluate this for one specific module, a one-sample t-test vs the control value $\log_2(1) = 0$ is done, using the values of this mean vector as elements of a sample. After correcting using Benjamini–Hochberg FDR, the resulting p-values can then be used to assess the module-trait association.
Note that the module eigengene cannot be used, as it represents the first principal component of the module log ratio expression matrix. As PCA Scores are centred, the mean of the module eigengene values will always be close to 0.
Similarly, the gene significance can be replaced with a one-sample t-test of the log-ratio expression values for one specific gene.
The other steps of the original, independent WGCNA pipeline can be used unchanged (including the module membership and the intramodular connectivity).

2.2 (iv) Pair-aware WGCNA

This method is similar to Li et al.'s approach. It neither assumes independence, nor does it work with ratios, but it accounts for the paired data once again using a linear mixed model. The decisive difference to the approach of Li et al. is that the modules are built from within-patient centered expression, so that correlations (and thus the network) are driven by within-patient repeated measures rather than between-patient differences.
After preprocessing, the expression value matrix is changed in a way that each patient's baseline level is removed. This means, first for every patient i and gene g, the patient mean $\overline{x}_{ig}$ is computed:

$$\overline{x}_{ig} = \frac{x_{\text{tumour},ig} + x_{\text{normal},ig}}{2}.$$

In the next step, this mean is subtracted from the expression values, i.e.

$$x_{ijg}^{\text{within}} = x_{ijg} - \overline{x}_{ig}.$$

By construction, for each patient i and gene g, the mean of x_{ijg}^{within} across $j \in \{\text{tumour}, \text{normal}\}$ is zero. Any further downstream analysis will be done on these within-patient values. The consequence of this is that the correlations between the within-patient expression values of the genes will be largely driven by how the gene co-varies within patients across the repeated measures (tumours vs. normal), rather than being driven by stable between-patient differences. (Note that not actually a rmcorr model is fit, but the idea of focussing on the within-subject signal is the same.)
The downstream analysis is the same as the pipeline proposed by Li et al. Important, however, is which data is used for which. Using these within-patient expression values, Soft threshold selection, construction of the adjacency matrix and topological overlapping measures using the unsigned method, average hierarchical clustering and cutting the dendrogram using `cutreeDynamic()`, resulting in the first set of modules is performed. For the further steps, for comparability and meaningfulness, the pooled matrix (with the original tumour and normal expression values) is used: The computation of the module eigengenes and possible merging of some modules if their eigengenes are strongly correlated, construction of a linear mixed model for every module and return of the t-values corresponding to the tumour coefficient is all done based on the original, pooled data. To identify the hub genes, the paired gene significance and module membership are computed based on the pooled data, while the intramodular-connectivity is derived from the adjacency matrix based on the within-patient expression values. Also this approach returns most importantly the modules and the corresponding t-values from the linear mixed models as well as the proposed hub genes with gene significance, module membership and intramodular-connectivity as test statistics.

2.3 (vi) Limma

Limma is a useful R package for analysing gene/miRNA expression data. It fits a linear model to each gene/miRNA across samples and then applies empirical Bayesian methods to make the results more stable. A very classical use case for Limma is comparing expression values over two different states and determining whether this change is statistically significant. This is exactly why we can use it on our datasets. We want to know how miRNA expression changes between normal and tumour samples.
The miRNAs which Limma will determine to be the most significant will be compared to the hub genes that we have received via method (i).
DiffCoExp is a statistical method used for finding differentially co-expression links (DCLs) and differentially coexpressed genes (DCGs). DCLs are pairs of genes which show a significant correlation change between two conditions. DCGs are specific genes which show up in more DCLs than given by chance.
When applying this method, the data is first split by condition and separate dataframes are created. These two dataframes are then given to the *diffcoexp()* function, which yields two new dataframes, DCLs and DCGs. DCLs are identified by computing correlations for each gene pair in both conditions and testing whether the correlation difference is significant, followed by multiple-testing correction to control the false discovery rate. DCGs are then obtained by counting, for each gene, how many significant DCLs it participates in and assessing whether this count is larger than expected under a null model, again reporting FDR-adjusted significance values.

2.4 (viii) Graphical Lasso

Graphical Lasso is a statistical method which estimates a sparse Gaussian graphical model. This is a network where nodes represent variables and edges represent conditional dependencies. This model is built by estimating the inverse covariance matrix, where many entries are zero, and the resulting network is therefore sparse, meaning it has many fewer connections between nodes than the highest possible number of connections. If two nodes are not connected, the corresponding variables are considered independent given all the other, while a connection implies dependence.

3 Further remarks

3.1 `disableWGCNAThreads()`

To ensure maximal reproducibility, we disallowed multi-threading in the WGCNA procedure for both datasets.

3.2 `blockwiseModules()`

Classic WGCNA has a runtime lying in $O(p^2)$ where p is the number of genes. Having close to 20'000 genes as in the second dataset makes things impossible to run on a normal computer. The solution to this problem is `blockwiseModules()`, which was implemented in the WGCNA methods for the second dataset. This function requires a parameter `maxBlockSize` and then assigns genes into blocks of size $\leq$ `maxBlockSize` so that the genes of the different blocks are weakly correlated. In the following, this between-block correlation is ignored. WGCNA is then executed on each block, and modules are created. Subsequently, modules across blocks with module eigengenes that are highly correlated are merged. The merged modules are then returned, including all information from the classic WGCNA.

4 Results

4.1 Analysis on the first Dataset GSE45238

For transparency and to show the actual outcome of the methods, no further manual merging was done of the modules. Also, for comparability, `cutreeDynamic()` was always executed with `minClusterSize = 30` and `deepSplit = 2` when used.

In the following comparison, the module assignment agreement ARI index, Jaccard index and Spearman correlation tests are used. As a last measure, the top hub genes of each method should be compared. To make this comparison biologically meaningful, it was decided to calculate for each method and each gene the kWithin value, so the sum of adjacency scores between the gene and other genes in the same module, normalise this as follows:

$$\text{kWithinNorm}_i = \frac{\text{kWithin}_i}{\text{card}(m) - 1}$$

and then order the genes by the just calculated measure for each method. Now the N genes of each method with the highest kWithinNorm value are compared, and the percentage of intersection is calculated.

4.1.1 (i) WGCNA on paired data

Doing the diagnostics for choosing the soft threshold returned the plot shown in Figure 1. The red line is at an R^2 value of 0.97, far higher than the usual $R^2 = 0.9$. To reproduce Li et al.'s paper as well as possible, this paper also chose $\beta = 7$ as the soft threshold, which is additionally also the peak in the scale independence plot.

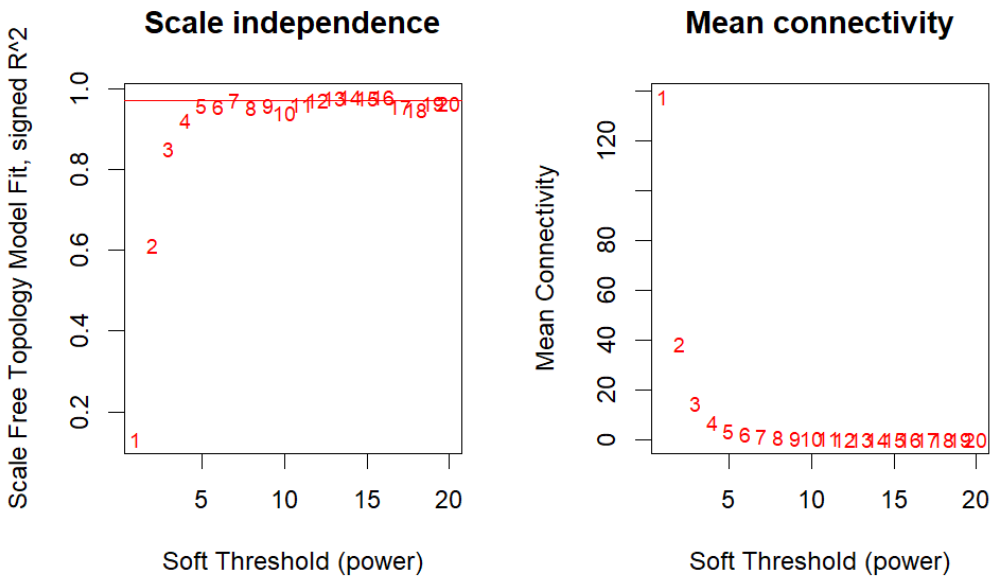

Figure 1. Scale independence plot and mean connectivity of WGCNA on paired data for the original (smaller) dataset

To exemplify the classic WGCNA, Figure 2 shows how the step of hierarchically clustering the genes, cutting them into modules and then merging similar modules look like. Characteristic for the smaller dataset is that the merging with a threshold of 0.25 actually does not change the modules.

Note again that the grey module is a module containing those genes for which it was not possible to assign them to a "real", meaningful module, and should therefore not be interpreted as having a biological meaning.

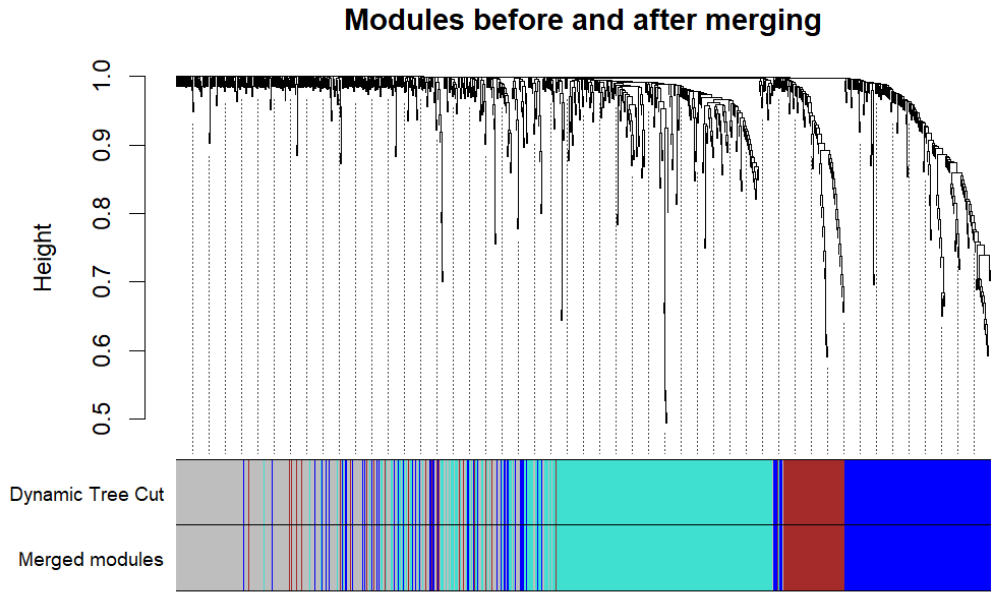

Figure 2. Dendrogram of the genes, the assigned module and the modules after the merging

The received modules have the following sizes:

Table 1. Module sizes from WGCNA on paired data (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
turquoise	294	35.4
grey	250	30.1
blue	201	24.2
brown	85	10.2

The linear mixed models, one fitted for each module, returned the following values for the coefficient of the tumour variable:

Table 2. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
turquoise	-0.21	**-19.28**	**6.87e-21**
grey	0.11	6.32	2.29e-07
blue	-0.01	-0.38	7.02e-01
brown	0.00	-0.11	9.15e-01

Remember that the grey module does not represent a real biological module. Therefore, the WGCNA method on paired data indicates that there is a very significant association between the turquoise module and the trait, with the expression values of the genes being lower in the tumour data compared to the normal data. This confirms the results of Li et al. The ten miRNAs in the turquoise with the highest intramodular connectivity are the following:

Table 3. The ten miRNAs in the turquoise module with the highest intramodular connectivity

miRNA	Module	kWithin	Module Membership (turquoise)	GS
ILMN_3167805	turquoise	7.48	0.867	11.228
ILMN_3167455	turquoise	7.08	0.887	12.730
ILMN_3166941	turquoise	6.53	0.858	9.727
ILMN_3167624	turquoise	6.31	0.845	9.378
ILMN_3167988	turquoise	6.22	0.843	10.057
ILMN_3167522	turquoise	6.22	0.826	9.131
ILMN_3168513	turquoise	6.04	0.881	14.642
ILMN_3168646	turquoise	5.83	-0.844	14.660
ILMN_3167031	turquoise	5.68	0.793	8.073
ILMN_3168716	turquoise	5.43	0.809	8.304

Note that this table gives a bit of a wrong impression. Gene significances go up to 16.4, and there are many genes in the module having a negative module membership. This is better displayed in the following plot:

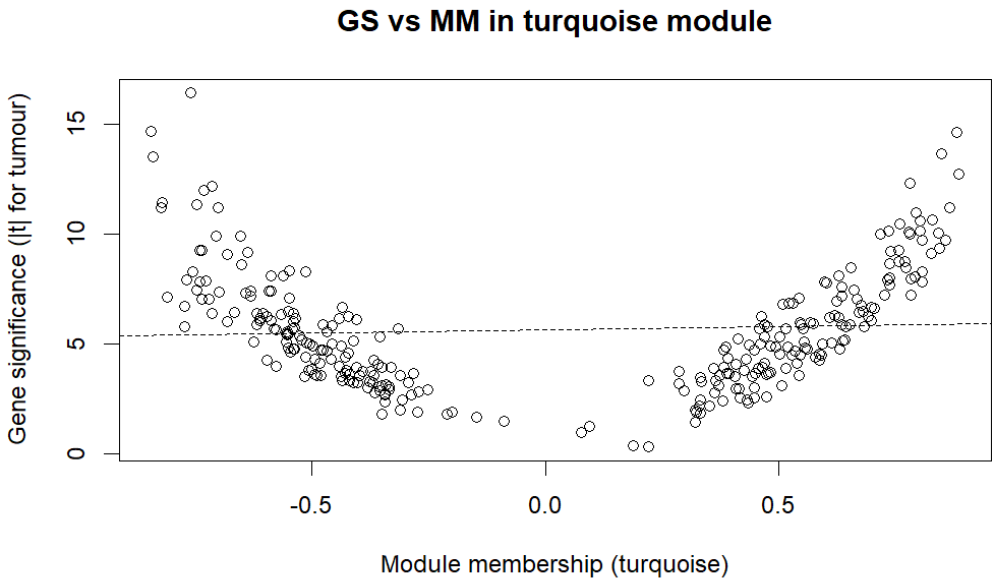

Figure 3. Gene significance vs. Module membership of elements in the turquoise module

It shows that the more meaningful genes are (de-) correlated with the module, the more association they have to the appearance of the tumour data, further indicating that the module-trait association is strong.

4.1.2 (ii) "Naive" WGCNA

As the first step of Naive WGCNA is exactly the same as for WGCNA on paired data, the soft threshold diagnostics were the same. Therefore, again $\beta = 7$ was chosen, and the modules were exactly the same as in the WGCNA on paired data method. This is also confirmed by the following table:

Table 4. Module sizes from Naive WGCNA (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
turquoise	294	35.4
grey	250	30.1
blue	201	24.2
brown	85	10.2

Also the naive Pearson correlation between the module eigengenes and the trait return a similar picture as the linear mixed model from before:

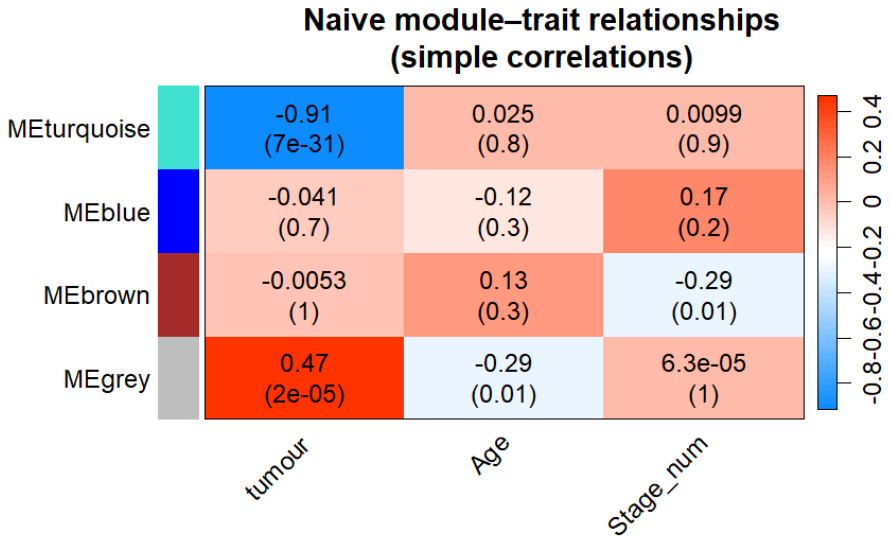

Figure 4. Correlations between the module eigengenes and the traits, most importantly the tumour variable on the left

Because the networks of the paired WGCNA and the naive WGCNA are the same, the kWithin values of the different genes in the modules will be congruent. Therefore, in the following, the five genes with the highest gene significance are compared (standard gene significance for the naive WGCNA and paired gene significance for the paired WGCNA):

Table 5. The five Genes with the highest gene significance in naive WGCNA (first five rows) and WGCNA on paired data (last five rows)

miRNA	Module	kWithin	Module Membership (turquoise)	GS (sGS/pGS)
ILMN_3168646	turquoise	5.83	-0.844	0.865
ILMN_3168513	turquoise	6.04	0.881	0.844
ILMN_3168273	turquoise	3.77	-0.761	0.839
ILMN_3168388	turquoise	4.84	-0.840	0.832
ILMN_3167455	turquoise	7.08	0.887	0.812
ILMN_3167805	turquoise	7.48	0.867	11.228
ILMN_3167455	turquoise	7.08	0.887	12.730
ILMN_3166941	turquoise	6.53	0.858	9.727
ILMN_3167624	turquoise	6.31	0.845	9.378
ILMN_3167988	turquoise	6.22	0.843	10.057

While there are some intersection points (e.g. ILMN_3167455), the different types of gene significance lead to greatly different influential genes. So, for the identification of hub genes, Li et al. approach shows advantages.
Note that any analysis not based on the gene significance, but on the adjacency scores, will return identical results in both methods:

Table 6. Overlap of the N hub genes with the highest kWithinNorm of paired WGCNA and naive WGCNA

Top N	Absolute Overlap	Relative Overlap
25	25	100%
50	50	100%
100	100	100%
500	500	100%

So, for this dataset, while naive and paired WGCNA return the same result for connectivity measures (by construction of the method), for the identification of hub genes, the results differ. Therefore, Li et al.'s method indeed surpasses the original method and offers different results.

4.1.3 (iii) WGCNA on the log ratios

Choosing the soft threshold was a bit delicate, as only for $\beta \geq 15$, the R^2 value was at 0.9 or above. It was thus decided to allow for R^2 to be 0.8 to have a smaller β. Finally, it was fixed that $\beta = 5$.

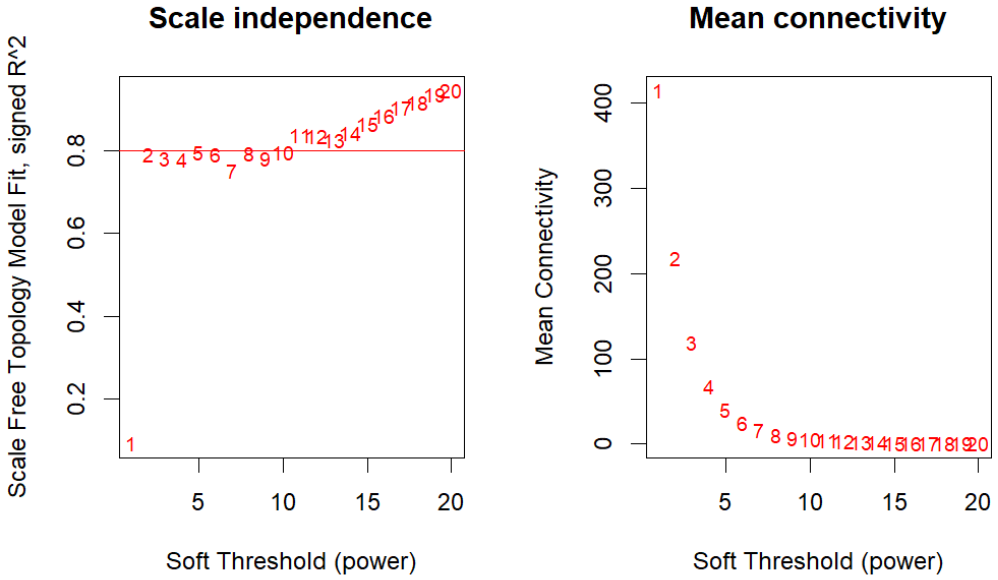

Figure 5. Scale independence plot and mean connectivity of WGCNA on log ratios for the original (smaller) dataset

The method resulted in only two modules. Maybe a more aggressive splitting of the dendrogram would have been a better choice (so, a larger `deepSplit`), but for comparability, `deepSplit = 2` was kept. Interestingly, no grey module was created:

Table 7. Module sizes from WGCNA on log ratios (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
turquoise	550	66.3
blue	280	33.7

For these two modules, the module-trait associations are now calculated using the above-explained method with the one-sample t-test:

Table 8. Module-trait associations based on the t-test method on the mean vector (original, small dataset)

Module	Sample Size	Mean of the Mean vector	t-value	p-value	FDR
blue	38	-0.189	-3.973	0.000315	0.000315
turquoise	38	0.096	3.973	0.000315	0.000315

Note that the sample size corresponds to the number of patients (without outliers). Also, while it seems like a bug that the p-values are exactly the same and the t-values added up are 0, this appears to be a result of the normalisation of the data applied after the log2-transformation. If not quantile normalisation, but a different one had been chosen, they would not have added up to 0. To keep preprocessing the same for all methods, this was accepted.
The table shows that both modules show a significant module-trait relationship. While the expression values of the turquoise module are higher in the tumour samples, the opposite is the case in the blue module (the ratio $\frac{\text{tumour}}{\text{normal}}$ was taken.
The following identification of the hub genes results in the following table (ordered by kWithin, top 5 genes per module):

Table 9. The five genes in both modules, each with the highest intramodular connectivity

miRNA	Module	kWithin	Module Membership	GS	p-value	p-value (adjusted)
ILMN_3167052	blue	31.9	0.800	-7.24	1.38e-08	1.68e-07
ILMN_3168757	blue	30.5	0.749	-5.80	1.16e-06	7.65e-06
ILMN_3166935	blue	30.5	0.759	-4.27	0.000132	0.000545
ILMN_3168866	blue	29.9	0.720	1.17	0.248	0.367
ILMN_3168815	blue	28.7	0.655	2.01	0.0518	0.102
ILMN_3168183	turquoise	82.1	0.955	-0.878	0.386	0.523
ILMN_3168373	turquoise	81.8	0.940	-2.25	0.0306	0.0659
ILMN_3168541	turquoise	81.2	0.963	0.554	0.583	0.697
ILMN_3168603	turquoise	80.8	0.958	-1.72	0.0931	0.167
ILMN_3167503	turquoise	80.2	0.950	-1.78	0.0825	0.151

Interesting to see here is that the turquoise module seems to be much more clustered. Also, genes with a very high intramodular connectivity in the turquoise module do not seem to have such a big gene significance, which is exactly contrary to the genes in the blue module. This makes sense as for both module to have the same t-value, the blue one with a bigger absolute mean vector needs to be spread further around.
When doing a Cartesian product with the modules of the WGCNA on paired data, the following heatmap is obtained:

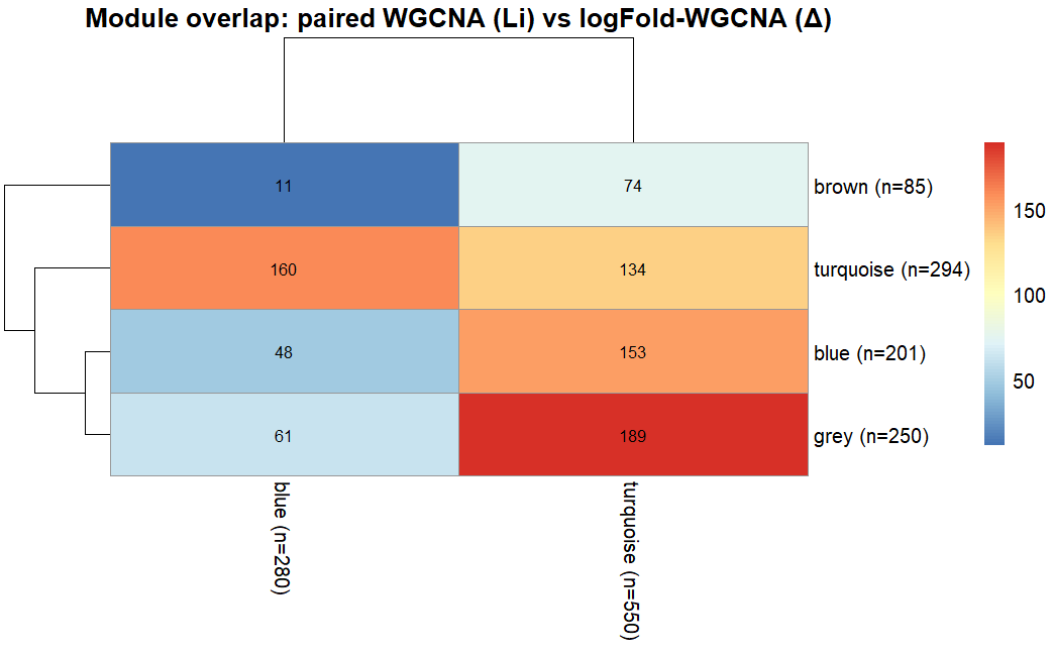

Figure 6. Cartesian product of the modules returned by paired WGCNA and WGCNA on log ratios

That the module assignment agreement is ARI = 0.03 confirms what can be guessed from the heatplot as well: there is not much of an agreement between the methods.
The Jaccard indices between the turquoise module of the paired WGCNA to the two modules of the log ratio WGCNA are $\text{Jaccard}_{\text{turquoise, blue}} = 0.386$ and $\text{Jaccard}_{\text{turquoise, turquoise}} = 0.189$. As the bigger Jaccard index is greater than 0.3, at least some correspondence is implied.
The TOM edge-weight agreement plot as well as the TOM-based agreement of the hub genes, are omitted as paired WGCNA uses the unsigned method while log ratio WGCNA needs the signed method, implying distorted results independent from the real biological relation.
At last, the overlap of the N highest hub genes should be examined. The resulting table, calculated as explained above, is the following:

Table 10. Overlap of the N hub genes with the highest kWithinNorm of paired WGCNA and log ratio WGCNA

Top N	Absolute Overlap	Relative Overlap
10	4	40 %
25	12	48 %
50	33	66 %
100	70	70 %
500	389	77.8 %

The resulting overlap is relatively low, indicating quite different results. This will be examined when comparing with the conventional ground truth.

4.1.4 (iv) Pair-aware WGCNA

As the scale independence plot first exceeded $R^2 = 0.9$ at $\beta = 5$, the soft threshold was fixed like this.
When executing the method, the following modules are returned (note that `minClusterSize` was still set to 30, but as the grey module is just the group of the genes that cannot be assigned, it may be smaller than this size):

Table 11. Module sizes from Pair-aware WGCNA (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
turquoise	554	66.7
blue	263	31.7
grey	13	1.6

Like for log ratios WGCNA, there are two big meaningful modules, one having around 2/3 of the genes and the other one having 1/3.
The Module-trait associations (calculated using the linear mixed model) look as follows:

Table 12. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
turquoise	-0.208	**-18.688**	**1.95e-20**
blue	-0.0116	-0.462	6.47e-01
grey	0.00177	0.0672	9.47e-01

Contrary to the log ratio WGCNA, only one of the two modules has a significant module-tumour association. Genes in the turquoise module have lower expression values in the tumour tissue. Interestingly, the mean of the blue module is also negative, though not significant.
When searching for the hub miRNA in the turquoise module, the following are returned:

Table 13. The ten miRNAs in the turquoise module with the highest intramodular connectivity

miRNA	Module	kWithin	Module Membership (turquoise)	GS
ILMN_3168513	turquoise	47.569	0.881	14.642
ILMN_3167455	turquoise	45.620	0.891	12.730
ILMN_3168646	turquoise	45.482	-0.834	14.660
ILMN_3167805	turquoise	45.278	0.871	11.228
ILMN_3168707	turquoise	44.896	0.785	12.304
ILMN_3168388	turquoise	43.239	-0.825	13.526
ILMN_3167729	turquoise	43.224	0.861	13.676
ILMN_3168273	turquoise	41.971	-0.751	16.411
ILMN_3166941	turquoise	41.554	0.846	9.727
ILMN_3168749	turquoise	40.876	-0.819	11.202

As both, the paired WGCNA and the pair-aware WGCNA have only one module with a significant module-trait association, the following table shows the size of the overlap between the top hub genes of the turquoise module of both methods:

Table 14. Overlap of the N hub genes with the highest intramodular connectivity in each of the turquoise module of paired WGCNA and pair-aware WGCNA

Top N	Absolute Overlap	Relative Overlap
10	5	50 %
25	20	80 %
50	38	76 %
100	86	86 %
200	173	86.5 %

So, while the modules seem to be very similar, the very top hub genes coincide only moderately. The following heatplot again shows how similar the modules are:

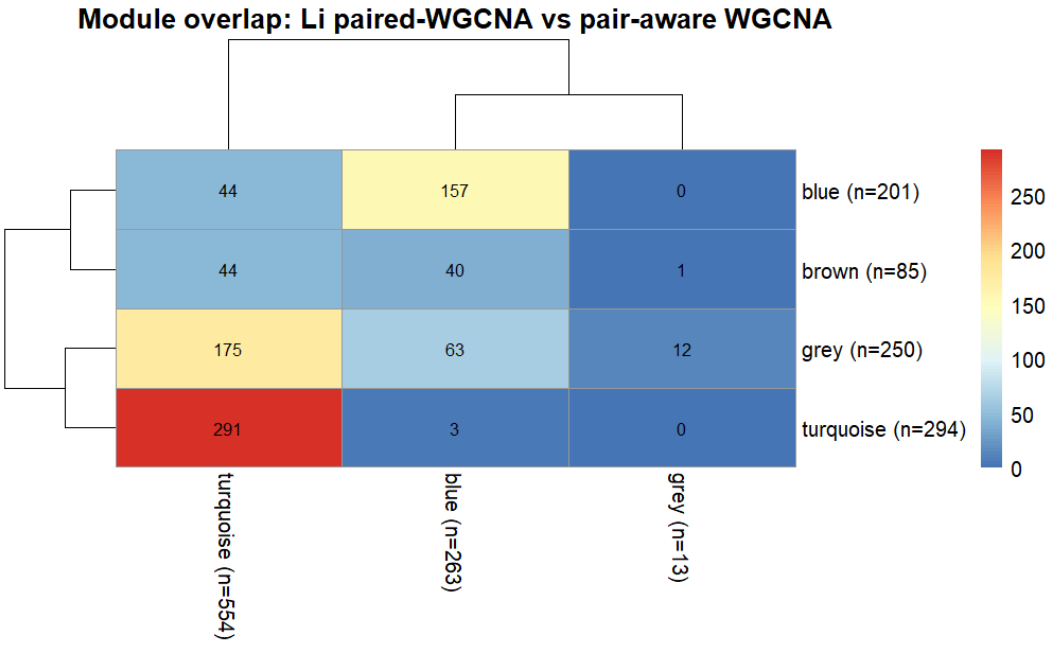

Figure 7. Cartesian product of the modules returned by paired WGCNA and pair-aware WGCNA

The agreement of the edges of the TOM values of both networks shows a similar picture:

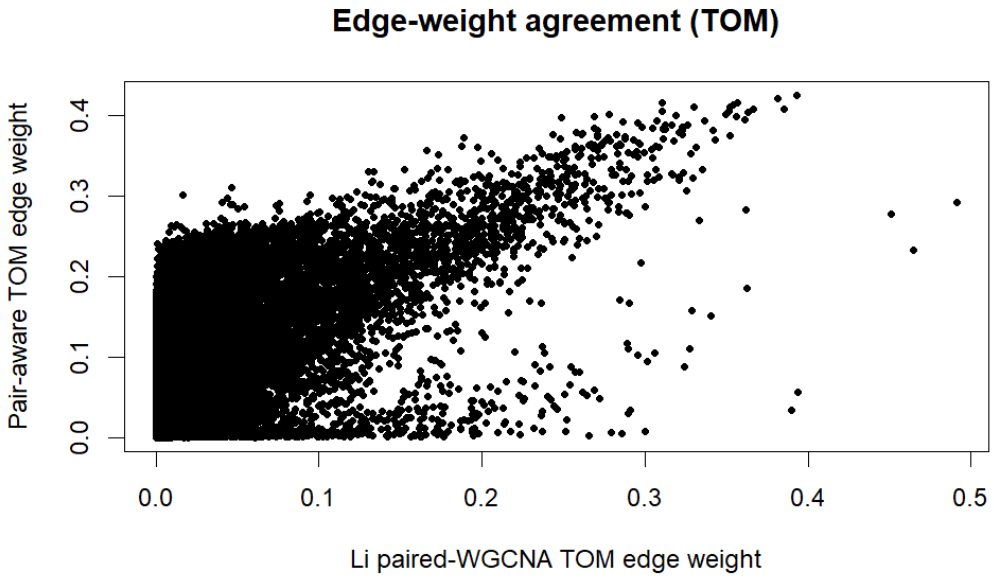

Figure 8. Edge-weight agreement plot of the networks created by the paired WGCNA and pair-aware WGCNA

A Spearman's rank correlation test on the possible correlation of these edge weights yields a p-value $< 2.2e-16$, indicating strong correlation.
The Jaccard index between the two turquoise modules is 0.5224417. This result, as bigger than 0.5, implies that the modules are nearly the same, confirming the previous knowledge once more.
We can thus conclude that the pair-aware WGCNA, very similarly to paired WGCNA returns one significant module, which is a subset of the significant module of paired WGCNA, consisting of its most important genes. The order of the hub genes is mostly consistent, though the very top hub genes differ slightly.

4.1.5 (v) Separate WGCNA networks and module preservation

In the appendix.

4.1.6 (vi) Limma

Using a paired Limma model, the differential miRNA expression of tumour vs normal was studied. The method yielded 366 significant miRNAs ($FDR < 0.05$). Limiting additionally to $|\text{logFC}| > 1$, Limma identified 80 miRNAs, of which 34 are upregulated, and 46 are downregulated. These thresholds are more closely shown in the volcano plot in figure 9, where the purple coloured points represent the miRNA which pass them.
Figure 9 also includes the MDS plot. It clearly shows that normal and tumour samples form two distinct clusters, which means that the condition is a strong driver for variation.
Additionally, table 15 shows the 6 most significant miRNAs by Limma.

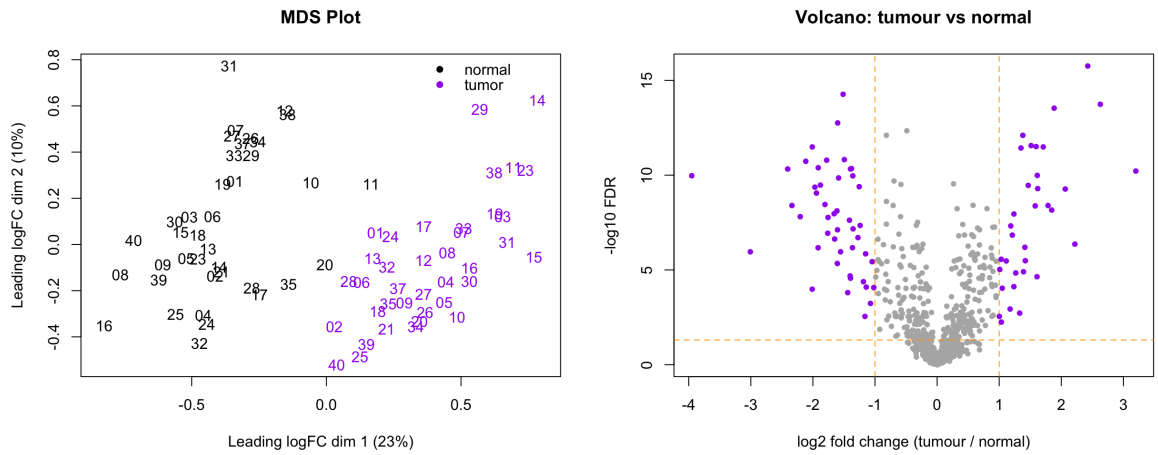

Figure 9. Results Limma for first dataset

Table 15. Top miRNAs by Limma

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
ILMN_3168513	-1.51	11.77	-14.81	1.30×10^{-17}	5.40×10^{-15}	30.04
ILMN_3167455	-1.60	11.94	-12.94	1.06×10^{-15}	1.75×10^{-13}	25.67
ILMN_3167729	-0.49	13.49	-12.49	3.27×10^{-15}	4.53×10^{-13}	24.55
ILMN_3168707	-0.82	6.37	-12.18	7.09×10^{-15}	7.90×10^{-13}	23.77
ILMN_3167805	-2.01	9.02	-11.47	4.45×10^{-14}	3.25×10^{-12}	21.94
ILMN_3167818	-1.49	8.32	-10.82	2.55×10^{-13}	1.51×10^{-11}	20.20

When comparing Limma to the original paired data WGCNA pipeline, the first question asked was: Do the miRNAs deemed important by both methods overlap? Limma already gives an ordered list (by adj.p.val) of important miRNAs, for paired WGCNA, a wgcna-score was defined as follows: $S_{wgcna}(i) = GS_i * |MM_i|$. GS_i defines the gene-significance of each gene and MM_i defines the module membership. In table 16, the number of miRNAs which are in the top N most important miRNAs of both methods can be seen. Enrichment beyond chance was assessed using a one-sided Fisher's exact test. From the table, it can be clearly seen, that the results are very strong and highly significant. So, there is a clear agreement between the top-N miRNAs given by these two methods.

Table 16. Fisher's exact test results for overlap among top-N miRNAs between paired WGCNA and (iv) Limma

N	Overlap	p-value	Odds ratio	95% CI (lower)
200	184	$< 2.2 \times 10^{-16}$	433.87	229.30
150	137	$< 2.2 \times 10^{-16}$	517.68	257.83
100	89	$< 2.2 \times 10^{-16}$	507.70	233.80
50	44	$< 2.2 \times 10^{-16}$	894.16	304.98

Next, it was examined which paired WGCNA modules contained the most miRNAs flagged by Limma. These results are seen in table 17. These values show, that for the turquoise module, around 90% of the miRNAs are DE. As the turquoise module has been shown to be the most associated with the phenotype by the corresponding LMM, these results show a correlation between DE and phenotype association. The remaining modules show very high FDR values, and so those results are not significant.

Table 17. DE miRNAs by Limma in the paired WGCNA modules

Module	Module size	# DE miRNAs in module	Fraction DE	FDR
turquoise	294	274	0.932	8.05×10^{-110}
brown	85	22	0.259	1.000
blue	201	36	0.179	1.000
grey	250	34	0.136	1.000

The correlations seen in table 18 indicate a moderate positive association between Limma evidence for differential expression and WGCNA-derived measures of intramodular hubness (kWithin) and module membership. The very strong correlation for the combined score $GS*|MM|$ is expected, since GS already encodes phenotype association and therefore aligns closely with Limma's differential expression signal.

Table 18. Spearman rank correlations between limma significance and paired-WGCNA membership measures

Comparison	Spearman ρ	p-value
$kWithin$ vs. limma significance	0.413	$< 2.2 \times 10^{-16}$
$\|MM\|$ vs. limma significance	0.382	$< 2.2 \times 10^{-16}$
$GS \cdot \|MM\|$ vs. limma significance	0.915	$< 2.2 \times 10^{-16}$

4.1.7 (vii) DiffCoExp

DiffCoExp was applied to the first dataset by estimating separate correlation matrices and clustering miRNAs based on differences in coexpression. The method was run using Spearman correlations with BH adjustment and default thresholds of $|r| \geq 0.5$, correlation $q \leq 0.1$, $|\Delta r| \geq 0.5$, differential-correlation $q \leq 0.1$, and DCG $q \leq 0.1$.
440 differentially coexpressed links (DCLs) were identified (see table 19), of which 101 were same signed (correlation sign remains the same between conditions), 335 were different signed (correlation sign differs between conditions) and 4 were switched opposites (correlation sign reverses between conditions, i.e., a positive correlation in one condition becomes negative in the other and vice versa). The distribution of correlation changes indicated that the strongest differential links had correlation differences up to $|\Delta r| = 1.13$, with corresponding adjusted p-values as low as $q = 9.18 \times 10^{-6}$.
Furthermore, 28 differentially coexpressed miRNAs (DCGs) were identified (see table 20). The top-ranked DCGs included ILMN_3168167, ILMN_3168700, and ILMN_3167182, with 22, 20, and 26 associated DCLs, respectively. These DCGs represent candidate miRNAs with extensive rewiring of coexpression relationships between tumour and normal tissue. In table 20, the six most significant DCGs are shown. CLs refers to how many coexpression links this gene has been a part of, and DCLs refers to how many of those pass the differential coexpression threshold.

Table 19. Top differentially coexpressed links (DCLs) from DiffCoExp

Gene.1	Gene.2	cor_normal	cor_tumour	cor_diff	q_{diff}	type
ILMN_3168717	ILMN_3167363	0.33	0.96	0.63	9.18×10^{-6}	same signed
ILMN_3167609	ILMN_3167363	0.33	0.95	0.62	2.78×10^{-5}	same signed
ILMN_3168009	ILMN_3167363	0.35	0.95	0.60	1.26×10^{-4}	same signed
ILMN_3168656	ILMN_3167807	0.91	0.07	-0.83	1.94×10^{-4}	same signed
ILMN_3168663	ILMN_3166974	0.95	0.39	-0.56	3.60×10^{-4}	same signed
ILMN_3168808	ILMN_3167363	0.41	0.95	0.54	3.60×10^{-4}	same signed

Table 20. Top differentially coexpressed miRNAs (DCGs) from DiffCoExp

Gene (probe)	CLs	DCLs	q (FDR)
ILMN_3168167	92	22	3.25×10^{-15}
ILMN_3168700	81	20	2.68×10^{-14}
ILMN_3167182	179	26	4.59×10^{-13}
ILMN_3166957	78	15	4.81×10^{-9}
ILMN_3167807	167	16	2.86×10^{-5}
ILMN_3167148	155	15	5.02×10^{-5}

For the comparison between DiffCoExp and paired WGCNA, the first question was how the top candidates overlap. For DiffCoExp, the DCL degree is defined as the number of DCLs a gene participates in, and this measure was compared to the intramodular connectivity $kWithin$ from paired WGCNA. A Spearman rank test yields a correlation of around $\rho \approx 0.09$ with $p = 0.11 > 0.05$. Therefore, there is no statistically significant relationship between the kWithin and the DCL degree of a gene. In plot 10, the two values are plotted against each other.

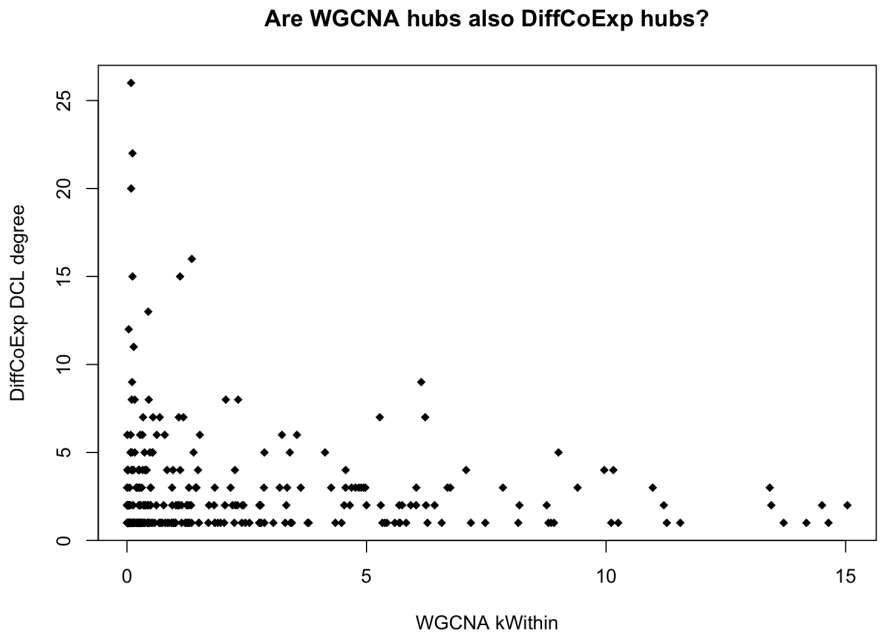

Figure 10. DGL degree vs kWithin

Applying an LMM on paired WGCNA showed that the turquoise module showed the most tumour association. 68.2% of DCLs had at least one endpoint in the turquoise module, while only 22.7% were within it completely. Taken together, these results suggest that tumour-associated rewiring is centred on the turquoise module, but a substantial portion of that signal corresponds to altered interactions between turquoise and other modules.

Lastly, to check whether the DiffCoExp DCL pairs are also connected in the paired-WGCNA network, the WGCNA TOM weights for each DCL edge were compared to the TOM weights of all miRNA pairs. The TOM values for DCL edges were clearly higher than the background distribution. The median TOM was 6.14×10^{-4} for DCL edges compared to 5.45×10^{-5} for all pairs (about 11 times higher). This suggests that many correlation changes occur on pairs that already show above-average co-expression connectivity in the WGCNA network.

4.1.8 (viii) Graphical Lasso

Sparse miRNA association networks were inferred on the first dataset, separately for normal and tumour samples, using Graphical Lasso (Glasso) with StARS-based regularisation selection. This yielded a network with 15'460 edges and a density of 0.0449 for normal samples, where edges refer to a non-zero value between two miRNAs in the network. For tumour samples the result shows slightly more edges, 15'832, and a slightly higher density, 0.046.
Comparing the StARS-selected Glasso networks, 13'996 edges were gained, and 13'624 edges were lost in tumour relative to normal, indicating extensive rewiring of conditional miRNA associations, with only a small net increase in connectivity (+372 edges) in tumour.
Table 21 shows the six top miRNAs based on the degree gained from normal to tumour. The degree is the number of edges touching a given miRNA.

Table 21. Top degree miRNAs from Graphical Lasso

miRNA	degree (N)	degree (T)	strength (N)	strength (T)	Δdegree	Δstrength
ILMN_3167452	28	54	1.32	1.60	26	0.28
ILMN_3168788	17	41	1.01	1.39	24	0.38
ILMN_3168097	32	55	1.29	1.63	23	0.35
ILMN_3168701	24	47	0.85	1.41	23	0.56
ILMN_3168706	20	43	1.07	1.47	23	0.40
ILMN_3167509	15	37	0.85	1.29	22	0.45

Similarly to the comparison for DiffCoExp, the first thing examined when comparing Glasso is whether the edges defined by Glasso fall into the paired WGCNA modules. The results showed that in both conditions, within-module edges were

strongly enriched relative to a normal edge baseline (see table 22). A Fisher test with a significant p-value confirmed that the within-module fraction was slightly higher in normal than in tumour, indicating a small shift towards more between-module edges in tumour.

Table 22. Within-module enrichment of Glasso edges with respect to paired-WGCNA modules

condition	total_edges	within_edges	within_fraction	expected_fraction	enrichment	p_value
normal	15460	6669	0.43	0.28	1.52	$< 10^{-300}$
tumour	15832	6634	0.42	0.28	1.47	3.94e-301

Table 23 details which edges from which modules are linked. The within module edges are represented along the diagonal. One can clearly see, that the within-module edges are pronounced while the module with the most between-module connections is the turquoise one. These results hold for both the normal and the tumour network.

Table 23. Glasso edges between paired-WGCNA modules

	Normal				Tumour			
Module pair	blue	brown	grey	turquoise	blue	brown	grey	turquoise
blue	1760	436	1389	2149	1680	538	1595	2131
brown		655	711	857		609	709	876
grey			1675	3249			1655	3349
turquoise				2579				2690

Glasso hubness, defined as node degree in the sparse conditional-dependence network, showed moderate agreement with paired WGCNA hubness, with Spearman correlations of 0.421 in normal and 0.415 in tumour. Consistently, the overlap between the top 100 hubs from Glasso and the top 100 WGCNA hubs was 38 miRNAs in both conditions. Overall, this indicates that Glasso and paired WGCNA identify partly overlapping but not identical sets of hub miRNAs, and the level of concordance is similar in normal and tumour.

4.2 Analysis on the second Dataset GSE62043

In the following, whenever possible the above explained function `blockwiseModules()` was used with `minModuleSize = 70` and `maxBlockSize = 12000`.

4.2.1 WGCNA on paired data

After consulting the Scale independence plot, the soft threshold was set to $\beta = 9$ as this is the first value for which $R^2 \geq 0.9$.
Using `blockwiseModules()`, the following modules are obtained:

Table 24. Module sizes (number of genes) from the paired WGCNA network

Module	Number of Genes	Amount of Genes (% of total)
grey	9247	47.19
turquoise	3371	17.20
blue	1648	8.41
brown	1312	6.70
yellow	1215	6.20
green	1029	5.25
red	493	2.52
black	323	1.65
pink	198	1.01
magenta	172	0.88
purple	160	0.82
greenyellow	131	0.67
tan	104	0.53
salmon	97	0.50
cyan	95	0.48

Almost half of the genes were not successfully assigned. Nevertheless, because of the great number of genes, many other reasonably sized modules are left for further analysis. In the following, the grey module is dropped from processing.

Applying a linear mixed model results in the following Module tumour associations:

Table 25. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
green	0.114	**17.9**	**8.85e-32**
turquoise	-0.113	**-16.7**	**1.32e-29**
salmon	0.0994	**15.3**	**5.66e-27**
brown	-0.0987	**-12.5**	**1.66e-21**
tan	-0.0829	**-12.0**	**1.51e-20**
cyan	0.0617	**8.26**	**1.04e-12**
pink	0.0697	**7.81**	**8.82e-12**
magenta	0.0178	**5.79**	**9.77e-08**
red	-0.0416	**-5.16**	**1.44e-06**
purple	0.0429	**4.85**	**5.00e-06**
greenyellow	0.0382	**4.25**	**5.20e-05**
black	0.00544	1.56	0.123
yellow	0.0133	1.39	0.167
blue	0.00392	1.03	0.305

The resulting heatmap looks like:

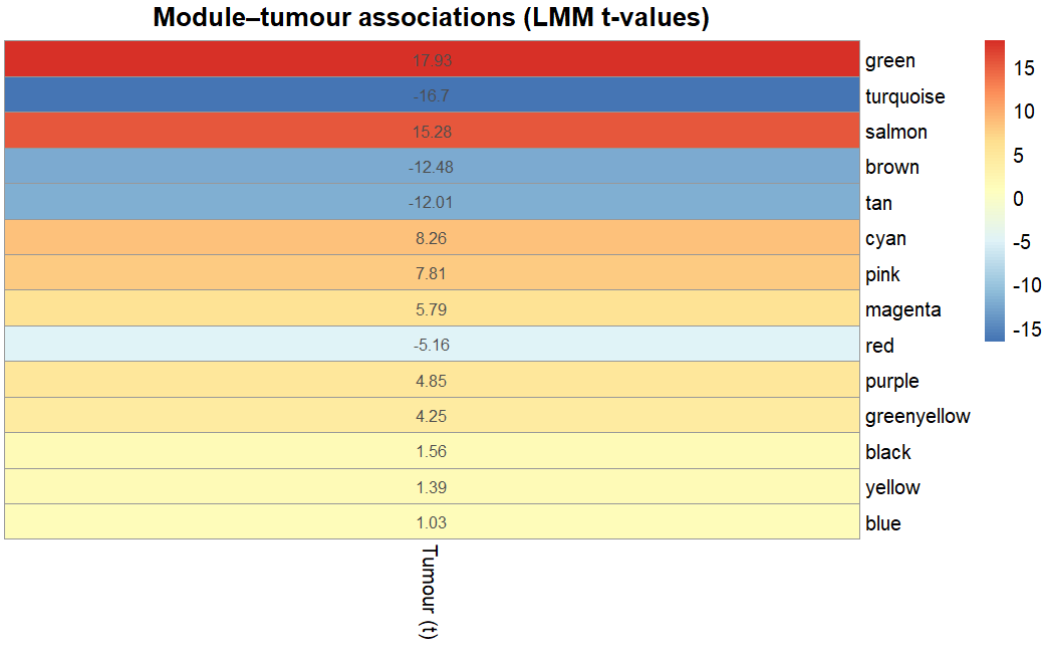

Figure 11. Heatmap of the modules corresponding to the t-value of the tumour-coefficient of the LMM

Contrary to the smaller dataset, there are now also significant modules whose genes seem to be more expressed in the tumour tissue.
From the three most significantly correlated modules (green, turquoise and salmon), now the five hub genes with the highest kWithin values each are retrieved:

Table 26. Top 5 genes (by kWithin) in green, turquoise, and salmon modules (ordered by kWithin)

Gene	Module	kWithin	Module membership (MM)	Gene significance (GS)
GMFG	turquoise	82.849	0.930	14.057
LCP2	turquoise	82.602	0.904	13.561
ARHGDIB	turquoise	82.082	0.916	11.556
CD53	turquoise	79.640	0.865	9.196
RCSD1	turquoise	79.465	0.906	11.444
SPC25	green	55.214	0.948	16.699
BUB1	green	54.136	0.948	14.231
MELK	green	53.957	0.941	15.936
CDCA5	green	52.737	0.943	17.437
NUF2	green	52.158	0.932	15.543
SCAMP3	salmon	7.472	0.936	16.279
SNX27	salmon	5.832	0.920	14.665
SCNM1	salmon	5.430	0.910	13.348
PRCC	salmon	4.748	0.897	11.012
VPS72	salmon	4.570	0.895	15.120

Note that because the turquoise (3371 genes) is much bigger than the green (1029 genes) and the salmon (97 genes) modules, it is not surprising to see that the kWithin values of the genes in the turquoise modules are much better. Related to their size, the salmon module shows the highest kWithinNorm values. So, it appears that the salmon module represents a precise, small biological module with a strong correlation to the presence of tumour tissue.
The turquoise module, still strongly anticorrelated with the trait, is the one with the most genes (except for the grey one). The Gene significance-Module Membership of it passes the sanity check:

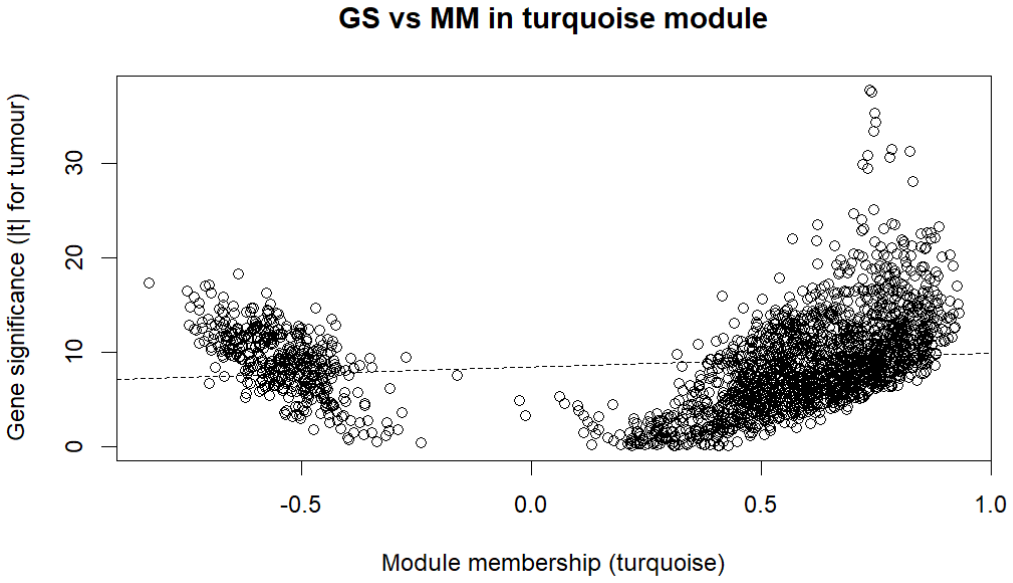

Figure 12. Gene significance vs. Module membership of elements in the turquoise module

Note that as the total number of genes is huge, only the gene significance for the 8000 genes with the highest variance across samples has been calculated. Genes in the turquoise not having such a big variance were dropped. This also introduces some bias to this plot, it should therefore be considered with a grain of salt.
Nevertheless, paired WGCNA here worked well to spot modules of genes having a significant association with the presence of tumour tissue.

4.2.2 (ii) "Naive" WGCNA

In the processing of the data, `blockwiseModules()` is used. Note that this function induces a bit of randomness, as does `pickSoftThreshold()` as well. To make the networks as similar as possible, the softpower of the naive WGCNA

is manually set to $\beta = 9$ as well. The scale independence plot shows that the choice is not bad, also for the data itself:

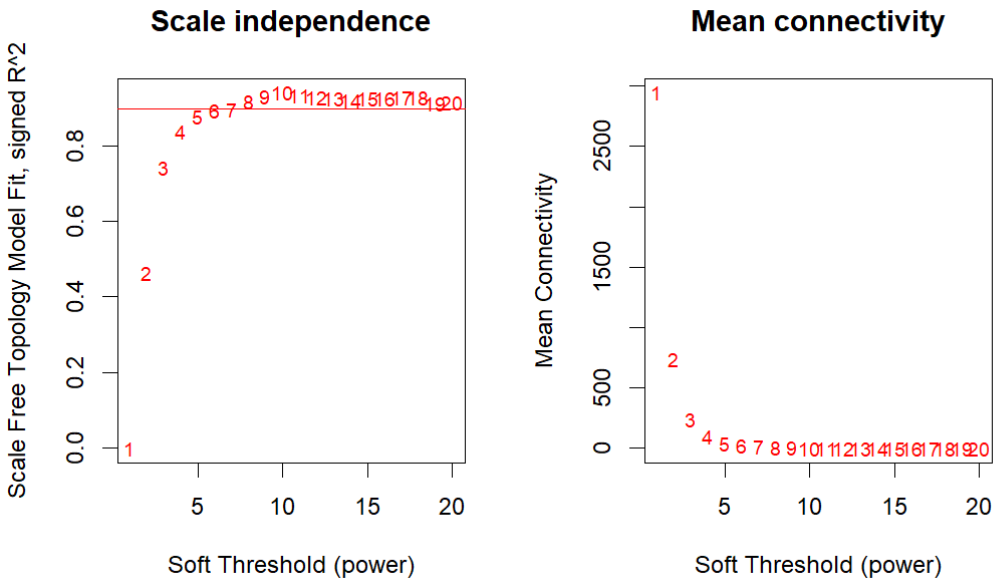

Figure 13. Scale independence plot and mean connectivity of naive WGCNA for the bigger dataset

Then the naive method results in the following modules:

Table 27. Module sizes (number of genes) from the naive WGCNA network

Module	Number of Genes	Amount of Genes (% of total)
grey	9247	47.19
turquoise	3371	17.20
blue	1648	8.41
brown	1312	6.70
yellow	1215	6.20
green	1029	5.25
red	493	2.52
black	323	1.65
pink	198	1.01
magenta	172	0.88
purple	160	0.82
greenyellow	131	0.67
tan	104	0.53
salmon	97	0.50
cyan	95	0.48

So, exactly the same modules (as wanted) were obtained. This is more strongly shown in this heatmap:

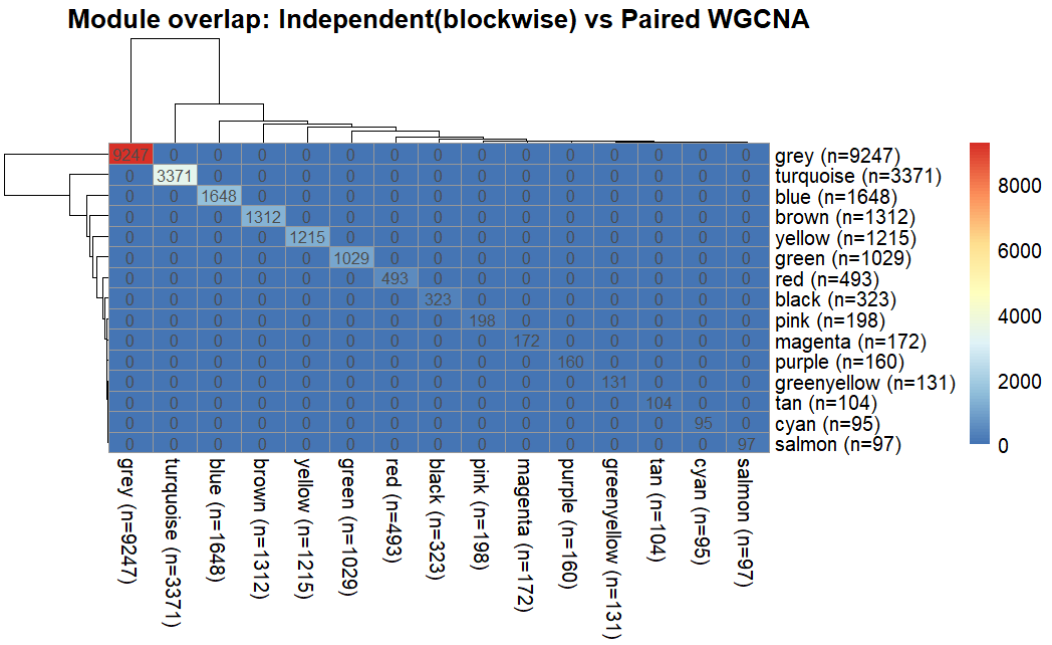

Figure 14. Heatmap of the cartesian product of the modules between naive WGCNA and paired WGCNA

The modules are thus identical.
The heatplot of the correlation between the module eigengenes and the trait is very similar to the resulting t-values from the linear mixed model before:

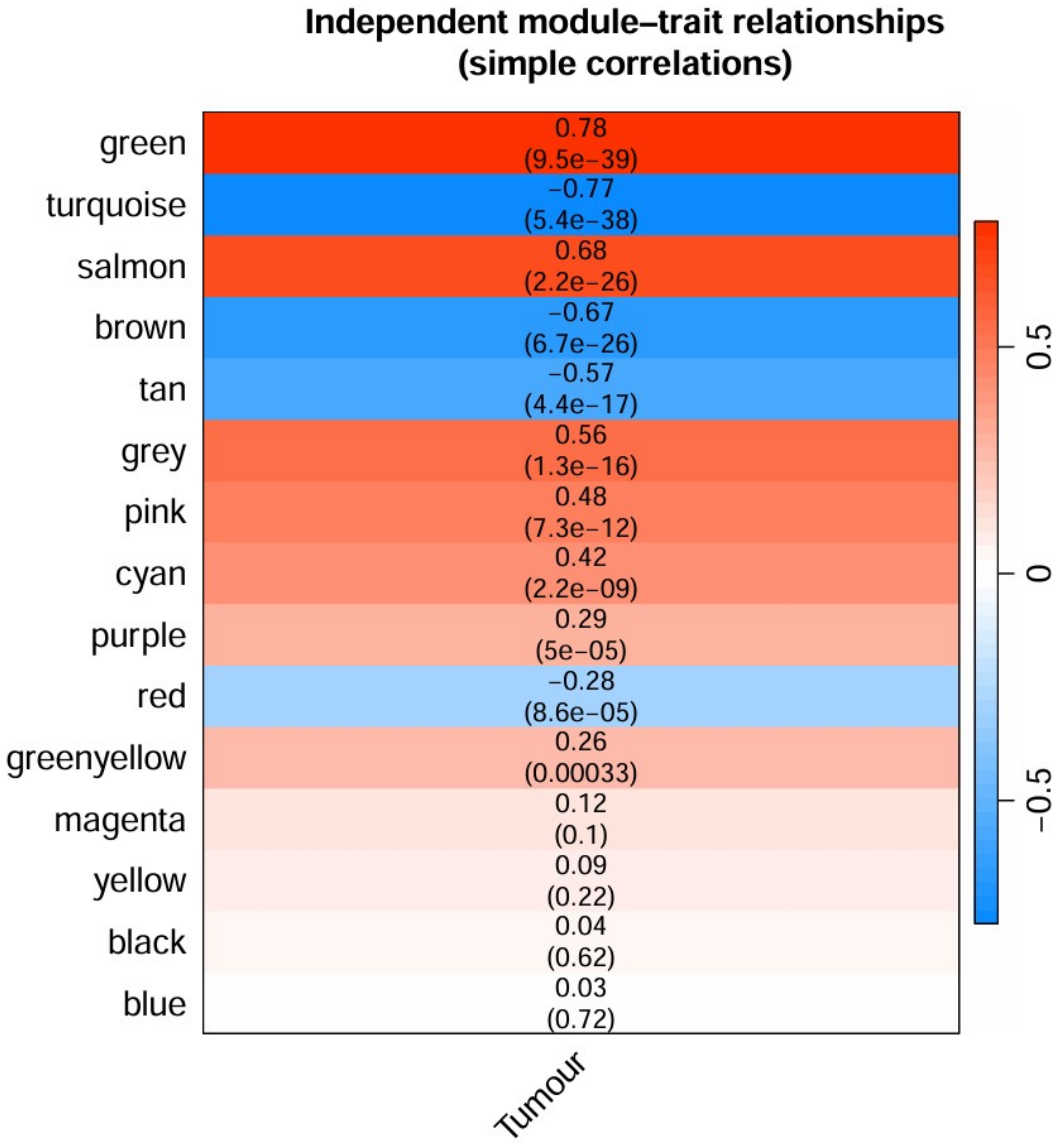

Figure 15. Heatmap of the correlation values (with p-values) of the bigger dataset

Note that here, the grey module was still included to check its correlation.
Indeed, the order of the significance is mostly preserved, with the five most associated modules being exactly the same. To check how similar the networks are, the kWithinNorm values of paired WGCNA and naive WGCNA are tested by the Spearman's rank correlation test. The returned sample estimated $\rho = 1$ implies identical networks, which is also confirmed by the following table:

Table 28. Overlap of the N hub genes with the highest kWithinNorm of paired WGCNA and naive WGCNA

Top N	Absolute Overlap	Relative Overlap
25	25	100%
100	100	100%
500	500	100%
2000	2000	100%

Because the networks of the paired WGCNA and the naive WGCNA are thus again exactly the same, the kWithin values of the different genes in the modules will be congruent. Therefore, again, the five genes with the highest gene significance are compared (standard gene significance for the naive WGCNA and paired gene significance for the paired WGCNA) for each of the three modules, green, turquoise and salmon:

Table 29. Top 5 genes by gene significance (GS) in selected modules: naive vs paired WGCNA

Module	Rank	Gene (naive)	GS (naive)	Gene (paired)	GS (paired)
green	1	CENPW	0.8075	PTTG1	23.24
green	2	CCNB1	0.8005	CDKN3	21.28
green	3	BIRC5	0.7987	BIRC5	20.71
green	4	UBE2T	0.7973	CCNB1	20.12
green	5	RACGAP1	0.7935	SAC3D1	19.98
turquoise	1	OIT3	0.9184	CLEC4G	37.74
turquoise	2	CLEC4G	0.9075	MARCO	37.47
turquoise	3	CRHBP	0.9018	FCN3	35.28
turquoise	4	CXCL14	0.9018	CRHBP	34.33
turquoise	5	ECM1	0.8996	STAB2	33.33
salmon	1	CCT3	0.7732	CCT3	19.07
salmon	2	FAM189B	0.7444	FAM189B	18.58
salmon	3	EFNA4	0.7113	GBA	17.79
salmon	4	GBA	0.7097	KRTCAP2	17.76
salmon	5	SCAMP3	0.7045	SCAMP3	16.28

So, while there are some genes appearing in both methods in the top 5 genes ordered by gene significance, there is still some discrepancy, implying that paired WGCNA really is needed when working with paired data.
Naive WGCNA and paired WGCNA return very similar results, but when it comes to individual gene significance, the highlighted genes differ.

4.2.3 (iii) WGCNA on the log ratios

The soft threshold was set to $\beta = 9$, the first value that led to $R^2 \geq 0.9$. Note that this is the same soft threshold as used for naive WGCNA and paired WGCNA.
The method resulted in the following modules:

Table 30. Module sizes (number of genes) from the log ratio WGCNA network

Module	Number of Genes	Amount of Genes (% of total)
grey	6282	32.06
turquoise	3099	15.82
blue	2136	10.90
brown	1974	10.07
yellow	1366	6.97
green	1337	6.82
red	1061	5.41
black	516	2.63
pink	483	2.46
magenta	247	1.26
purple	226	1.15
greenyellow	218	1.11
tan	198	1.01
salmon	130	0.66
cyan	112	0.57
midnightblue	111	0.57
lightcyan	99	0.51

Again, the module-trait associations are calculated using the one-sample t-test for each of the modules:

Table 31. Module-trait associations based on the t-test method on the mean vector (GSE62043, Δ-expression)

Module	Sample Size	Mean of the Mean vector	t-value	p-value	FDR
lightcyan	93	-0.771	-21.0	6.62e-37	1.06e-35
yellow	93	0.397	17.3	1.12e-30	8.93e-30
red	93	-0.266	-11.2	6.17e-19	3.29e-18
pink	93	-0.296	-10.9	3.36e-18	1.35e-17
magenta	93	0.394	10.6	1.10e-17	3.53e-17
tan	93	-0.482	-9.45	3.30e-15	8.80e-15
grey	93	0.031	8.22	1.25e-12	3.026e-12
salmon	93	0.443	8.16	1.67e-12	3.82e-12
cyan	93	-0.305	-8.07	2.54e-12	5.07e-12
brown	93	-0.301	-7.94	4.72e-12	8.39e-12
turquoise	93	-0.250	-7.33	8.65e-11	1.38e-10
greenyellow	93	0.180	5.50	3.32e-07	4.83e-07
black	93	0.169	5.18	1.33e-06	1.77e-06
purple	93	0.112	4.16	7.07e-05	8.70e-05
green	93	-0.0490	-3.82	2.41e-04	2.75e-04
midnightblue	93	0.0566	2.94	4.14e-03	4.41e-03
blue	93	0.0296	1.03	3.05e-01	3.05e-01

Surprisingly, all but one module offer significant module-trait associations. This result is probably also related to the increased sample size.
For the two modules with the smallest p-values by a wide margin, we identify the hub genes (ordered by kWithin, top 5 genes per module):

Table 32. The five genes in each module with the highest intramodular connectivity (lightcyan and yellow; GS from one-sample t-test on Δ-expression vs 0).

Gene	Module	kWithin	Module Membership	GS	p-value	p-value (adjusted)
SIGLEC7	lightcyan	18.950	0.902	-14.571	1.25e-25	7.32e-24
LILRA1	lightcyan	17.886	0.886	-14.323	3.75e-25	1.94e-23
LILRA4	lightcyan	17.492	0.869	-11.566	1.24e-19	2.31e-18
LILRP2	lightcyan	16.187	0.847	-11.334	3.75e-19	6.44e-18
SIGLEC11	lightcyan	16.074	0.858	-18.054	5.42e-32	1.1e-29
BUB1	yellow	113.733	0.882	14.230	5.67e-25	2.86e-23
KIF18B	yellow	112.942	0.888	14.004	1.56e-24	7.2e-23
C15orf42	yellow	109.244	0.839	12.846	3.08e-22	9.29e-21
RAD54L	yellow	108.159	0.860	14.481	1.86e-25	1.05e-23
EXO1	yellow	108.092	0.856	15.118	1.13e-26	8.14e-25

Note that the lightcyan is around 14 times smaller than yellow. This explains the differences in the values of kWithin. Nevertheless, it's impressive for a module of the size of yellow to have genes with such a high kWithin value, implying very strong connections to the rest of the module. Also, all of these ten genes exhibit very big gene significance values, indicating strong associations to the appearance of the trait (with the yellow genes having higher expression values in the tumour tissue, whereas the lightcyan are expressed less when the trait is present).
When doing a Cartesian product with the modules of the WGCNA on paired data, the following heatmap is obtained:

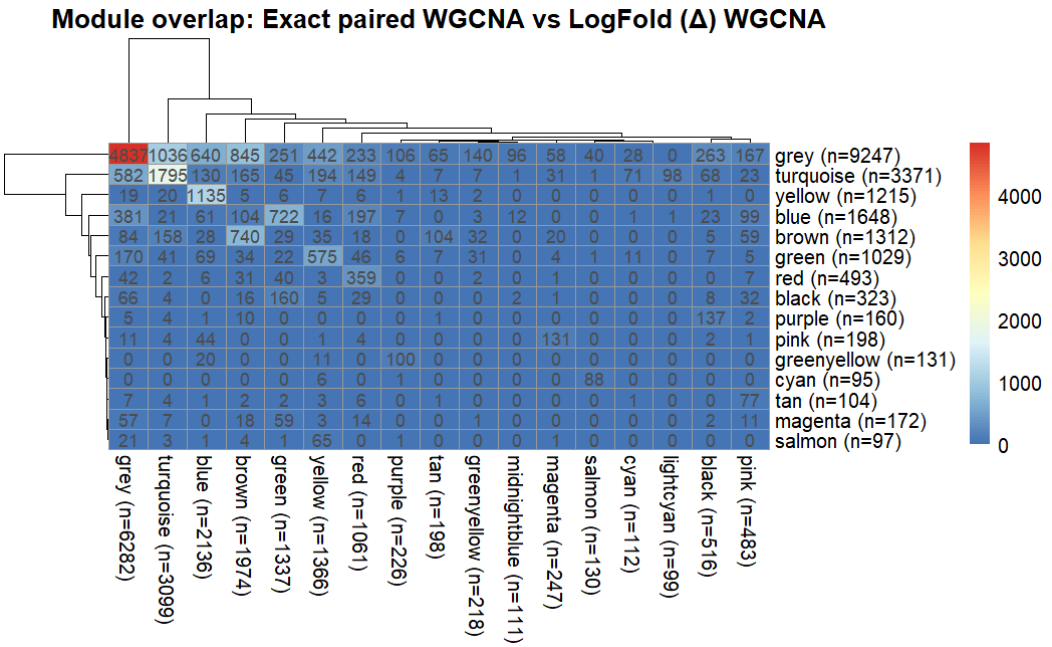

Figure 16. Cartesian product of the modules returned by paired WGCNA and WGCNA on log ratios

Remember that the grey modules are just the modules of unassigned genes, so it's not a big surprise that there is a rather big overlap, especially as the two modules are the largest in both methods. Besides, it seems like the turquoise modules are rather similar, the yellow paired modules seems to be represented by the blue log ratio module, the blue paired modules seems to share many genes with the green log ratio module, and the brown modules seem to be similar. There are two further associations (green <-> yellow and red -> red). It thus becomes visible that some of the structure of some modules is preserved over the methods.
The module assignment agreement is ARI = 0.238, implying some moderate agreement.
Again, because the methods of creating the adjacency scores are not the same, the edge-weight agreement plot and the TOM-based agreement of the hub genes do not make sense to be created.
Finally, the overlap of the N highest hub genes is calculated, based on the kWithinNorm values, yielding the following table:

Table 33. Overlap of the N hub genes with the highest kWithinNorm of paired WGCNA and log ratio WGCNA

Top N	Absolute overlap	Relative overlap
25	0	0 %
100	27	27 %
500	195	39 %
2000	1045	52.2 %

The overlap of the hub genes of the two methods is again quite low; especially when only considering the very top hub genes, the differences are striking.

4.2.4 (iv) Pair-aware WGCNA

The soft threshold was chosen to be $\beta = 7$, again the first value for which the scale independence plot exceeded $R^2 = 0.9$.
The following modules were obtained using pair-aware WGCNA:

Table 34. Module sizes from Pair-aware WGCNA (large dataset)

Module	Number of genes	Relative amount (%)
turquoise	5414	27.6
grey	3489	17.8
blue	2749	14.0
brown	1967	10.0
yellow	1263	6.4
green	1129	5.8
red	1113	5.7
black	627	3.2
pink	447	2.3
magenta	314	1.6
purple	249	1.3
greenyellow	225	1.1
tan	177	0.9
salmon	175	0.9
cyan	131	0.7
midnightblue	126	0.6

For the first time, the grey module is not the largest one, even containing less than 1/5 of the genes.
Using the linear mixed model approach, the following module-trait associations are modelled:

Table 35. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
turquoise	0.130	**25.732**	**8.29e-44**
brown	-0.088	**-10.419**	**3.01e-17**
salmon	0.082	**9.314**	**6.35e-15**
blue	-0.082	**-9.286**	**7.27e-15**
black	-0.073	**-8.105**	**2.17e-12**
green	0.062	**7.012**	**3.86e-10**
magenta	0.064	**6.561**	**3.09e-09**
purple	0.051	**5.721**	**1.31e-07**
red	-0.037	**-4.897**	**4.14e-06**
grey	0.019	**4.631**	**1.19e-05**
cyan	0.038	**4.136**	**7.81e-05**
midnightblue	0.028	**3.595**	**5.24e-04**
greenyellow	0.032	**3.487**	**7.51e-04**
yellow	-0.032	**-3.394**	**1.02e-03**
pink	0.017	**3.357**	**1.15e-03**
tan	0.016	1.535	1.28e-01

The significance of the association with the tumour of the huge turquoise module is remarkable. The genes seem to be far more strongly expressed in the tumour tissue compared to the normal tissue.
To take a closer look at the turquoise module, the ten genes with the highest kWithin value were calculated:

Table 36. The ten genes in the turquoise module with the highest intramodular connectivity

Gene	Module	kWithin	Module Membership (turquoise)	GS
STAB2	turquoise	535.927	-0.885	33.327
ECM1	turquoise	535.443	-0.906	28.033
PTH1R	turquoise	535.125	-0.903	23.435
CFP	turquoise	535.078	-0.894	30.552
DNASE1L3	turquoise	534.278	-0.917	29.370
OIT3	turquoise	531.608	-0.922	31.464
CCNB1	turquoise	531.344	0.914	20.121
CRHBP	turquoise	530.245	-0.909	34.326
PDE2A	turquoise	529.947	-0.913	21.106
MARCO	turquoise	520.633	-0.847	37.470

The comparison between the modules from paired WGCNA and pair-aware WGCNA shows that the turquoise pair-aware module actually gathered many genes from the turquoise module, as well as a lot of unassigned genes from the grey module of the paired WGCNA:

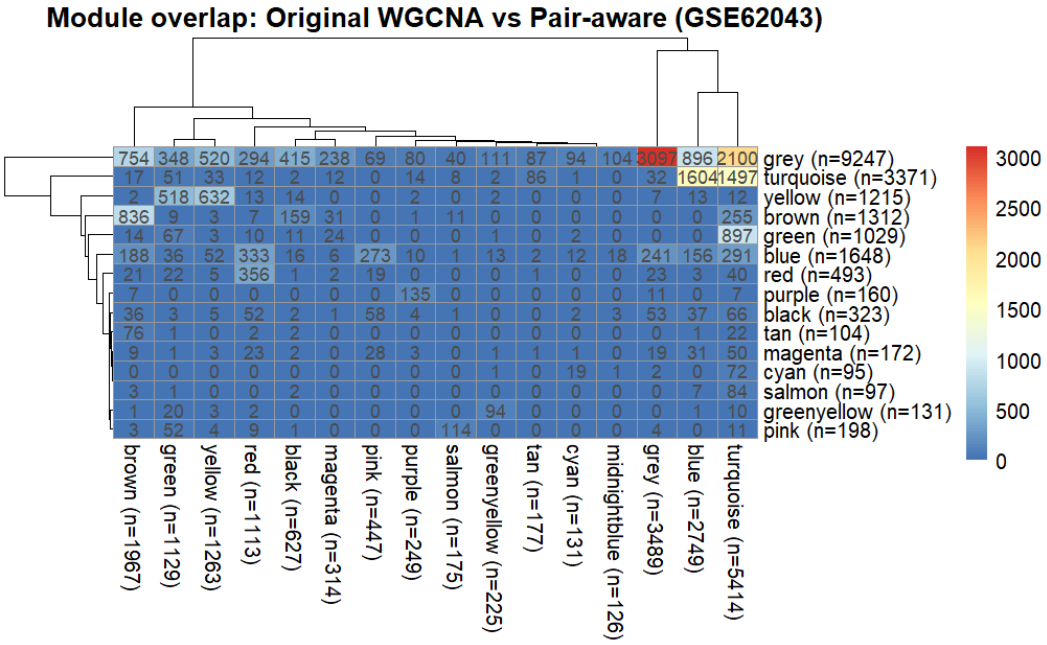

Figure 17. Cartesian product of the modules returned by paired WGCNA and pair-aware WGCNA

While the general ARI index of ARI 0.124 seems to indicate that there are no strong patterns between the modules of paired and pair-aware WGCNA, the calculation of the ARI score when excluding the grey modules shines a different light on the situation: ARI (excluding grey) = 0.3006409. So, there is at least some connection.
The hub gene table shows that the very top hub genes differ a lot, while when including more genes, the intersection becomes larger:

Table 37. Overlap of the N hub genes with the highest kWithinNorm of paired WGCNA and pair-aware WGCNA

Top N	Absolute Overlap	Relative Overlap
25	0	0 %
100	28	28 %
500	219	43 %
2000	1129	56 %

We conclude that pair-aware WGCNA again returns different results than paired WGCNA, including much more genes that were previously unassigned in meaningful modules as well. In this way, pair-aware surpasses paired WGCNA on this dataset.

4.2.5 (v) Separate WGCNA networks and module preservation

In the appendix.

4.2.6 (vi) Limma

On the second dataset, the same design matrix as for the first one was used when applying Limma, patient_id + tumour. The method yielded 14'031 significant genes, which is around 72% of all genes. When also applying the condition $|\text{logFC}| > 1$, Limma identified 790 genes. 656 (~83%) of those were downregulated in tumour and 134 (~17%) were upregulated in tumour. This disbalance can be seen in the volcano plot in figure 18. This figure also shows an MDS plot which clearly shows two clusters by condition.
Lastly, table 38 shows the 6 most significant genes from Limma (by adj.PVal).

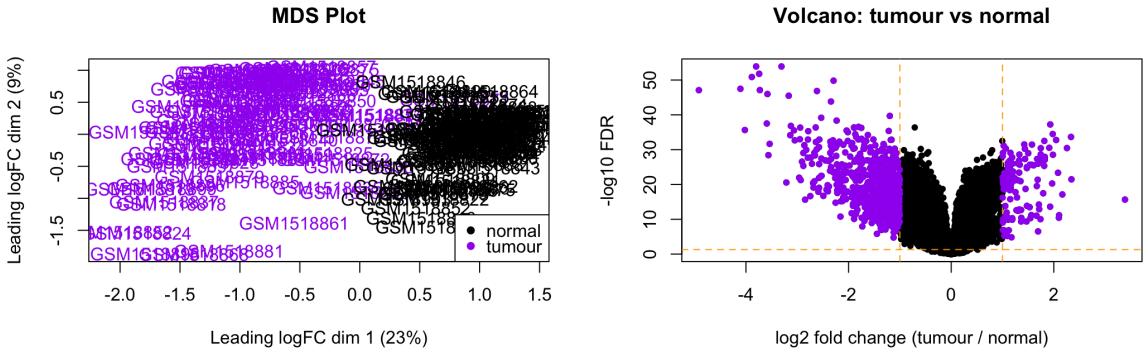

Figure 18. Results Limma for second dataset

Table 38. Top differentially expressed genes from Limma

Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
CLEC4G	-3.80	11.64	-38.10	6.64×10^{-59}	1.26×10^{-54}	123.31
MARCO	-3.31	12.67	-37.81	1.28×10^{-58}	1.26×10^{-54}	122.68
FCN3	-3.74	11.89	-35.61	2.47×10^{-56}	1.62×10^{-52}	117.59
CRHBP	-3.89	10.39	-34.65	2.68×10^{-55}	1.32×10^{-51}	115.27
STAB2	-2.29	10.55	-33.61	3.87×10^{-54}	1.52×10^{-50}	112.68
CXCL14	-4.11	10.96	-31.51	1.00×10^{-51}	3.28×10^{-48}	107.26

In the next step, these results were compared to the ones form paired WGCNA. Just like in the first dataset, the overlap of the top-N genes was computed. The importance score for paired WGCNA was computed the same way as for the first data set, $S_{wgcna}(i) = GS_i * |MM_i|$. GS_i. This outcome of the comparison can be seen in table 39. For different values of N, the overlap was very much statistically relevant and quite high, with a fraction of around 75% for all of them. However, the fraction of overlap was lower than what was observed for the first dataset.

Table 39. Fisher's exact test results for overlap among top-N genes between paired WGCNA and Limma

N	Overlap	p-value	Odds ratio	95% CI (lower)
50	36	$< 2.2 \times 10^{-16}$	1760.16	873.33
100	76	$< 2.2 \times 10^{-16}$	1277.57	758.64
150	115	$< 2.2 \times 10^{-16}$	931.17	599.52
200	154	$< 2.2 \times 10^{-16}$	714.83	505.07

Next in the comparison, it was examined which paired WGCNA modules the Limma DE genes fall into. These results are seen in table 40. Four of the 14 modules (excluding the grey module) yielded statistically significant FDR values, with a fraction of DE genes between 14.7% and 19.7%. When referencing the LMM from the paired WGCNA, it can be seen, that these four modules are within the six modules showing the strongest association to the phenotype. This suggests that the modules enriched for DE genes are also strongly phenotype-associated, consistent with Limma differential expression reflecting tumour–normal status.

Table 40. Differential-expression enrichment per WGCNA module (DE defined by Limma).

Module	Module size	# DE genes in module	Fraction DE	FDR
brown	1312	258	0.197	3.32×10^{-54}
turquoise	3371	390	0.116	7.64×10^{-26}
green	1029	98	0.095	0.0295
cyan	95	14	0.147	0.0375

The last thing examined was whether hub genes from paired WGCNA show differential expression significance from Limma. Gene significance for the tumour–normal phenotype showed an almost perfect association with Limma significance with $\rho = 0.998$ and $p < 2.2 \times 10^{-16}$, indicating that GS largely captures the same tumour–normal signal as the Limma contrast. Conversely, intramodular connectivity and module membership were only weakly to moderately associated with Limma significance (kWithin: $\rho = 0.298$; |MM| : $\rho = 0.215$; both $p < 2.2 \times 10^{-16}$). This suggests that DE strength is not simply explained by network centrality. The wgcna-score $GS_i * |MM_i|$ correlated strongly with Limma significance, $\rho = 0.957$, which is expected because it is dominated by GS and therefore reflects differential expression scaled by module centrality rather than an independent validation of DE.

4.2.7 (vii) DiffCoExp

For the second dataset, a subset of 5'000 genes with the highest variance across all samples was selected to simplify computation and reduce time. This means that any DCGs/DCLs involving lower-variance genes were not evaluated and could have been missed, so the results should be interpreted as restricted to high-variance genes.
Two separate correlation matrices were calculated, and clustering was performed based on differences in coexpression. The result was a list of DCLs (differentially coexpressed links) and DCGs (differentially coexpressed genes). The method was run using Spearman correlations with BH adjustment and thresholds of $|r| \geq 0.4$, correlation $q \leq 0.2$, $|\Delta r| \geq 0.2$, differential-correlation $q \leq 0.2$, and DCG $q \leq 0.2$.
The method yielded 706'111 DCLs, with 547'515 of those being same signed (correlation sign remains the same between conditions), 158'011 being different signed (correlation sign changes between conditions, with only one value passing the thresholds), and 585 being switched opposites (correlation sign reverses between conditions, with both values passing the threshold). The correlation difference showed absolute values between 0 and 1.21 with corresponding adjusted p-values reaching very small values which are flagged as essentially 0 by R (see table 41). This can be caused by having very different correlations in the two conditions or because this analysis was performed on a larger dataset.
Additionally, DiffCoExp identified 2'016 DCGs. The top-ranked DCGs can be seen in table 42. The genes with the highest adjusted p-values reached 1'295, 1'367 and 1'297 DCLs. For the genes shown in the table this yields a fraction of around 85-90% of their CLs, so the links they are a part of.

Table 41. Top differential co-expression links (DCLs)

Gene.1	Gene.2	cor_normal	cor_tumour	cor_diff	q_{diff}	type
NOTUM	GLUL	0.02	0.89	0.87	$< 1 \times 10^{-300}$	same signed
AXIN2	GLUL	0.01	0.90	0.89	$< 1 \times 10^{-300}$	same signed
SBF2	GLUL	0.12	0.93	0.81	$< 1 \times 10^{-300}$	same signed
ATP1B3	ADAMTS1	0.88	0.11	-0.76	$< 1 \times 10^{-300}$	same signed
DPYSL2	NDN	0.85	0.03	-0.83	$< 1 \times 10^{-300}$	same signed
MALL	LOC100132330	0.51	0.95	0.44	$< 1 \times 10^{-300}$	same signed

Table 42. Top differentially co-expressed genes (DCGs) ranked by FDR

Gene (probe)	CLs	DCLs	q (FDR)
VWF	1432	1295	4.74×10^{-311}
CYBRD1	1577	1367	4.13×10^{-283}
SAP30	1477	1297	3.15×10^{-281}
CLRN3	1186	1091	6.22×10^{-279}
CHRDL1	1115	1023	4.23×10^{-259}
DKK3	1507	1292	8.49×10^{-258}

In the comparison between DiffCoExp and paired WGCNA, the first step was computing the overlap among the top-N genes. For DiffCoExp, the DCL degree is defined as the number of DCLs a gene participates in, and this measure was compared to the intramodular connectivity $kWithin$ from paired WGCNA. For this analysis, genes assigned to the grey module were excluded.
A Spearman rank test yielded a correlation of around $\rho \approx 0.3$ with $p < 2.2 \times 10^{-16}$. This indicates a moderate positive relationship: genes that are more hub-like in paired WGCNA tend to be involved in more differential co-expression links. The p-value is very small, which is expected given the large number of genes tested. The two variables are plotted against each other in Figure 19; visually, the association appears noisy, consistent with a moderate effect size.

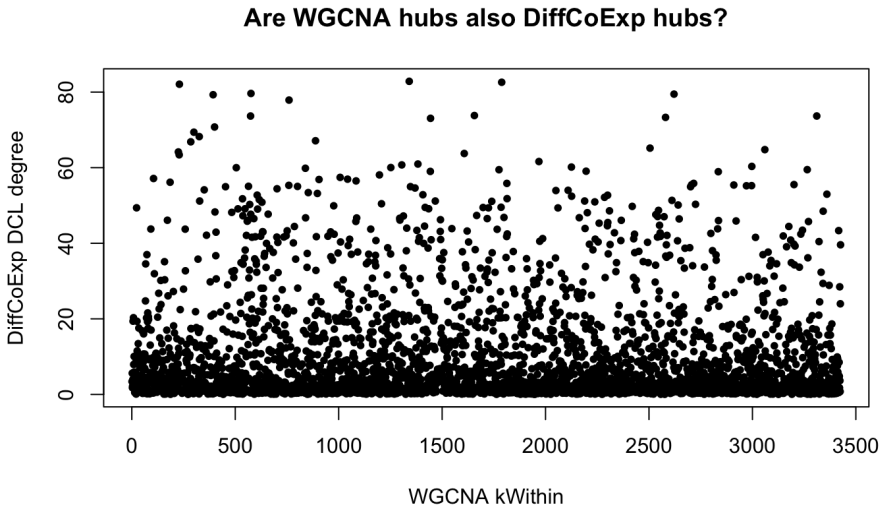

Figure 19. DGL degree vs kWithin

Next, each DCL edge was assigned to paired-WGCNA modules, enrichment in the tumour-associated green module was tested while adjusting for module size, as the green modules showed the most tumour association using the corresponding MLL in method (i). Overall, 14.22% of DCLs touched green, which is lower than expected under a random-pair null of 18.90%, corresponding to 0.75 times the expected fraction. In contrast, 2.46% of DCLs were within green, which is higher than expected compared with 0.99%, corresponding to about 2.49 times enrichment, with $p < 2.2 \cdot 10^{-16}$. This suggests that DCLs are not broadly concentrated around green, but DCLs that involve green are more likely to connect genes inside the same green module.
In the last step, after mapping each DCL endpoint to its paired WGCNA module and removing edges with missing module assignments, 40.15% of DCLs showed to be within-module and 59.85% were between-module. In counts, this corresponds to 183'473 within-module DCLs and 273'443 between-module DCLs. This indicates that a substantial fraction of correlation changes occurs across module boundaries rather than being confined to single modules.

4.2.8 (viii) Graphical Lasso

Using Glasso with StARS-based regularisation selection, a sparse gene network was built on the second dataset. This analysis was performed on the 500 genes which showed the most variance to simplify computation. It is important to keep in mind that these result do not apply to the whole dataset.
The analysis revealed a network with 7'012 edges (density 0.0562) for normal samples and 7'327 edges (density 0.0587) for tumour samples. An edge refers to a non-zero value between two genes in the network. This indicates a small overall increase in conditional dependence connectivity in tumour.
Additionally, 5'608 edges were gained in tumour, and 5'293 edges were lost in tumour. This corresponds to widespread rewiring with only a modest net change in edge count.
Table 43 shows the six top genes based on the degree gained from normal to tumour. The degree is the number of edges touching a gene.
For the comparison with paired WGCNA, the first thing considered was whether the edges defined by Glasso fall within or between paired WGCNA modules. The grey modules was excluded from this analysis, which left over 398 genes for the module–network comparison. Using these shared genes, the fraction of Glasso edges that connected two genes from the same WGCNA module was 0.618 in normal and 0.654 in tumour. This indicates that, in both conditions, Glasso edges preferentially occur within paired-WGCNA modules, with a slightly higher within-module tendency in tumour.
To account for different module sizes, an expected within-module edge fraction was computed under random edge placement, given the observed module sizes, which yielded an expected fraction of 0.347. Relative to this expectation,

Table 43. Top genes with the largest degree increase in tumour relative to normal from Glassso

gene	degree (N)	degree (T)	strength (N)	strength (T)	Δ degree	Δ strength
RELN	28	52	1.18	1.33	24	0.15
C3orf32	13	37	0.53	1.26	24	0.73
LCN2	20	43	0.96	1.46	23	0.50
PAGE4	19	40	0.86	1.32	21	0.47
HAL	26	47	1.24	1.54	21	0.30
HSD17B6	12	32	0.46	1.03	20	0.57

within-module edges were enriched 1.78 times in normal and 1.88 times in tumour. This supports that conditional dependence edges inferred by Glasso align with the WGCNA module structure more strongly than expected by chance, and that this alignment is marginally stronger in tumour.
Table 44 details which edges from which modules are linked. These matrices are symmetric, so only the values above the diagonal are shown. Along the diagonal are the within modules connections. For both networks, most links are found within the modules. The brown and turquoise modules contain the most links across both conditions. Overall, tumour shows a slight increase in within-module connectivity for several modules, while some large cross-module links involving turquoise become less frequent

Table 44. Glasso edges between paired-WGCNA modules

	Normal							Tumour						
Module pair	**brown**	**cyan**	**green**	**grey**	**purple**	**tan**	**turquoise**	**brown**	**cyan**	**green**	**grey**	**purple**	**tan**	**turquoise**
brown	1330	26	172	745	90	12	1064	1557	61	186	818	173	22	871
cyan		164	33	62	17	3	68		198	34	94	12	0	82
green			95	165	8	1	174			110	130	11	0	178
grey				409	105	14	838				407	119	5	676
purple					35	0	112					69	0	117
tan						1	6						1	12
turquoise							1263							1384

Lastly, the overlap between the top-N Glasso degree genes and paired WGCNA kWithin genes was computed. The Spearman correlation yielded $\rho = 0.24$ in normal and $\rho = 0.21$ in tumour. This indicates a weak to moderate positive association, meaning genes that are highly connected in the sparse Glasso network tend to be somewhat more connected within their WGCNA modules, but the agreement is limited.
When consider the top 100 genes across both methods, there was an overlap of 33 for normal and 31 for tumour. This is consistent with the correlation results and suggests only partial concordance between the hub rankings, with slightly weaker agreement in tumour.

5 Comparison with Ground Truth

Because there is no experimentally verified ground truth for differential modules or hubs in these datasets, literature-based validation heuristic were used to contextualize the biological plausibility of top-ranked candidates. Concretely, for each method the highest-ranked genes/miRNAs according to the method-specific importance measure (e.g., hubness, module membership, degree, or differential connectivity) were collected and it was checked whether they have been previously reported in the context of the corresponding cancer type. This reference set is not treated as definitive truth, but as an external sanity check that complements the statistical comparisons reported above.

Table 45. Literature support for method-selected miRNAs in oral cancer / OSCC, including the original Illumina probe IDs from GSE45238.

Method	Probe ID (as in experiment)	miRNA (miRBase name)	Literature support
paired WGCNA	ILMN_3167805	hsa-miR-487b	2/5 ⋆⋆
paired WGCNA	ILMN_3167455	hsa-miR-30a	3/5 ⋆⋆⋆
paired WGCNA	ILMN_3166941	hsa-miR-376c	3/5 ⋆⋆⋆
paired WGCNA	ILMN_3167624	hsa-miR-136	1/5 ⋆
paired WGCNA	ILMN_3167988	hsa-miR-411	1/5 ⋆
paired WGCNA	ILMN_3167522	hsa-miR-154	3/5 ⋆⋆⋆

Method	Probe ID (as in experiment)	miRNA (miRBase name)	Literature support
paired WGCNA	ILMN_3168513	hsa-let-7c	3/5 ⋆⋆⋆
paired WGCNA	ILMN_3168646	hsa-miR-21*	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	ILMN_3167031	hsa-miR-127-3p	1/5 ⋆
paired WGCNA	ILMN_3168716	hsa-miR-337-3p	1/5 ⋆
naive WGCNA	ILMN_3168646	hsa-miR-21*	5/5 ⋆⋆⋆⋆⋆
naive WGCNA	ILMN_3168513	hsa-let-7c	3/5 ⋆⋆⋆
naive WGCNA	ILMN_3168273	hsa-miR-503	1/5 ⋆
naive WGCNA	ILMN_3168388	hsa-miR-7	4/5 ⋆⋆⋆⋆
naive WGCNA	ILMN_3167455	hsa-miR-30a	3/5 ⋆⋆⋆
WGCNA on log ratios	ILMN_3167052	hsa-miR-495	1/5 ⋆
WGCNA on log ratios	ILMN_3168757	hsa-miR-145*	4/5 ⋆⋆⋆⋆
WGCNA on log ratios	ILMN_3166935	hsa-miR-329	1/5 ⋆
WGCNA on log ratios	ILMN_3168866	hsa-miR-340	1/5 ⋆
WGCNA on log ratios	ILMN_3168815	hsa-miR-361-3p	1/5 ⋆
WGCNA on log ratios	ILMN_3168183	hsa-miR-129-5p	1/5 ⋆
WGCNA on log ratios	ILMN_3168373	hsa-miR-518d-3p	0/5 –
WGCNA on log ratios	ILMN_3168541	hsa-miR-548b-5p	0/5 –
WGCNA on log ratios	ILMN_3168603	hsa-miR-298	0/5 –
WGCNA on log ratios	ILMN_3167503	hsa-miR-609	0/5 –
pair-aware WGCNA	ILMN_3168513	hsa-let-7c	3/5 ⋆⋆⋆
pair-aware WGCNA	ILMN_3167455	hsa-miR-30a	3/5 ⋆⋆⋆
pair-aware WGCNA	ILMN_3168646	hsa-miR-21*	5/5 ⋆⋆⋆⋆⋆
pair-aware WGCNA	ILMN_3167805	hsa-miR-487b	2/5 ⋆⋆
pair-aware WGCNA	ILMN_3168707	hsa-miR-921	0/5 –
pair-aware WGCNA	ILMN_3168388	hsa-miR-7	4/5 ⋆⋆⋆⋆
pair-aware WGCNA	ILMN_3167729	hsa-miR-30c	3/5 ⋆⋆⋆
pair-aware WGCNA	ILMN_3168273	hsa-miR-503	1/5 ⋆
pair-aware WGCNA	ILMN_3166941	hsa-miR-376c	3/5 ⋆⋆⋆
pair-aware WGCNA	ILMN_3168749	hsa-miR-455-3p	1/5 ⋆
Limma	ILMN_3168513	hsa-let-7c	3/5 ⋆⋆⋆
Limma	ILMN_3167455	hsa-miR-30a	3/5 ⋆⋆⋆
Limma	ILMN_3167729	hsa-miR-30c	3/5 ⋆⋆⋆
Limma	ILMN_3168707	hsa-miR-921	0/5 –
Limma	ILMN_3167805	hsa-miR-487b	2/5 ⋆⋆
Limma	ILMN_3167818	hsa-miR-432	1/5 ⋆
DiffCoExp	ILMN_3168167	hsa-miR-187	1/5 ⋆
DiffCoExp	ILMN_3168700	hsa-miR-886-5p	1/5 ⋆
DiffCoExp	ILMN_3167182	hsa-miR-518c	0/5 –
DiffCoExp	ILMN_3166957	hsa-miR-182*	2/5 ⋆⋆
DiffCoExp	ILMN_3167807	hsa-miR-661	0/5 –
DiffCoExp	ILMN_3167148	hsa-miR-769-5p	0/5 –
Graphical Lasso	ILMN_3167452	hsa-miR-617	0/5 –
Graphical Lasso	ILMN_3168788	hsa-miR-888	0/5 –
Graphical Lasso	ILMN_3168097	hsa-miR-658	0/5 –
Graphical Lasso	ILMN_3168701	hsa-miR-200c*	2/5 ⋆⋆
Graphical Lasso	ILMN_3168706	hsa-miR-331-5p	1/5 ⋆
Graphical Lasso	ILMN_3167509	hsa-miR-363*	1/5 ⋆

The same was done for the second dataset.

Table 46. Literature support for HCC involvement of genes returned by different methods (GSE62043). Scores reflect how strongly each gene is conventionally linked to hepatocellular carcinoma (HCC) in prior literature.

Method	Gene (symbol)	Literature support
paired WGCNA	GMFG	1/5 ⋆
paired WGCNA	LCP2	1/5 ⋆
paired WGCNA	ARHGDIB	2/5 ⋆⋆
paired WGCNA	CD53	1/5 ⋆
paired WGCNA	RCSD1	1/5 ⋆
paired WGCNA	SPC25	4/5 ⋆⋆⋆⋆
paired WGCNA	BUB1	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	MELK	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	CDCA5	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	NUF2	4/5 ⋆⋆⋆⋆
paired WGCNA	SCAMP3	3/5 ⋆⋆⋆
paired WGCNA	SNX27	2/5 ⋆⋆
paired WGCNA	SCNM1	4/5 ⋆⋆⋆⋆
paired WGCNA	PRCC	3/5 ⋆⋆⋆
paired WGCNA	VPS72	4/5 ⋆⋆⋆⋆
paired WGCNA	PTTG1	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	CDKN3	4/5 ⋆⋆⋆⋆
paired WGCNA	BIRC5	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	CCNB1	5/5 ⋆⋆⋆⋆⋆
paired WGCNA	SAC3D1	4/5 ⋆⋆⋆⋆
paired WGCNA	CLEC4G	3/5 ⋆⋆⋆
paired WGCNA	MARCO	4/5 ⋆⋆⋆⋆
paired WGCNA	FCN3	3/5 ⋆⋆⋆
paired WGCNA	CRHBP	4/5 ⋆⋆⋆⋆
paired WGCNA	STAB2	3/5 ⋆⋆⋆
paired WGCNA	CCT3	4/5 ⋆⋆⋆⋆
paired WGCNA	FAM189B	3/5 ⋆⋆⋆
paired WGCNA	GBA	2/5 ⋆⋆
paired WGCNA	KRTCAP2	3/5 ⋆⋆⋆
WGCNA on log ratios	SIGLEC7	2/5 ⋆⋆
WGCNA on log ratios	LILRA1	2/5 ⋆⋆
WGCNA on log ratios	LILRA4	1/5 ⋆
WGCNA on log ratios	LILRP2	1/5 ⋆
WGCNA on log ratios	SIGLEC11	2/5 ⋆⋆
WGCNA on log ratios	BUB1	5/5 ⋆⋆⋆⋆⋆
WGCNA on log ratios	KIF18B	4/5 ⋆⋆⋆⋆
WGCNA on log ratios	C15orf42 (TICRR)	2/5 ⋆⋆
WGCNA on log ratios	RAD54L	4/5 ⋆⋆⋆⋆
WGCNA on log ratios	EXO1	4/5 ⋆⋆⋆⋆
pair-aware WGCNA	STAB2	3/5 ⋆⋆⋆
pair-aware WGCNA	ECM1	3/5 ⋆⋆⋆
pair-aware WGCNA	PTH1R	1/5 ⋆
pair-aware WGCNA	CFP	2/5 ⋆⋆
pair-aware WGCNA	DNASE1L3	4/5 ⋆⋆⋆⋆
pair-aware WGCNA	OIT3	3/5 ⋆⋆⋆
pair-aware WGCNA	CCNB1	5/5 ⋆⋆⋆⋆⋆
pair-aware WGCNA	CRHBP	4/5 ⋆⋆⋆⋆
pair-aware WGCNA	PDE2A	3/5 ⋆⋆⋆
pair-aware WGCNA	MARCO	4/5 ⋆⋆⋆⋆
Limma	CLEC4G	3/5 ⋆⋆⋆
Limma	MARCO	4/5 ⋆⋆⋆⋆
Limma	FCN3	3/5 ⋆⋆⋆

Continued on next page

Method	Gene (symbol)	Literature support
Limma	CRHBP	4/5 ⋆⋆⋆⋆
Limma	STAB2	3/5 ⋆⋆⋆
Limma	CXCL14	4/5 ⋆⋆⋆⋆
DiffCoExp	VWF	3/5 ⋆⋆⋆
DiffCoExp	CYBRD1	1/5 ⋆
DiffCoExp	SAP30	2/5 ⋆⋆
DiffCoExp	CLRN3	0/5 –
DiffCoExp	CHRDL1	2/5 ⋆⋆
DiffCoExp	DKK3	3/5 ⋆⋆⋆
Graphical Lasso	RELN	4/5 ⋆⋆⋆⋆
Graphical Lasso	C3orf32	0/5 –
Graphical Lasso	LCN2	4/5 ⋆⋆⋆⋆
Graphical Lasso	PAGE4	1/5 ⋆
Graphical Lasso	HAL	1/5 ⋆
Graphical Lasso	HSD17B6	3/5 ⋆⋆⋆

From this limited sample, it seems like paire-aware WGCNA, naive WGCNA and paired WGCNA performed similarly, whereas WGCNA on log ratios, DiffCoExp and Graphical Lasso did not capture many miRNAs/genes known to be associated with the phenotype. Note that the latter two did not focus on finding hub genes. Limma captured a moderate amount of relevant miRNAs in the first dataset but literature support is much higher for the second. This showed to be an overall trend.
It is important to understand that this is just one attempt at a ground truth. There are no clear labels whether a gene is associated with a phenotype. Statistical methods can help us identify tendencies and create hypotheses, but there is no actual answer.

6 Discussion

As naive and paired WGCNA have by definition always the same modules, the quality of their results can only be measured in the importance of the returned hub genes. While here, no big differences were found in the very limited study of how much conventional knowledge supports the hypothesis of the genes being hub genes, mathematically speaking, Li et al.'s paired WGCNA approach is necessary to handle paired data.
While it first seemed that WGCNA on log ratios achieved outcomes that looked in appearance similar to the ones of paired WGCNA, though with very different results, the study (though limited) assessing the hub genes returned with conventional knowledge showed in both datasets that the method did not output the most associated genes. Thus, this study cannot confirm the effectiveness of this form of analysis.
Pair-aware WGCNA did perform very well. When applied to the first dataset, the results of the hub genes were quite similar to those received by paired WGCNA. On the second dataset, this differed more. However, both datasets were unified by the fact that the grey module was much smaller in the pair-aware WGCNA than in the paired WGCNA, a feature that is very much wished for. So, this study indicates that pair-aware WGCNA can be used instead of paired WGCNA, leading to more assigned genes.

Limma is a useful tool when assessing the outcome of paired WGCNA because it provides an independent differential-expression based benchmark for phenotype association. For the first dataset, the results showed a large overlap between the top-N hub genes from both methods, and for all tested N, Fisher's exact test yielded high significance. These findings are consistent with the turquoise module containing a large fraction of differentially expressed genes by Limma, as it showed the highest tumour association for paired WGCNA.
On the second dataset, the same pattern was visible, though less distinct. The overlap of hub genes showed high fractions across all values for N. Differential expression enrichment was also not uniform across modules, and the modules enriched for Limma hits are among those with the strongest tumour associations in the paired WGCNA mixed model analysis. These results show, that paired WGCNA assigns phenotype-relevant labels in a way that broadly aligns with a classical differential expression analysis.

DiffCoExp provides a different angle than Limma by focusing on changes in correlation structure rather than shifts in mean expression. Thus, comparing DiffCoExp to paired WGCNA offers an additional perspective on phenotype-associated signal. For the first dataset, no statistically significant relationship between paired WGCNA intramodular connectivity and DiffCoExp link degree was found. Here, DiffCoExp link degree refers to the number of significant differentially coexpressed links incident to a miRNA/gene. This implies that hubness within coexpression modules does

not necessarily identify miRNAs whose correlation neighbourhood rewires most strongly between tumour and normal, and paired WGCNA hub labels should therefore not be interpreted as a substitute for differential coexpression. At the same time, DiffCoExp rewiring is not independent of paired WGCNA structure. Many differential links involve miRNAs from the phenotype-associated turquoise module, and differential links tend to have higher TOM connectivity in the paired WGCNA network than random miRNA pairs. This suggests that rewiring often occurs among pairs that are already coexpressively connected in the baseline network.
For the large dataset, a moderate relationship between differential link degree and kWithin was shown, which indicates that WGCNA hub-like genes can be more involved in rewiring when sample size and signal are higher, but the association remains noisy, and many correlation changes occur between modules rather than within a single module. Overall, DiffCoExp supports paired WGCNA as a meaningful structural partition, but it also highlights that phenotype-associated biology can manifest through altered interactions between modules, not only through within-module hub genes.

Glasso infers sparse conditional dependence networks, so agreement with paired WGCNA is not expected to be perfect since WGCNA is based on marginal correlation and topological overlap. Still, the Glasso edges showed a clear alignment with paired WGCNA modules in both datasets, with a substantial enrichment of within-module edges relative to what would be expected from module sizes. This supports the idea that paired WGCNA modules capture biologically meaningful groupings that remain apparent even under a conditional dependence view of the data.
However, hub prioritisation differed between methods and varied across datasets, indicating that Glasso and paired WGCNA capture different notions of centrality. Moreover, the inferred Glasso networks exhibited extensive edge turnover between normal and tumour, consistent with widespread rewiring of conditional dependence structure. Overall, these results suggest that paired WGCNA and Glasso agree more strongly on modular organisation than on the ranking of individual hub features.

7 Appendix

7.1 Separate WGCNA networks and module preservation

For reasons of space, a further WGCNA variant method was moved to the appendix. The method did not perform well at all on the smaller, original dataset, clustering more than 2/3 of the genes into the unassigned grey modules, with the remaining modules being associated non-significantly with the trait. It performed a bit better on the second dataset, but having still more than 40 % unassigned genes is still not ideal for a network and module-based method. Also, performance issues prevent the method from being rerun quickly and make working with it a tedious matter, especially on the bigger dataset. Therefore, the method cannot be recommended.
Nevertheless, to present the insights also of the failed approach, here, the method is first explained, and then the results obtained when being applied on the two datasets are described (only main results for the second dataset due to unimportance and big performance problems).

7.1.1 Explanation of the method

This method, probably the most costly of them all, is based on the idea to build separate WGCNA networks on the tumour data and on the normal data and then comparing the created modules.
After preprocessing, the data is split into two matrices, one with the normal data for each patient and one with the tumour data of each patient. Both matrices in themselves contain independent data, as every row contains data from another patient. We can thus apply the original, independent WGCNA method (with the unsigned method for the adjacency scores) on each matrix individually. Note that both matrices contain column-wise the information on the same genes.
Important to mention is that for comparability of the networks later, the same soft threshold is chosen for both networks, even if the scale independence diagnostics indicate otherwise.
In a first step, `blockwiseConsensusModules()`, a function that treats the two inserted matrices (one for tumour data and one for the normal data) as independent (not paired!), is applied. It calculates the adjacency matrices and the topological overlap measures for both datasets. Next, it builds a gene–gene similarity matrix that represents the similarity in both tumour and normal. This can be imagined as follows: genes with strong similarity scores in both tumour and normal tissue are assigned a strong similarity score in the consensus network, while having a weak similarity score in at least one of the separate networks is allocated a weak score in the consensus network. Next, the function builds a hierarchical clustering on the basis of the consensus dissimilarity (1 - consensus similarity). The resulting dendrogram is then cut to return a number of modules. Also here, a merging step is implemented: If two modules have module eigengenes that are strongly correlated, they are merged. The function outputs then the merged, meaningful modules. In conclusion, the function tries to build modules on the basis of both adjacency matrices, favouring those gene combinations that have high adjacency scores in both matrices.
Using these consensus modules, the module eigengenes of the modules based on the two separate expression value matrices (one for tumour data, one for normal data) can be calculated, to have a quick sanity check whether the consensus modules make sense on the individual subdatasets. Additionally, the module eigengenes of the consensus

modules using the whole data (tumour and normal) are computed. On these consensus module eigengenes based on all data, now, a linear mixed model is built. As for the linear mixed model in the approach of WGCNA on paired data, m represents a module, i the patient, and j again the indicator whether the sample is tumorous or not. The model is then again the following:

$$\mathrm{ME}_{ij}^{(m)} = \beta_{0,i}^{(m)} + \beta_1^{(m)}\,\mathrm{tumour}_j + \beta_2^{(m)} \cdot \mathrm{age}_i + \beta_3^{(m)} \cdot \mathrm{stage}_i + \varepsilon_{ij}^{(m)},$$

where $\mathrm{ME}_{ij}^{(m)}$ is the value of the module eigengene of module m for the sample j of patient i. Again, as before, the crucial test statistic is the t-value of $\beta_1^{(m)}$, indicating how statistically significant the consensus module-trait relationship is, thus answering whether the module's average eigengene level change between normal and tumour (within patients).
In the next step, a new question is examined: Do the modules' network structures (adjacency scores and topological overlap measures) stay the same between tumour and normal? The function `modulePreservation()`, using the consensus modules as input, evaluates how coherent these modules are in the original subdatasets (tumour and normal). For this, it first constructs networks (i.e. adjacency scores and topological overlap measures) using the unsigned method on both subdatasets. For each of the consensus modules, it then calculates some "preservation statistics" for both constructed networks. These statistics assess whether genes are strongly connected in the modules, and whether the same genes are the hub genes in both subdataset networks. This is evaluated on both sides (one subdataset is taken to be the reference set, while the other is the test set, then this is switched). The statistics are then combined to the Zsummary. To compute this Zsummary, a bit of randomness is included. Precisely, for each module, the function picks a set of genes (randomly chosen out of all genes) of the same size and applies the same preservation statistics. This is repeated for `nPermutations` times, a hyperparameter that should be chosen carefully as each permutation is quite expensive. Using these random results, a null mean and standard deviation is calculated to calculate the ZSummary as the following Z-score:

$$\frac{\mathrm{Observed} - \mathrm{NullMean}}{\mathrm{NullStandardDeviation}}.$$

If the ZSummary is larger than 10, the modules are assumed to be strongly preserved, while a ZSummary smaller than 2 indicates weak or no preservation. Everything in between is considered moderate preservation. `modulePreservation()` often includes a "gold" module, which is a random gene set used as a control, respectively, baseline for preservation statistics. It is not biologically meaningful and is usually excluded from downstream interpretation (often along with "grey" which is the module of unassigned genes).
Because the consensus modules were created based on modules on two separately constructed networks, the function here serves as a sanity check and high ZSummary values are expected. To see whether this is also the case when consensus modules are not built based on the modules of the networks on the two subdatasets, the following steps are taken:
Until now, all WGCNA methods were part of larger, more complicated functions, resulting in the fact that never modules for two separately constructed networks were obtained. Therefore, now, the original, independent WGCNA is once again "manually" executed on both subdatasets. For simplicity of the code, the entire WGCNA method was implemented into a single function `run_single_wgcna()`. The resulting modules of the WGCNA done on the tumour data and on the normal data are then compared using the function `modulePreservation()`. Again, the comparison is done on both sides, and again, the ZSummary is returned for each module. Now, the crucial question is whether the modules are less preserved or not. Besides ZSummary, medianRank is another statistics indicating the preservation of modules. It is a rank-based order statistic, based on multiple preservation statistics, therefore more robust, but also with less statistical power.
The different ZSummary (and medianRank) statistics, as well as the t-values from the linear mixed model, are the results of this method.

7.1.2 Analysis on the smaller, original dataset (GSE45238)

The consensus modules that are built considering two networks built on the subdatasets (tumour and normal data) result in the following modules:

Table 47. Module sizes of the consensus modules from the separate WGCNA approach (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
grey	563	67.8
turquoise	193	23.3
blue	74	8.9

Here, it is also the case that the grey is the group for all the unassignable genes. The results, with more than 2/* not being assigned to any useful module are thus not very satisfying.

The situation is not getting much better when looking at the Module-tumour association based on the t-values of the linear mixed models using the original, pooled data (with both subdatasets included):

Table 48. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
grey	0.211	19.8	2.77e-21
blue	0.0124	1.12	0.268
turquoise	0.00597	0.249	0.805

While the grey module should not be looked at as a coherent module, the other two modules do not have significant module-trait associations (at the usual 95 % level). The WGCNA approach has thus failed in this case, and no further meaningful analysis can be done in the direction of the modules and hub genes.
Nevertheless, we can compare how well preserved the consensus modules were across the subdatasets, meaning measuring the connectedness of the genes in these modules in networks built separately on the subdatasets. As this is calculated with one reference subdataset and one test subdataset, the following table is divided in two halves:

Table 49. The preservation statistics for the consensus modules (first: tumour as reference, normal as test; second: vice versa)

Module	Zsummary	medianRank
turquoise	30.07	2
blue	18.11	1
grey	1.79	4
turquoise	27.17	2
blue	18.81	1
grey	1.61	4

Note that the gold module used for internal processing (with no meaning) has been removed from these tables (this is why the medianRank = 3 is missing).
While the grey module seems to incorporate any genes which are not closely clustered and thus reveals no (or only very little) evidence for preservation, the other two modules are strongly preserved.
When building two separate networks, for comparability, the same soft threshold is chosen. After consulting the scale independence plots, $\beta = 4$ is chosen as it's the first value for which both networks fulfil the scale-free topology criterion with $R^2 \geq 0.9$:

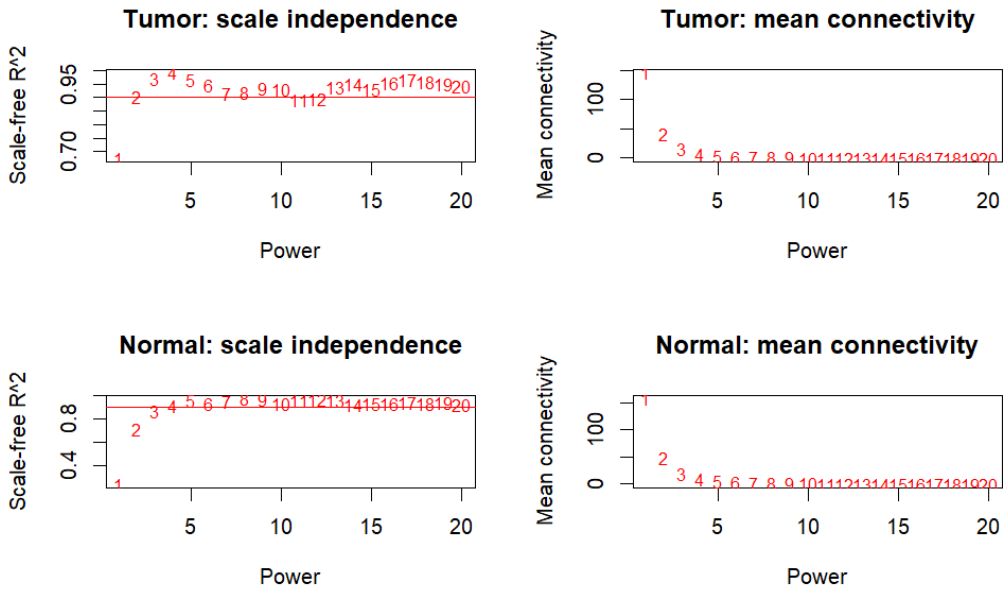

Figure 20. Scale independence plot for the two subdatasets

For these separately built networks, once again the preservation statistics between the different modules are calculated:

Table 50. The preservation statistics for the modules created by separate networks built on the subdatasets (first: tumour as reference, normal as test; second: vice versa)

Module	Zsummary	medianRank
turquoise	23.38	2
blue	14.69	1
brown	6.17	3
yellow	3.18	5
grey	-1.22	6
turquoise	22.75	2
blue	13.28	3
brown	6.81	6
yellow	6.57	1
green	4.32	4
grey	-0.68	7

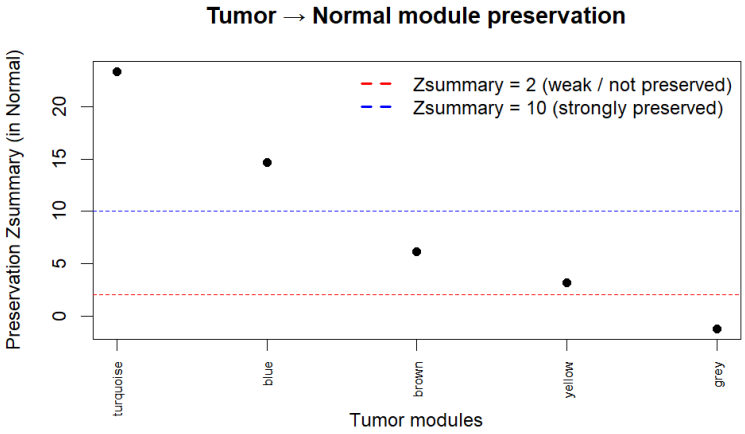

Figure 21. Plot of the ZSummary values for different modules (Normal -> Tumour)

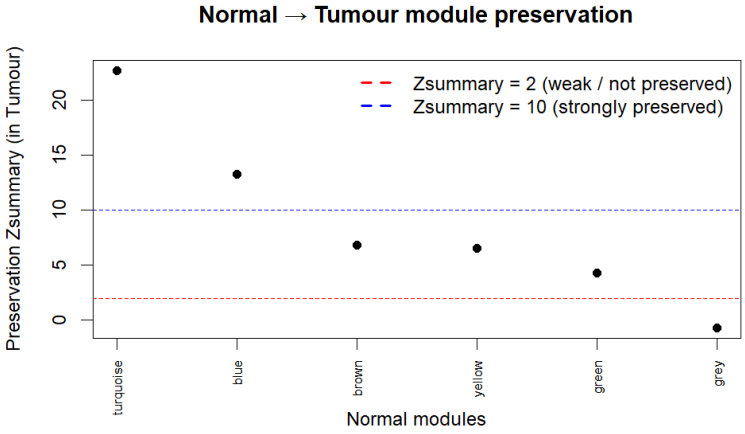

Figure 22. Plot of the ZSummary values for different modules (Tumour -> Normal)

Again, the gold module is omitted. Unsprisingly, the grey modules are not preserved at all. Of the other modules, some are strongly preserved, others, especially smaller ones, only moderately. The following heatmap supports this and gives further insight into the size of the modules:

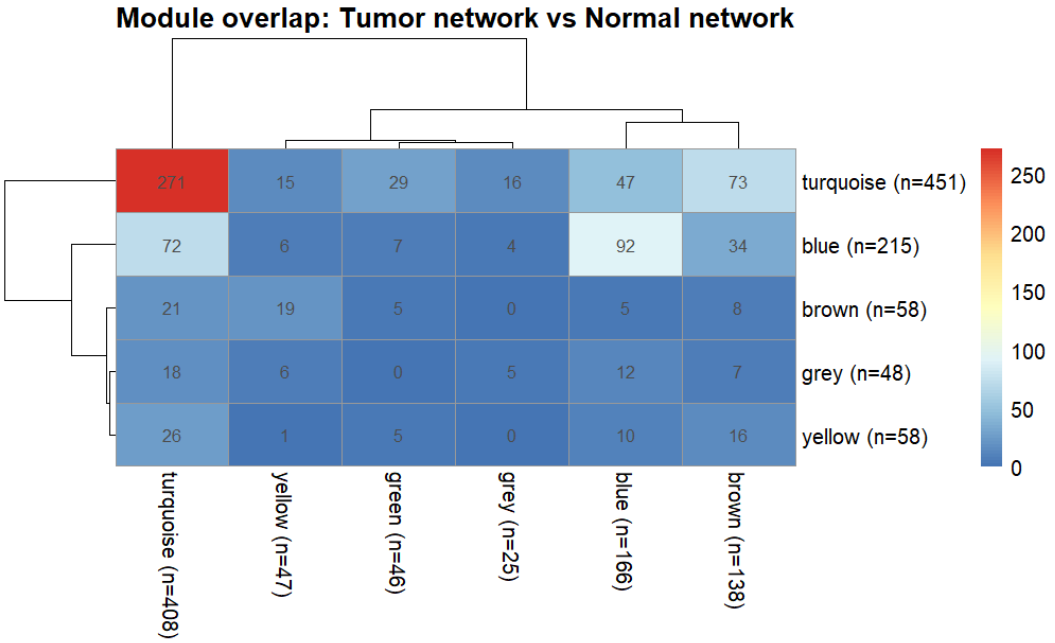

Figure 23. Cartesian product of the modules returned by the two separately built networks

In conclusion, the method did not give any great insights.

7.1.3 Analysis on the larger dataset (GSE62043)

We here just have a look at the consensus modules due to the slow (multiple hours) execution of the further code. The following consensus modules were obtained:

Table 51. Module sizes of the consensus modules from the separate WGCNA approach (original, small dataset)

Module	Number of miRNAs	Percentage of miRNAs
grey	8044	41.1
turquoise	3033	15.5
blue	2444	12.5
brown	1338	6.8
yellow	944	4.8
green	805	4.1
red	655	3.3
black	475	2.4
pink	425	2.2
magenta	275	1.4
purple	235	1.2
greenyellow	210	1.1
tan	178	0.9
salmon	170	0.9
cyan	148	0.8
midnightblue	114	0.6
lightcyan	102	0.5

For the modules, the following module-trait associations were computed using linear mixed models:

Table 52. Module-tumour associations (LMM t-values)

Module	β_{tumour}	$t_{\textbf{tumour}}$	p_{tumour}
grey	0.127	24.5	4.69e-42
greenyellow	0.107	16.4	4.93e-29
cyan	-0.101	-13.7	6.37e-24
salmon	-0.0986	-13.1	8.12e-23
yellow	-0.0876	-10.4	3.61e-17
green	-0.0809	-9.80	5.97e-16
brown	-0.0803	-9.23	9.61e-15
magenta	-0.0705	-8.19	1.46e-12
red	-0.0669	-7.63	2.11e-11
lightcyan	0.0590	6.40	6.53e-09
black	-0.0396	-5.04	2.36e-06
midnightblue	0.0498	4.88	4.45e-06
tan	0.0385	4.29	4.44e-05
purple	0.0132	4.11	8.70e-05
turquoise	-0.00412	-1.00	0.318
blue	0.00867	0.915	0.363
pink	0.00291	0.826	0.411

These results are better, as there are still some modules which have a significant association with the trait. However, due to performance issues, running further analysis is very tedious. It was thus decided not to further pursue the possibilities of the method.

7.2 Datasets and Preprocessing

7.2.1 GSE45238: Identification of oral carcinoma miRNA involved in Wnt/b-catenin signaling

The dataset was contributed by Shiah et al. and used in the original paper of Li et al. to exemplify the introduced method of applying WGCNA on paired data. It contains the data of 40 OSCC patients (who are all male), each providing a sample of tumour specimens and a sample of non-tumour epithelium. The dataset includes the expression vectors of 858 microRNAs in each of these samples.

In the preprocessing, first, non-standardised miRNAs are removed, namely, control miRNAs (none) and miRNAs not using the 12th version of the miRNA target library (28) are searched and removed. Next, as the data has not been normalised yet, the data is transformed using the log to base 2, after adding one, namely:

$$Y_{ij} = \log_2(X_{ij} + 1) \text{ for } 1 \leq i \leq 830, j \in \{1, 2\}$$

as well as adjusted using a quantile normalisation across columns (Limma's `normalizeBetweenArrays()` was applied).

For quality control, a PCA of the samples (based on their transformed expression values), and a dendrogram of the average hierarchical clustering (using the Euclidean distance between the normalised expression values of the miRNA) was done. The findings of Li et al. were confirmed. Namely, while in the PCA no samples seem very distant from the other samples, the dendrogram shows that a cluster only consisting of two samples is added at the very end (above the red line in Figure 25). These two samples were considered to be outliers, and together with their paired values were removed in the following. Indeed, those two outliers were also removed in the preprocessing done by Li et al.

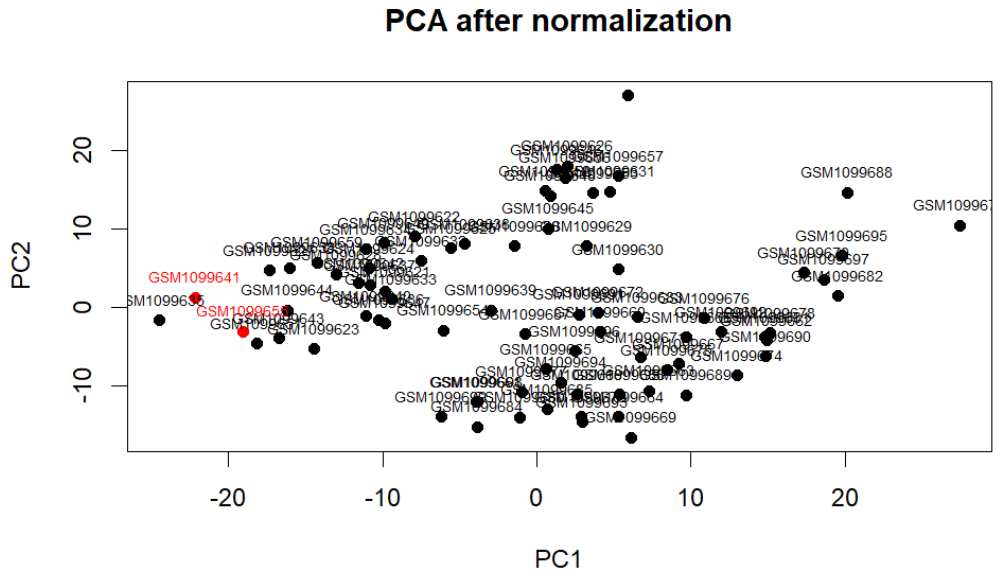

Figure 24. PCA of the samples, with the two outliers of the dendrogram marked in red

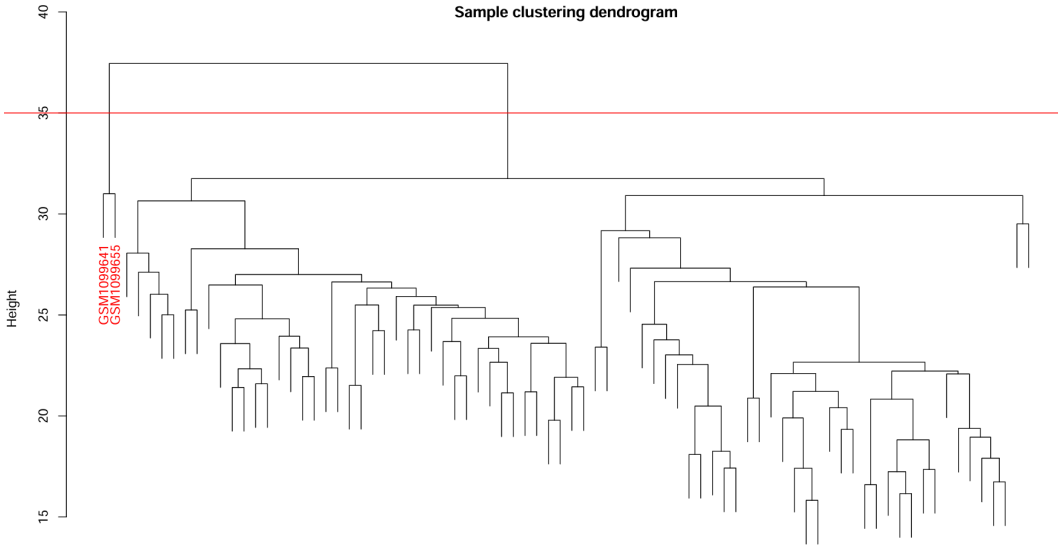

Figure 25. Dendrogram of the samples, with the two outliers at the very left removed

Finally, new, simplified patient ID (from 1 to 40) and condition variables ("tumour" or "normal") are added. This ends the preprocessing of the dataset, and the different methods are applied.

7.2.2 GSE62043: Primary and adjacent non-malignant tissue samples of 100 HCC patients (mRNA)

This dataset, contributed by Thurnherr et al., contains paired samples of 100 patients (female and male), each featuring the expression profiles of tens of thousands of genes. The pairs always consist of a sample of *adjacent non-malignant (non-tumour) tissue* as well as a sample of tissue affected by HCC (a form of liver cancer). Adjacent non-malignant tissue will be referred to as 'normal' for brevity. This dataset was chosen as it should confirm the capability of Li's WGCNA on bigger datasets. As the series matrix file contains the normalised tumour/normal ratios, in this paper, the raw data of this dataset was chosen as a starting point, complicating preprocessing a lot.

First, the metadata of the series matrix was extracted, naming all the genes and samples. Next, the raw data is downloaded into a folder, checked for accordance with the metadata, and gunzipped to `.txt` files (for limma to work). The metadata is reordered to match the order of the raw files.

These raw Agilent two-colour microarray files are now read into R as a limma "RG" object, with red (Cy5) corresponding to the tumour tissue and green (Cy3) corresponding to the *adjacent non-malignant tissue*.

Now, the column names of the red foreground intensity matrix `RG$R` and the green foreground intensity matrix `RG$G` are set to GSM IDs (Sample IDs), and the names of the genes *(Agilent probe IDs at this stage)* become the row names. As done normally, the "RG" object is then normalised - first, doing a *model-based background correction* of the background values, then fixing within-array bias and making the average values (A-values) comparable across arrays.
Next, we reconstruct the log2-intensities. As Limma stored

$$M = \log_2(R/G)$$
$$A = 0.5 \cdot \log_2(R \cdot G),$$

we calculate

$$\log R = A + M/2 \text{ for red/Cy5/tumour tissue}$$
$$\log G = A - M/2 \text{ for green/Cy3/}\textit{adjacent non-malignant tissue}$$

to finally arrive at the log2-expression estimates per probe per sample for each dye channel. Contrary to the Series matrix, this gives us not just the tumour/*non-malignant* ratio.
After dropping control probes to only remain with expression data, we now build the "tumour" and "normal" samples from the two dyes. In this experiment, RNA from the tumour sample was labelled with the red dye (Cy5), and RNA from the adjacent non-tumour sample was labelled with the green dye (Cy3), so red-channel signal corresponds to tumour RNA and the green-channel signal to non-tumour RNA. Therefore, the expression values of the `RG$R` (Cy5/red) channel are treated as the tumour sample and the ones of the `RG$G` (Cy3/green) channel as the adjacent non-tumour (non-malignant) sample for each array. This is done by adding a "_T" resp. "_N" to the values and combining `RG$R` and `RG$G` to a single matrix. Also, for confirmation, it is checked that more samples have their tissue:ch2 (corresponds to the green dye Cy3) metadata labelled as “normal” or “adjacent” than tissue:ch1. Due to empirical reasons, a final quantile normalisation was omitted. So, we end up with an expression matrix with 200 columns, corresponding to a pair of samples for each of the 100 patients.
In the next step, annotations from the GPL annotation tables are downloaded and added to the expression matrix, symbols are cleared (e.g. whitespace is removed) and multiple probes per gene are collapsed using `limma::avereps()`. We end up with an expression matrix containing gene information (mRNA expression values) row-wise and sample information column-wise (first 100 tumour samples, then 100 normal samples). The matrix is of dimension 19595 x 200.
Now, sample IDs are created and split into patient IDs (GSM) and condition ("T" and "N"), trait vectors containing information about the age and the genders are built, and the matrix `sample_info` with all information about the samples (one row per sample) is created. Finally, the order of the columns of the expression matrix is adapted to that of `sample_info`.
As for the smaller dataset, a quality control step is included. A PCA of the samples, where each sample is coloured according to its conditio,n shows possible outliers (cf. Figure 26). To make the process more statistically objective, the Sample-Sample correlation was computed and then transformed into a connectivity measure (by observing how connected each sample is, namely, adding the absolute values of the correlations and standardising). Samples, with a sample connectivity $Z.k$ being smaller than $Z.k < -2.5$ are considered outliers. These seven samples (belonging to the patients GSM1518815, GSM1518820, GSM1518844, GSM1518848, GSM1518858, GSM1518895, GSM1518814) were also highlighted in a dendrogram (created based on distances = 1 - correlation) and the PCA from before (cf. Figure 27 and Figure 28). Both figures, especially the dendrogram, confirm the choice of outliers. In the following, the paired samples of these patients is removed. The expression matrix is now of dimension 19595 x 186.

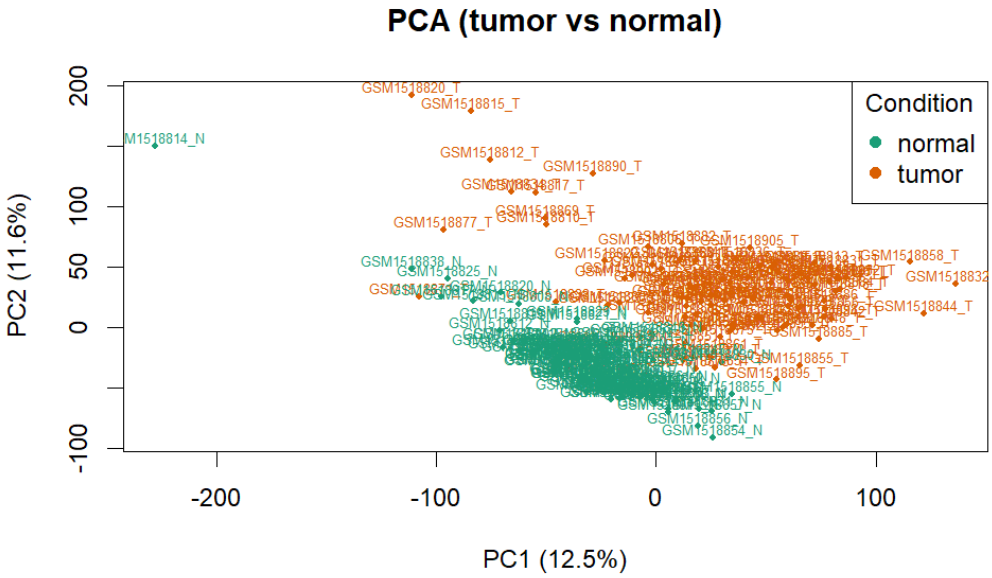

Figure 26. PCA of the samples, coloured according to condition

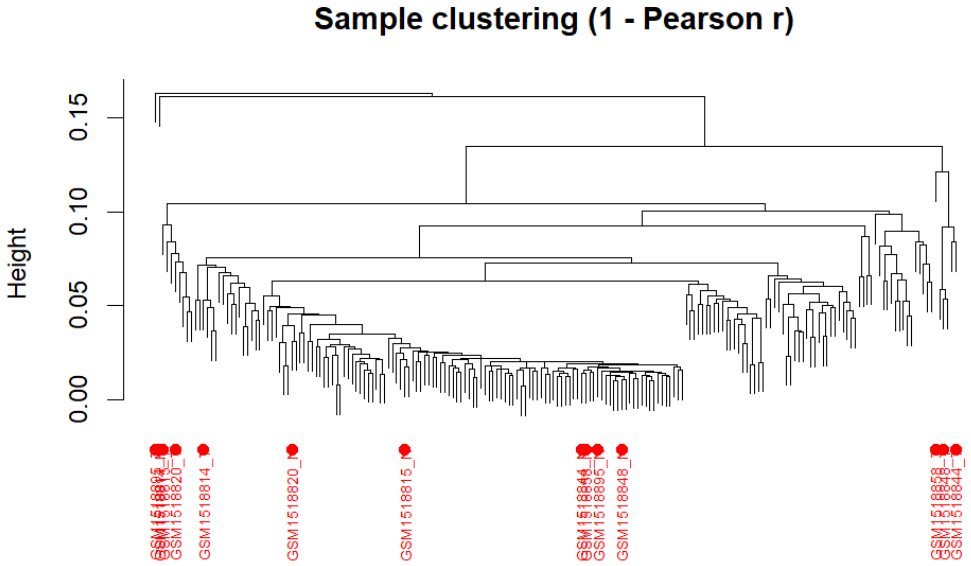

Figure 27. Dendrogram of the outliers, outliers marked

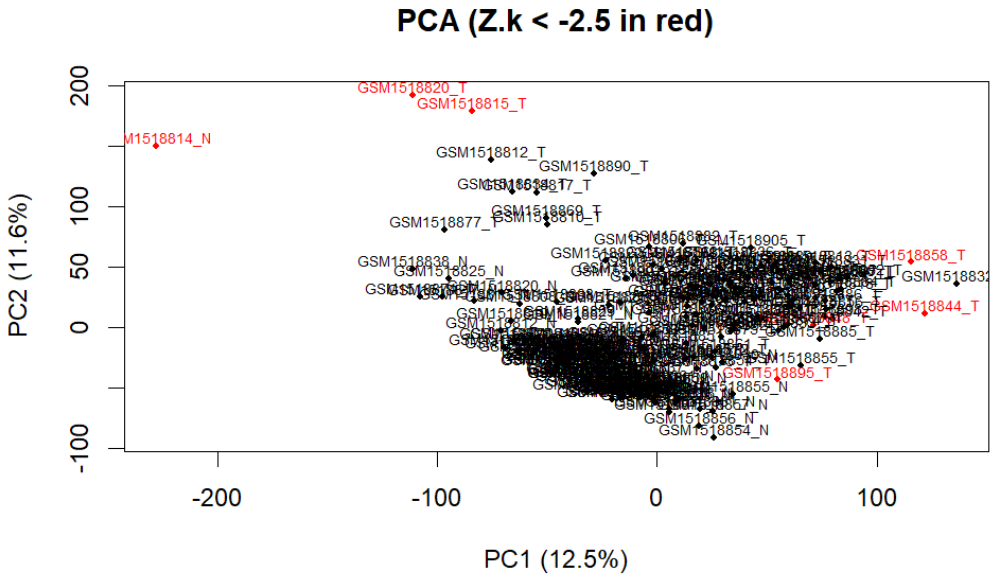

Figure 28. PCA of the samples, removed outliers in red

The preprocessing is concluded by transposing the expression matrix to get a WGCNA-ready sample x genes matrix, saving all preprocessing information (e.g. the $Z.k$ threshold) into an information list, and exporting all created data into an .RData file that was then imported into new R files to do further analysis.

7.3 (i) WGCNA on paired data

For this method, the goal was to recreate the proposed pipeline for WGCNA on paired data from [1]. Li et al. proposed the following pipeline (cf. Figure 29):

Step 0: Preprocessing of data. First, the data is preprocessed. Bad samples and bad genes like outliers or low-quality information, should be removed. The data should be normalised and, if wanted, transformed with e.g. log2 when counts or intensities are present. WGCNA needs a sample x genes expression matrix.

Step 1: Construction of gene co-expression network. For any two genes, the Pearson correlation

$$\text{cor}(x_i, x_j)$$

of the expression values x_i, x_j of these two genes is calculated (one of the crucial points of the paper of Li et al. was that no matter whether the data is independent or paired, the Pearson correlation can be used in this step). This gene-gene similarity score is then converted using a soft power. In the "unsigned" method, the entries of the adjacency matrix are calculated by raising the absolute value of the correlation to the soft threshold β, i.e.

$$a_{ij} = \left|\text{cor}(x_i, x_j)\right|^{\beta}.$$

The now constructed adjacency matrix represents the gene co-expression network of the data.

How to determine β is an art in itself. In this paper, `pickSoftThreshold()` was used to compute a diagnostic plot. It computes the adjacency matrix for each proposed β and subsequently, each gene's connectivity

$$k_i = \sum_{j \neq i} a_{ij}.$$

Then, the distribution of the k_i is observed, and a linear relationship on the log-log scale is fit. From this fit, the R^2 "scale-free topology model fit index" (Scale independence plot) is returned and plotted afterwards for each β. Indeed, the function checks how closely the adjacency values depending on the β fulfil the Scale-free topology criterion, which assumes that $P(k)$, the probability of each gene having a connectivity of k, decays as a power law for k big enough, i.e.

$$P(k) \sim k^{-\gamma} \text{ for some } \gamma > 0.$$

Besides plotting the Scale independence, a plot for the mean connectivity is also plotted, which is of lesser importance, though.

Note that there is also a "signed" method for which negative and positive correlations give different adjacency values. This will be used and further explained in the "WGCNA on the log ratios" (iii) method.

Step 2: Identify modules by using hierarchical clustering. Using the adjacency matrix from the last step, the goal is now to create a measure of distance between the genes to be able to apply hierarchical clustering. While it would be possible to just take the similarity scores from before, this is normally considered too noisy for WGCNA. The solution is to use Topological Overlap Measure (TOM). TOM not only consider whether two genes have similar expression values, but also whether two genes connect to the same other genes, so whether there are similar to the same sets of genes. These TOM Scores TOM_{ij} between genes i and j can be calculated from the adjacency matrix A from before.

For the hierarchical clustering, the distance $d_{ij} = 1 - \text{TOM}_{ij}$ is used. Normally, average hierarchical clustering is the standard in the WGCNA pipeline and was also consistently used for this paper. The resulting dendrogram is then split to give meaningful modules. In this paper, the function `cutreeDynamic()` was used. The two crucial parameters `deepSplit` and `minClusterSize` decide how aggressively the tree is cut into smaller clusters and how many genes there need to be at least in a module (to prevent tiny modules). Adapting these parameters, the function returns a set of modules of genes.

This set of modules undergoes another manually added check of merging. For this, the Module Eigengenes are calculated. Specifically, the module expression matrix (sample x genes-in-module expression matrix) is created, and PCA is applied to it. The resulting first principal component is then considered to be the Module eigengene ME_i of the module i, and used as a representation of the module.

These module eigengenes can now also be correlated (Pearson correlation) with each other to determine how closely related two modules i and j are. Namely, if

$$\text{cor}(\text{ME}_i, \text{ME}_j)$$

has a value close to 1, the two modules seem to be very similar and might be merged. To decide whether the merging should be done, an average hierarchical clustering is done using the distances $d_{ij} = 1 - \text{cor}(\text{ME}_i, \text{ME}_j)$ and then cut at some threshold height, which was often chosen to be 0.25. While this changed some module assignments in the bigger dataset, it often left things unchanged when working with the smaller dataset. The result of this step is a set of disjoint modules that form a partition of the space of all genes.

Note: Often, in this step, a grey or golden is artificially created by the functions, consisting of the remaining genes that could not be assigned in any of the other modules.

Step 3: Relate modules to phenotypes via a linear mixed-effects model. The goal is now to determine whether a gene module seems to have an effect on whether the sample is tumour or not. To account for the paired design, a Linear mixed model is used. For each module m with module eigengene $\text{ME}^{(m)}$, the following model is created:

$$\text{ME}_{ij}^{(m)} = \beta_{0,i}^{(m)} + \beta_1^{(m)} \cdot \text{Tumour}_j + \beta_2^{(m)} \cdot \text{Age}_i + \beta_3^{(m)} \cdot \text{Stage}_i + \epsilon_{ij},$$

where i is the patient and $j \in \{0,1\}$ indicates whether the sample is tumorous ($j = 1$) or normal ($j = 0$). Also, $\text{Tumour}_j = \delta_{1j}$. ϵ_{ij} are the errors that underlie the typical assumptions (homoscedasticity, mean $= 0$, normally distributed, etc.). Note also that the intercept is patient depending. The aim is to test whether $\beta_1 = 0$. In R, this is implemented by

```
lmer(MEturquoise ~ Tumour + Age + Stage_num + (1 | patient_id)).
```

The crucial test statistics here are the t-value and p-value of $\beta_1^{(m)}$ (resp. the coefficient of `tumour`). The larger the absolute t-value, the stronger the evidence against the null hypothesis (i.e., that the coefficient equals zero, equivalently that the influence of tumour on the expression values of the genes in the modules is nonexistent), and the smaller the p-value. So, the module–trait association is more statistically significant.

In the second dataset (GSE62043), while the stage of the cancer is not a given variable, the gender is now of interest (before, all patients were males). This leads to the slightly different linear mixed model:

$$\text{ME}_{ij}^{(m)} = \beta_{0,i}^{(m)} + \beta_1^{(m)} \text{Tumour}_j + \beta_2^{(m)} \text{Age}_i + \beta_3^{(m)} \text{Gender}_i + \varepsilon_{ij}^{(m)}. \tag{1}$$

Again, the coefficient of the tumour variable $\beta_1^{(m)}$ and its t-value and p-value are the points of interest.

Step 4: Identify hub genes. To find the most important genes in the modules, a number of measures is calculated.

When working with WGCNA, the gene significance is very meaningful. As there is a specialised gene significance for paired data, the usual gene significance will be referred to as the "standard gene significance", while the one for paired data will be called the "paired gene significance".

The standard gene significance (sGS) is a popular measure when working with WGCNA. While it is not used in WGCNA on paired data, it is an important component in other methods (like e.g. Naive WGCNA) and introduced here for completeness. It is calculated by correlating the expression values x_i of a gene i with the Trait and then taking the absolute value, i.e.

$$\text{sGS}_i = |\text{cor}(x_i, \text{Trait})|.$$

"Trait" must be imagined to be a boolean vector, indicating with 1 and 0 whether the sample is tumorous or not. It is the analogue of correlating the module eigengene with the trait (see "Naive method"), but on the scale of the individual gene, not of entire modules. The closer this value is to 1, the more related the gene and the trait seems (note that it is also possible that the gene appears related to when the trait is absent).
Similarly to the way the linear mixed model was included to evaluate the module-trait association in paired WGCNA, the paired gene significance (pGS) analogously introduces a linear mixed model to assess the gene-trait association. The model is similar to before:

$$x_{ij}^{(g)} = \beta_{0,i}^{(g)} + \beta_1^{(g)} \cdot \text{Tumour}_j + \beta_2^{(g)} \cdot \text{Age}_i + \beta_3^{(g)} \cdot \text{Stage}_i + \epsilon_{ij}^{(g)},$$

where now g is the gene, i the patient, and j an indicator whether the sample is tumorous or not. So, $x_{ij}^{(g)}$ is the expression profile of gene g in the sample j of patient i. Again, the intercept is patient depending. The paired gene significance pGS_g is then the absolute value of the t-value of $\beta_1^{(g)}$ (resp. the coefficient of Tumour). As before, it measures the statistical significance of the gene-trait association.
Another useful measure is the module membership $\text{MM}_{g,m}$ of a gene g to a module m. It computes how related the expression vectors of an individual gene and the one of a module (normally the module to which the gene was assigned) are and is often used as a "hubness" measure. For this, the Pearson correlation of the gene expression values and the module eigengene is calculated, i.e.

$$\text{MM}_{g,m} = \text{cor}(x_g, \text{ME}_m).$$

If $\left|\text{MM}_{g,m}\right|$ is large, the gene g behaves like a typical pattern of the module m and becomes a candidate for a hub gene of the module. If, additionally, the paired gene significance is high, the gene is considered a hub gene of the module.
Another measure to find hub genes of the modules is the intra-module connectivity. In this paper, it was used to order the proposed hub genes of each module. The intra-module connectivity, or also k-Within_g statistic called for a gene g, is the sum of all the adjacency scores between the gene and every other gene in the same module, i.e. if g is assigned to module m,

$$\text{k-Within}_g = \sum_{h \in m \setminus \{g\}} a_{gh}.$$

It, therefore, gives a quantity for the similarity of the gene g with the rest of its module m, and gives another possibility to classify the importance of the genes, and to find hub genes.
Hub genes (or hub miRNAs) can be seen as central connectors within a module: because they are strongly linked to many other genes, changes in their expression often reflect (or drive) broader changes in the whole module. For that reason, they are useful candidates for follow-up, for example, as biomarkers and as potential intervention targets if they sit upstream in a disease-relevant pathway, like in this tumour-normal setting. They are the result of the whole WGCNA process.

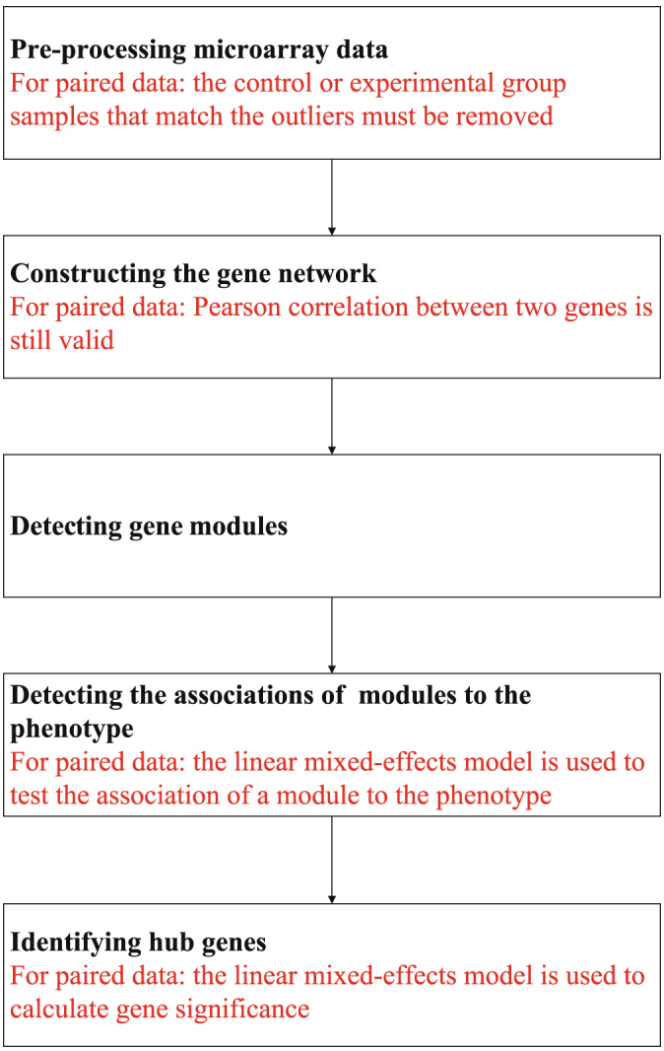

Figure 29. Naive WGCNA workflow in black, additions by Li et al. in red ([1])

Note: Because the adjacency matrix with size $(\#\text{genes})^2$ is needed, the pipeline lies in the runtime of $O(p^2)$, where p is the number of genes.

7.4 (ii) "Naive" WGCNA

In this method, the regular WGCNA pipeline for non-paired data is reproduced. The samples will be wrongly treated as independent, to get a comparison to the more sophisticated method proposed by Li et al.
As mentioned, Li et al. showed that in the first step, the simple Pearson correlation between the genes can be used no matter whether the data is independent or not. Thus, the first two steps after the preprocessing are the same in Naive WGCNA and WGCNA on paired data (Naive WGCNA also uses the unsigned method for the adjacency values). This also leads to the fact that the modules proposed by the method are exactly the same.
Differences between the two methods are only shown in later steps:
Step 3: Correlating module eigengenes with the trait is enough. If the samples are independent, no linear mixed model is needed. Instead, the module eigengenes are calculated as before (first principal component of the module expression matrix). Then, for every module, this module eigengene is now correlated (again Pearson correlation) with the phenotypic trait, namely, the condition on whether the sample is tumorous or not, i.e.

$$|\text{cor}(\text{ME}_i, \text{Trait})|\,.$$

Again, "Trait" must be imagined to be a boolean vector, indicating with 1 and 0 whether the sample is tumorous or not. Note that the Pearson correlation is in this situation merely a standardised difference between groups. It is the analogue to the standard gene significance (explained above) on the level of modules. The closer this value is to 1, the more statistically significant is the module-trait association. These correlation values can also be supported with p-values (standard Pearson-correlation p-values computed using the Student's t test for correlation) to indicate their significance.

Step 4: Standard gene significance, Module membership and Intramodular connectivity to identify hub genes. In the last step, the aim is once again to find the hub genes.
Again, the linear mixed model for calculating the paired gene significance is omitted in favour of the standard gene significance sGS_i, where

$$\text{sGS}_i = |\text{cor}(x_i, \text{Trait})|$$

as explained above.
The other two measures, namely the module membership and the intra-module connectivity (resp. k-Within statistic), work exactly as for the paired data. Again, a gene is considered a hub gene if its standard gene significance and module membership are rather high. Again, the ordering depends on the k-Within statistic.
Note: Another common quantity when working with WGCNA is the module significance (MS) of a module m. It is defined to be the mean of all gene significances in the module, i.e.

$$\text{MS}_m = \text{mean}_{g \in m}(\text{GS}_g).$$

Also this is used as a measure for the module-trait association. However, as the above-introduced methods are preferred, in this paper, it was chosen not to include the module significance in the pipeline.

8 Data availability

The two datasets are available online by the National Library of Medicine. All code that was created for this study can be found in this public GitHub Repository: https://github.com/Smile4heWorld/WGCNA_comparison.git

References

[1] Jianqiang Li, Doudou Zhou, Weiliang Qiu, Yuliang Shi, Ji-Jiang Yang, Shi Chen, Qing Wang, and Hui Pan. Application of weighted gene co-expression network analysis for data from paired design. *Scientific Reports*, 8(1):622, 2018.