Data summary

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 10	Day 14	Number of samples	Number of available time points
EC_wild type	1	NA	3	3	1	1	1	1	2	1	14	9
EC_CCR2KO	1	1	1	1	NA	1	1	1	1	NA	8	8
FAP_wild type	1	1	3	3	2	1	1	1	1	1	15	10
FAP_CCR2KO	1	1	1	2	1	1	1	NA	1	NA	9	8
Muscle progenitors_wild type	NA	1	1	2	NA	1	NA	1	1	NA	7	6
Muscle progenitors_CCR2KO	1	NA	NA	2	1	1	1	NA	1	NA	7	6
Inflammatory cells_wild type	NA	2	3	3	1	1	1	2	1	NA	14	8
Inflammatory cells_CCR2KO	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0

Table 1. Data summary of skeletal muscle tissue gene expression

- Missing values (NA) are because the population could not be purified since there were only few of them.
- Green entries shows experiments with more replicates.

Quality control process for RNA-Seq data

Sample-sample correlation heatmap

We compute the samples correlation in a pairwise comparisons and then draw the heatmap in two different ways.
1. For a specific cell type, so we can see which days are more similar to each other and discover the potential outliers. Fig. 1 shows the heatmap of inflammatory cells wild type where day 6 seems very distinct from the other days. Replicates are usually grouped together, showing no potential outliers through replicates. Other days experiments which need to be considered are day 7 in EC KO and day 1 in FAP KO. Heatmaps are attached.
2. For all samples together, so that we could recognize which samples are more similar to each other. We would expect to see samples of a specific cell type to be grouped together in the dendrogram provided in the heatmap. As we can see in Fig. 2, all EC and FAP samples are grouped together. However, some of muscle progenitor samples and inflammatory cells samples are also grouped together, addressing similarities in such samples. The other point is that there are not much dissimilarities between knocked out and wild type since they are all in the same group according to the heatmap dendrogram.

Filtering low expressed gene

We filter lowly expressed genes to reduce the effect of experimental noise. We look at the distribution of gene expression across all samples through all cell types. To choose the cutoff value, we assume that the distribution will be following the mixture of two normal distributions, one for lowly expressed genes and the other one for the other genes. As shown in Fig. 3, we choose the cutoff value around 0.7 (log scale) where genes are following the second Gaussian distribution. Then we keep only genes which are expressed above this threshold (0.7) in at least 50% of the experiments. We start from 25K genes and end up with around 11K genes with this criteria.

Quantile normalization

We perform quantile normalization so that we make the distribution of genes across different samples to be identical.

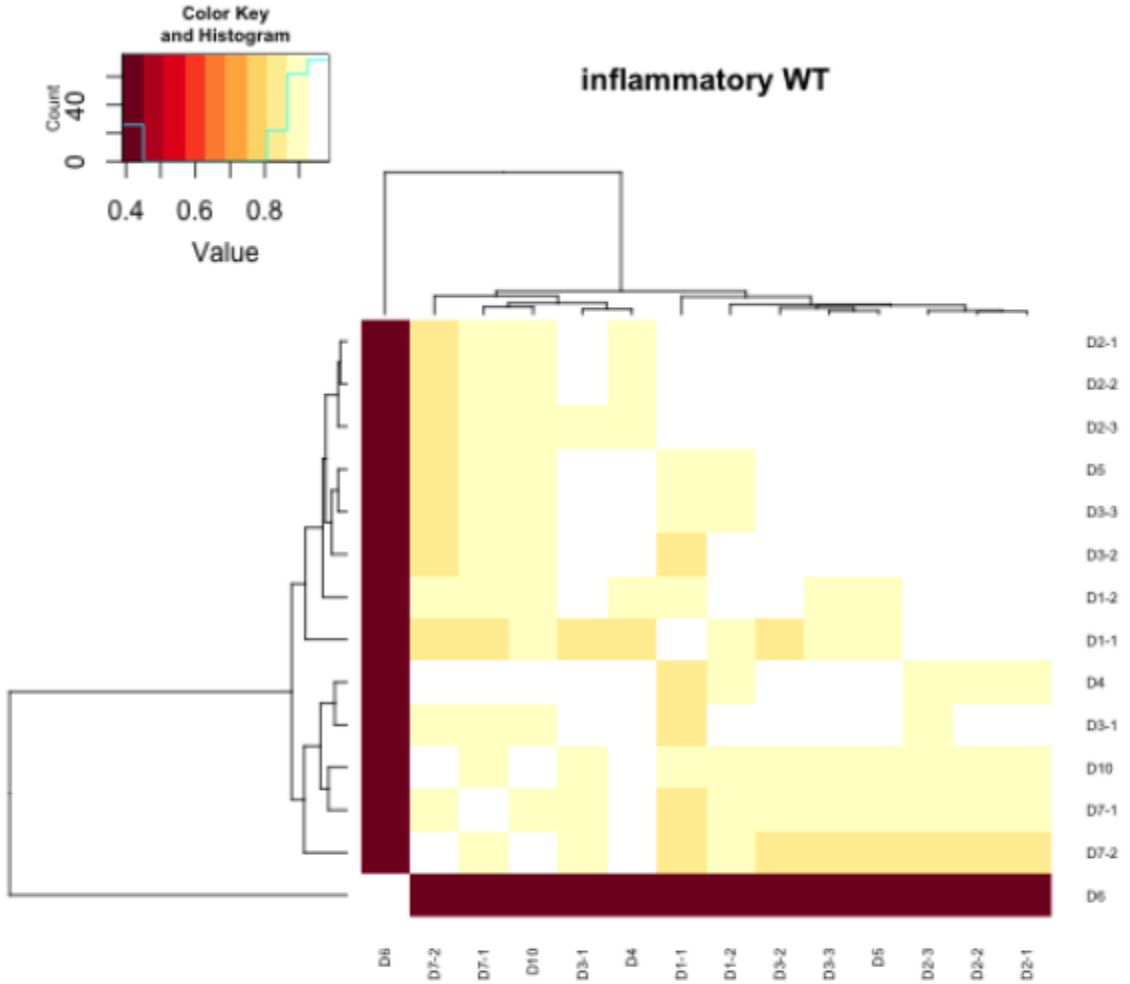


Figure 1. Sample-sample correlation heatmap for Inflammatory cells wild type

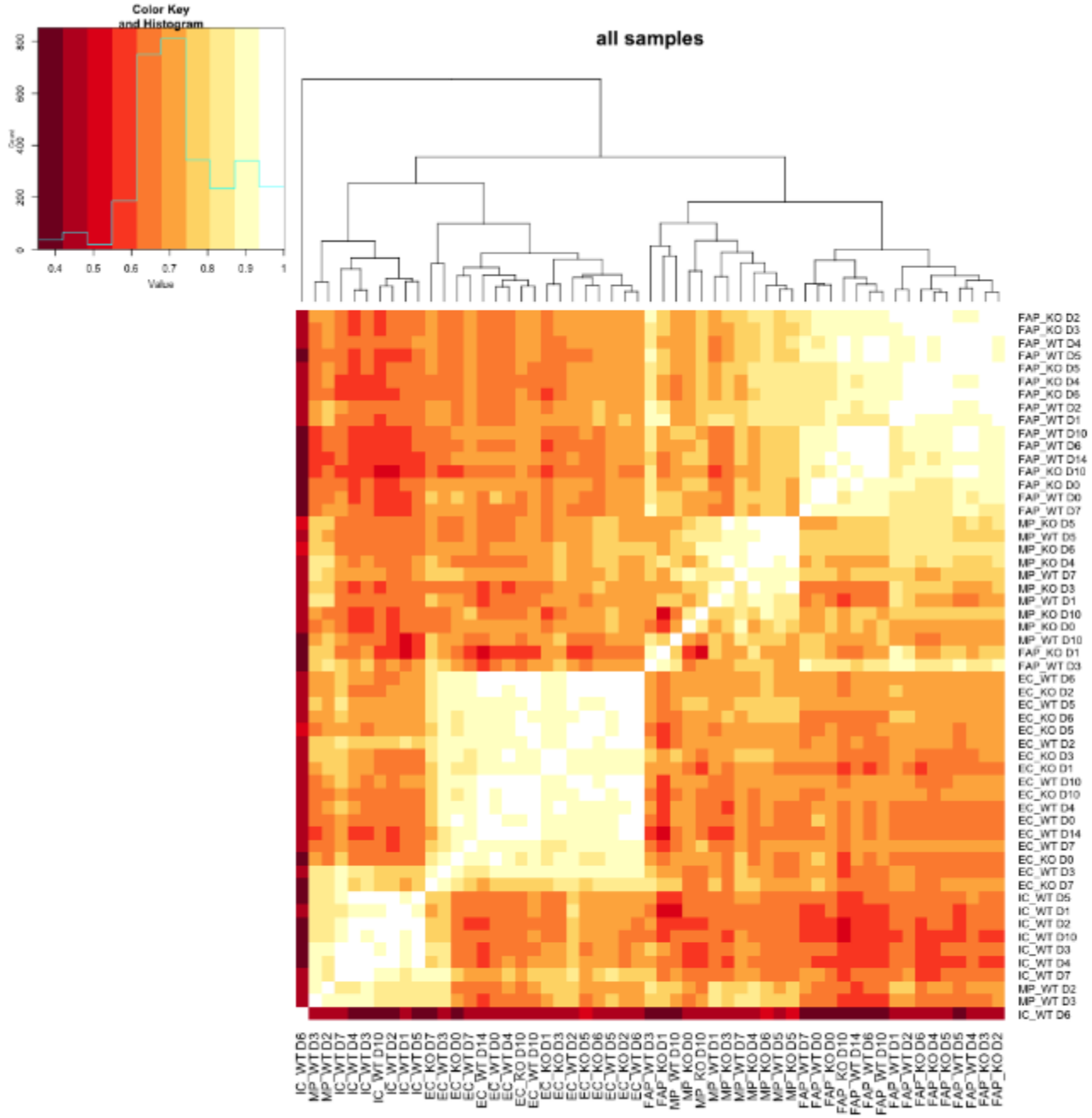


Figure 2. Sample-sample correlation heatmap of all samples

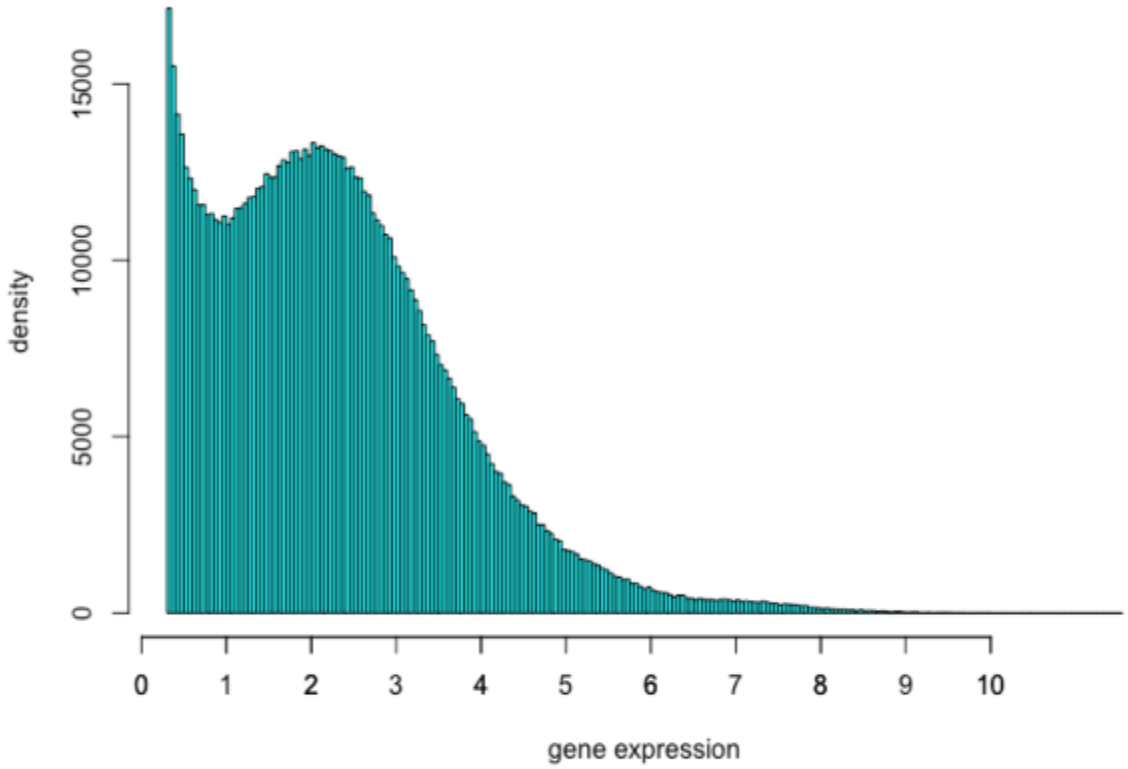


Figure 3. Gene expression distribution of all samples

Imputing missing day experiments

According to the primary goal of this project, identifying cell-to-cell interactions through co-expressed analysis, we are interested in discovering genes with highly similar temporal patterns. Assessing co-expression analysis based on limited number of time points (6-10) is hard. So we come up with the idea of imputing missing day experiments. We model each expression profile as a cubic spline (piecewise polynomial) using B-Spline basis function that is estimated from the observed data. Through this, we get smoothing/approximating splines rather than interpolating splines. In the ideal case, the smoothing model will help us predict the unobserved day experiments as well as re-estimating the observed day experiments so that we could reduce the noise. We use four control points and estimate the coefficients by solving D = SC in the least-squares sense. D, S and C are data points, B-Spline matrix and control points, respectively. To evaluate the smoothing splines, we leave one experiment day out and fit the smoothing spline based on the other available days and calculate R2 scores. As it is clear in Table 2, which is the model evaluation for FAP KO, predictions are far from the observed data. Two challenges come up are: a) models are overfitting due to complex temporal patterns b) removing key time points where peak/dip is occurring screws up everything. Overall, due to insufficient number of time points, we can not have powerful smoothing models and solving this problem itself, requires more time points (more time points between two consecutive days of experiment).

Day0	Day1	Day2	Day3	Day4	Day5	Day6	Day10
-37	0.14	-8.9e3	-2.5e3	-1.3e4	-2.9e3	-5.2e4	-5.7e7

Table 2. R^2 score of predicting missing day experiments for FAP knocked out

Cluster/module analysis

Going back to the cell-to-cell interaction, a naive approach would take into account gene-to-gene co-expressions which means considering all expressed genes from cell A to all expressed genes to another cell, B. This analysis would require a huge sample size (time points in our case) to have enough statistical power to find significant gene-to-gene co-expression signals. The possible solution is reducing the dimensionality of the problem by clustering the genes or narrowing genes down to the known pathways. In this way, we do cluster-to-cluster co-expression analysis, assessing which clusters from cell A are interacting with which clusters from cell B.

We perform weighted gene co-expression network analysis (WGCNA) on each of the cells to find the highly correlated genes (modules). To construct a network, it calculates the Pearson correlation of all pairs of genes in the network. Then it weights the correlations by raising the absolute values to the power of β , struggling with the noise of such data, which would empower the strong correlations and punish the weak ones. The power β, is determined according to the scale free topology criterion. Afterwards, it finds the modules through hierarchical clustering.

Module preservation across cell types or cell conditions

Once we have the modules, we look for how well a specific module M from cell A is preserved in cell B which means that how well genes involved in M are co-expressed in cell B. We perform this analysis for cell conditions (wild type vs CCR2 knocked out) by module preservation statistics provided in WGCNA. This helps us to know once we knock out CCR2, how the gene co-expression network and thus the modules would change. The modules that are not preserved are interesting since they might be responsible for specific functions that would be affected by knocking out CCR2 gene. We test the module preservation in two ways, one is when the modules are found in WT network and the reference network is KO, so examining if the modules in WT condition are preserved in KO condition. The other one is the opposite way. For muscle progenitors, we can not find a scale free network even for very big βs. Results of preservations are for EC and FAP are shown in through Fig. 4 to Fig. 7. Each circle corresponds to a cluster. Clusters above the green dashed line, between green and blue dashed line and below the blue dashed line show strong, mild and no preservation, respectively. As we can see in Fig. 4 and Fig. 5, we find fewer modules in KO than in WT. Furthermore, more number of modules found in KO are preserved than modules found in KO. We can see in Fig. 6 and Fig. 7 that most of the modules of FAP are preserved across WT and KO.

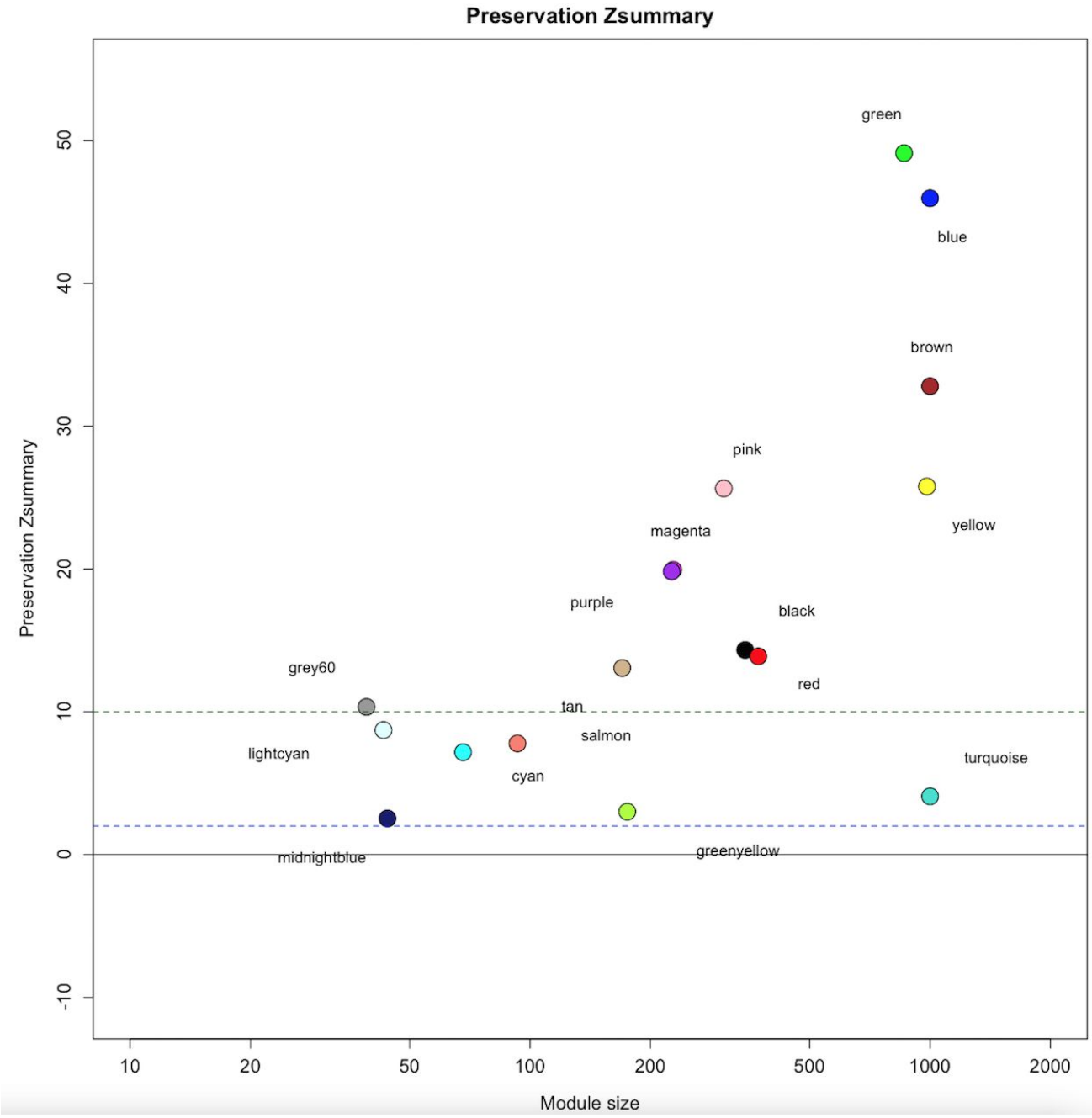


Figure 4. Preservation of EC KO modules in EC wild type co-expression network

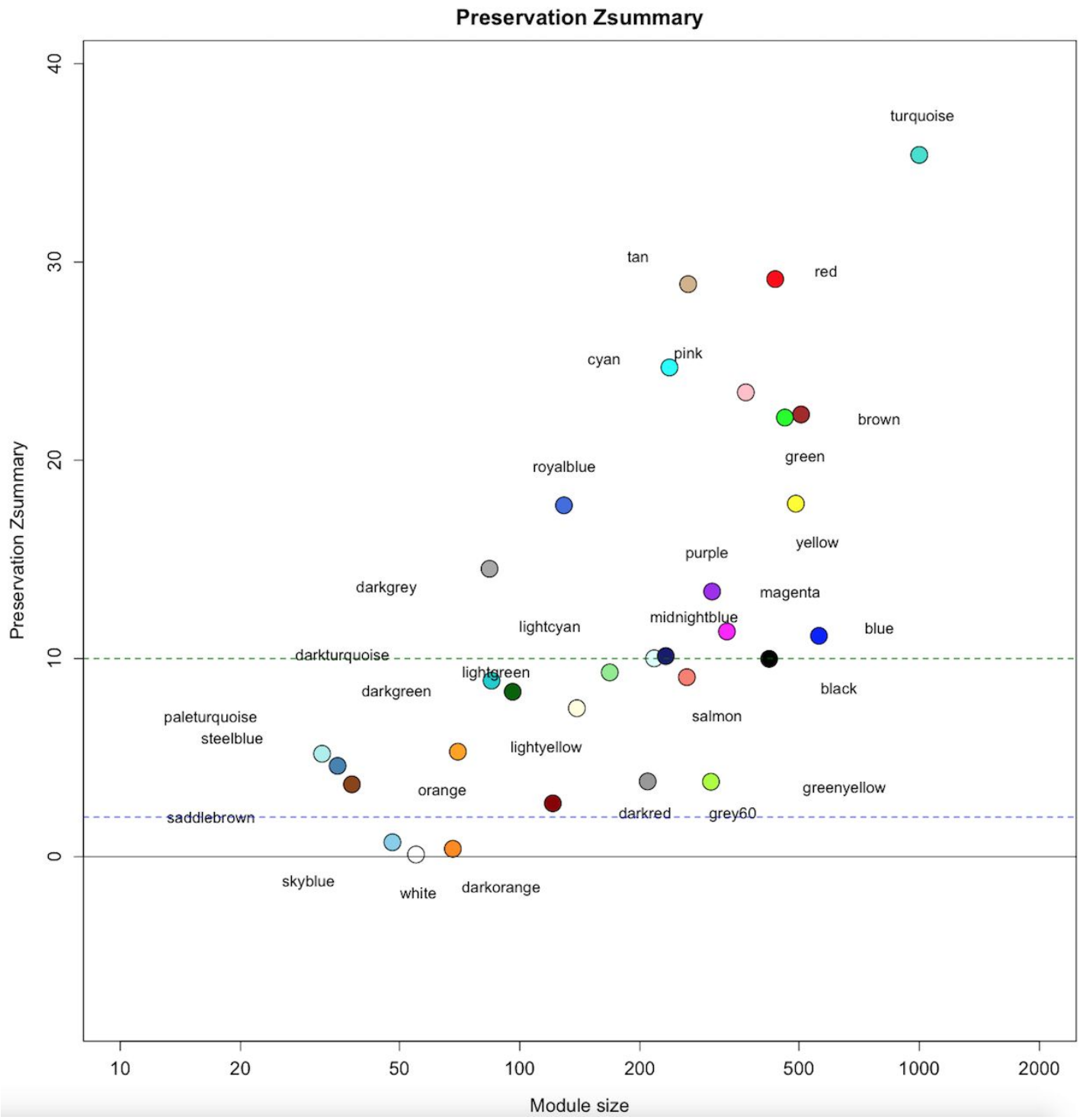


Figure 5. Preservation of EC wild type modules in EC KO co-expression network

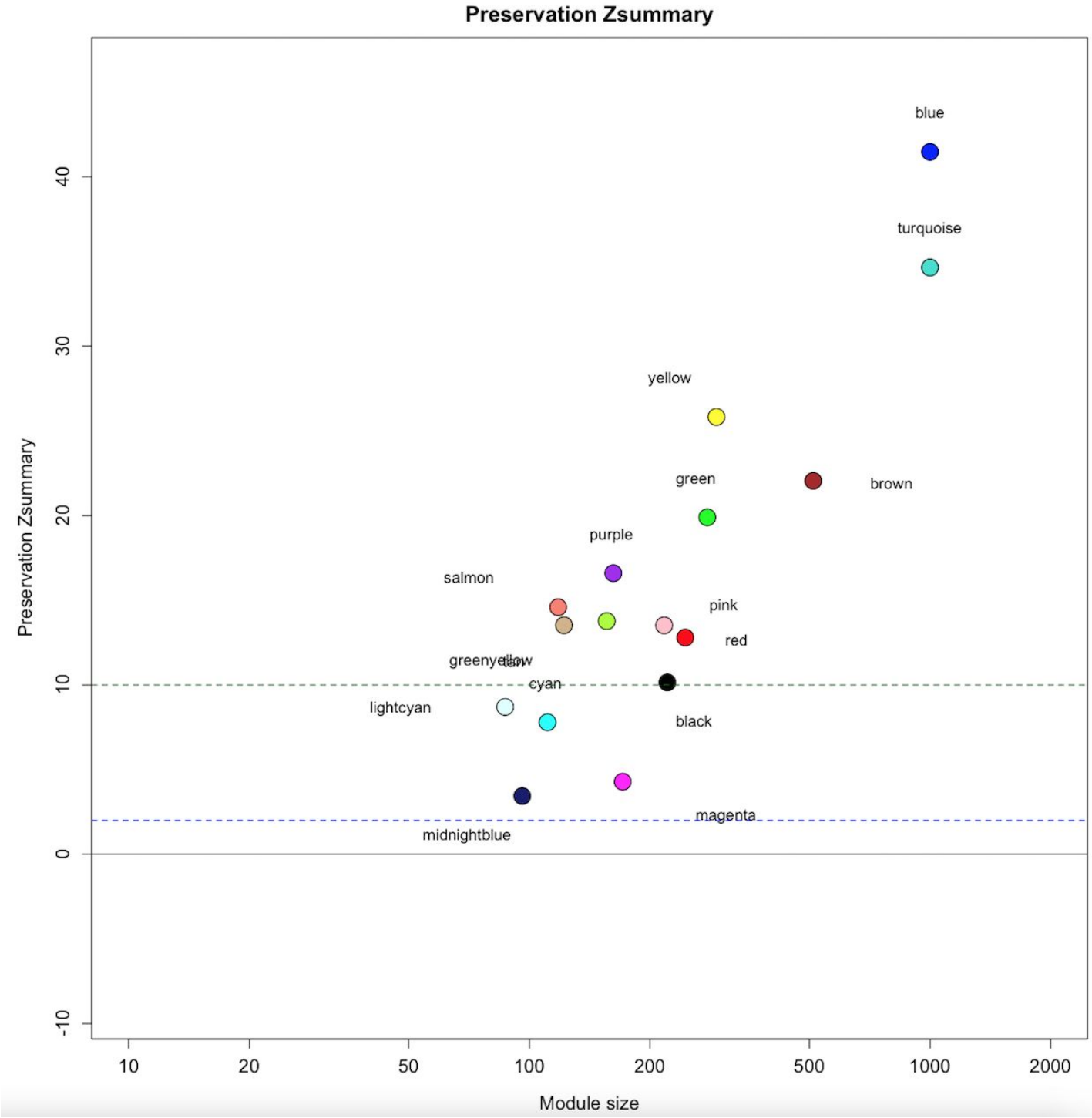


Figure 6. Preservation of FAP KO modules in FAP wild type co-expression network

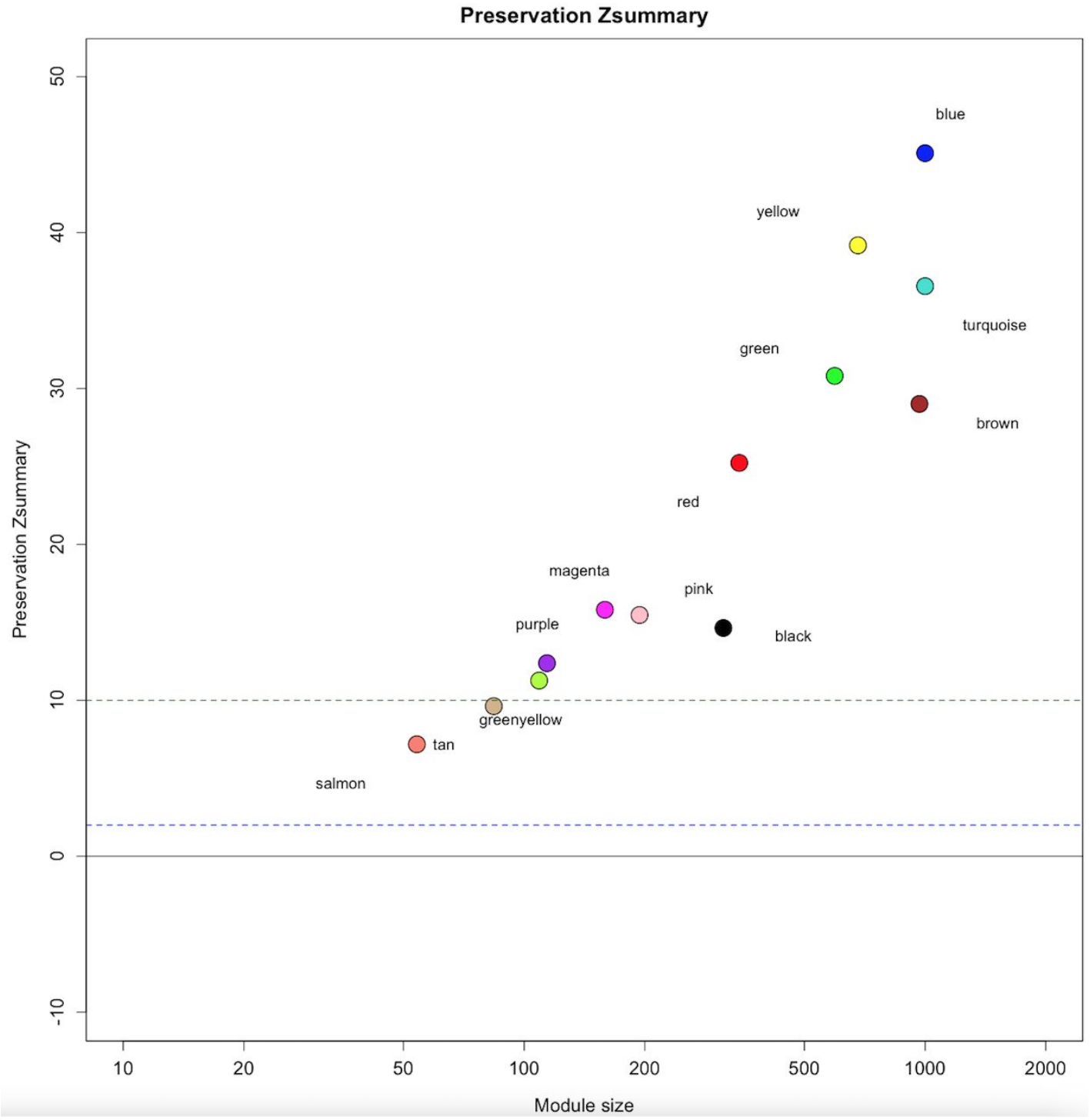


Figure 7. Preservation of FAP wild type modules in FAP KO co-expression network

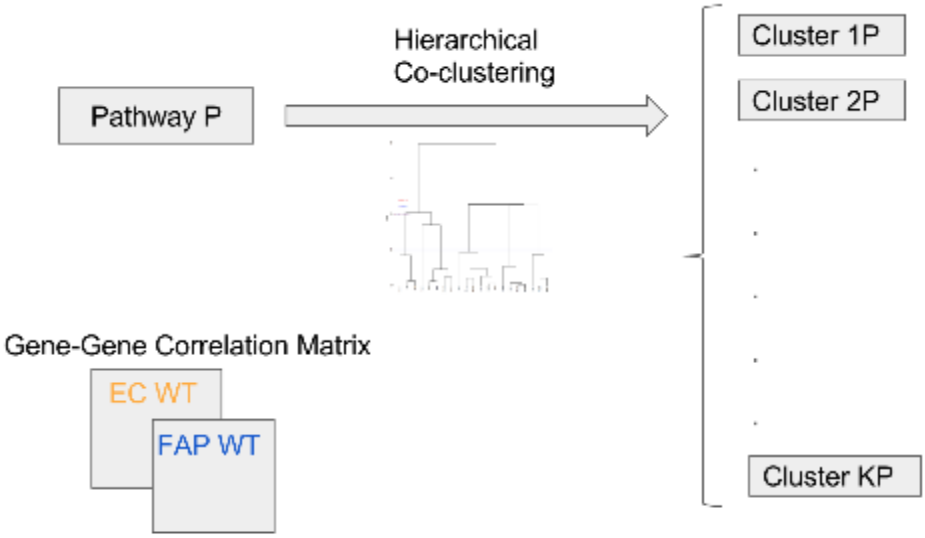
Cell-to-cell communication

We start from known receptor to ligand associations. If we take into account only ligand and receptor expression, we might not find signals since receptor activations does not necessarily change its expression profile. So in ideal case, we focus on the downstream effector of the receptor and try to find ligand receptor associations by performing co-expression analysis of ligands and downstream effector. The method of this part consists of three sections: a) starting from a receptor, get all the KEGG pathways that contain the receptor b) For each pathway, we perform hierarchical co-clustering and cluster the genes in the pathway. Co-clustering means that the gene-to-gene correlation matrix which is used as the similarity metric in hierarchical clustering, is calculated by averaging over such matrices across different cell types. This would lead the clusters to be co-expressed in each of the cell types separately. The current results are based on EC and FAP wild type. I cut the dendrogram at height 0.65. c) The clusters of a specific pathway (which also corresponds to a specific receptor) are analyzed if they are interacting with the corresponding ligand. The interactions are defined as the ligand and the cluster are correlated above the threshold (60%) on average. Fig. 8 shows different parts of ligand and downstream effector associations.

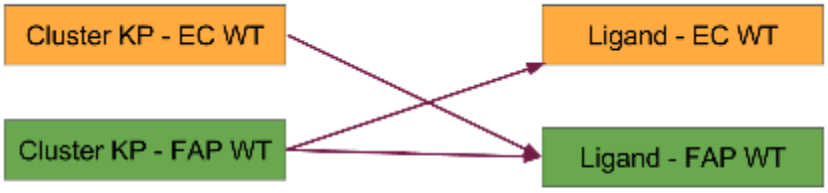
The table_clustering_atleast3.csv reports all the associations between ligand and cluster of size greater than 2 genes in downstream effector. The ‘score’ column shows the average correlation of ligand with the cluster. The patterns_atLeast10genes.pdf shows the temporal patterns of genes in such associations. I plotted only for the clusters with more than 10 genes. The two plots on the left side show the receptor downstream genes and ligand expression from up to bottom, respectively, in the cells they are associated together. The plot on the right side shows the receptor downstream genes in the other cell. As an example, in the first plot we can see that there is a cluster of 16 genes in the downstream effector of Itga9 receptor in EC wild type which are associated to the Vegfa ligand in EC wild type. The right plot shows the same 16 genes in the other cell, FAP wild type. Please note that genes are co-expressed in both cell types.

Figure 8. Cell-to-cell communication method

Part a) Get KEGG pathways for a receptor



Part b) co-clustering genes in pathways



Part c) finding ligand and downstream effector association

We start from 174 different KEGG pathways which cover 301 receptors. After clustering, we keep only clusters with more than two genes. Finally, the associations are found between 96 receptors and 57 ligands. Such receptors were associated with 139 KEGG pathways. Fig.9 and Fig.10 show the histogram of number of clusters per pathway (with mean 11) and number of genes per pathway (with mean 70 which are the genes after all quality control steps in our data), respectively.

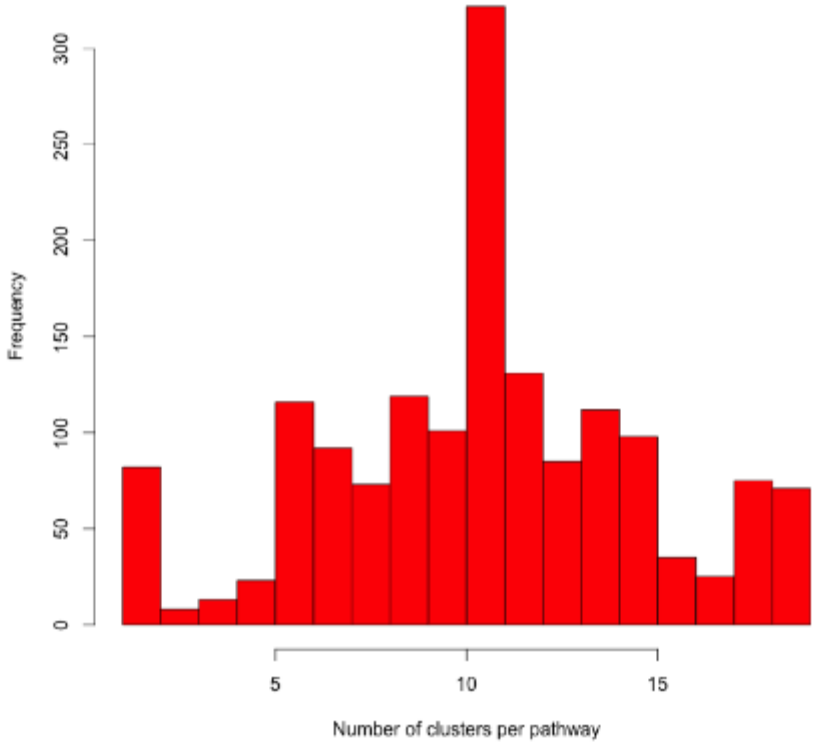


Figure 9. Histogram of number of clusters per pathway

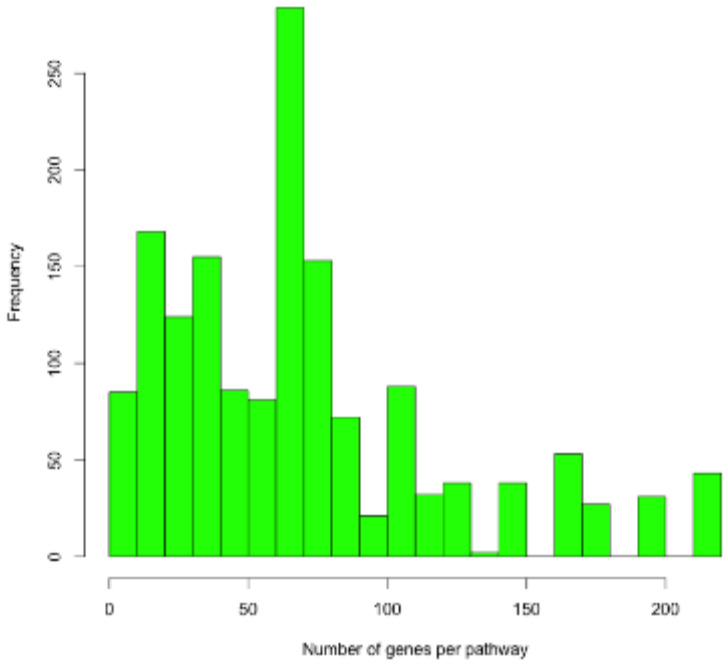


Figure 10. Histogram of number of genes per pathway