Gene expression variability and the prediction of clinical outcome in Chronic Lymphocytic Leukemia

Simone Ecker[*], Vera Pancaldi[*§], Daniel Rico[†] & Alfonso Valencia[†]

[*]These authors contributed equally to this work

[†]These authors contributed equally to this work

[§]Corresponding author

Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Calle Melchor Fernandez Almagro 3, 28029, Madrid, Spain

Emails: secker@cnio.es; vpancaldi@cnio.es; drico@cnio.es; avalencia@cnio.es

Abstract

Background: Chronic Lymphocytic Leukemia (CLL) presents two subtypes which have drastically different clinical outcomes. So far, these two subtypes are not associated to clear differences in gene expression profiles. Interestingly, recent results have highlighted important roles for heterogeneity, both at the genetic and at the epigenetic level in CLL progression.

Results: We propose to use gene expression variability across patients to investigate differences between the two CLL subtypes. We find that the most aggressive type of this disease shows higher variability of gene expression across patients and we elaborate on this observation to produce a method that classifies patients into clinical subtypes. Finally, we find that, overall, genes that show higher variability in the aggressive subtype are related to cell-cycle development and inter-cellular communication, probably related to faster progression of this disease subtype. We also find a non-trivial relationship between expression, methylation and gene expression variability, which we consider as a possible mechanistic basis for the apparent lack of a clear correlation between expression and promoter methylation reported in the literature.

Conclusions: We propose that gene expression variability should be considered on par with genetic variability in studying the link between tumour heterogeneity and phenotypes such as aggressiveness and resistance to therapy in CLL. Moreover, we start to outline interesting but complex relationships between the genotype, epigenotype and phenotype of potential biological and clinical consequences.

Keywords

Chronic Lymphocytic Leukemia, gene expression variability, noise, methylation, population/cellular heterogeneity, personalized medicine

Background

One of the outstanding challenges in biology is elucidating the relationship between genome, epigenome and phenotype. Notwithstanding the considerable progress that has been made in terms of mapping the epigenetic state of cells along with their transcriptome, it has often been hard to see the interdependencies between the two and their joint contribution to cellular behaviour. We are just starting to unravel the different genetic and non-genetic factors that are responsible for the incredible variability of phenotypes that can be observed in a population of cells.

Biological noise is emerging as an important factor influencing the phenotypic variability in cell populations. The first experiments measuring fluorescence of reporters in single bacteria [1] highlighted the presence of various sources of 'noise' that would contribute to the variability observed. Intrinsic noise, which is inherently caused by stochasticity in the biochemical phenomena that lead to gene transcription and affects each gene independently, and extrinsic noise, which causes fluctuations in the value of expression correlated amongst multiple genes [1]. In fact, biological phenomena are governed by randomness just like other physical systems on the small scale. For example, the production of mRNA happens in bursts whose regulation in size and frequency can control not only the average amount of RNA produced, but also the fluctuations in this value [2].

Recently, single-cell methods in yeast and mammalian systems have studied noise and cell-to-cell variability, which is now recognized to be at the basis of many interesting biological processes, for example p53 oscillations [3] and NF-kB pulses of localization in the nucleus [4, 5]. The gene expression variability at the single cell level is probably having an effect on the variability across different organisms in a population. Indeed, a strong correspondence between expression variability due to stochastic processes in single cells from the same population and variability of gene expression

in a population measured across different conditions is commonly observed. Multiple experimental investigations of this relationship have led to accept that common mechanisms are probably responsible for the two different types of variability of gene expression, connecting variability in a population to variability across a time courses [6, 7]. The conclusion from these studies is that variability across conditions in a time course, between different individuals that have slightly different genetic backgrounds, and variability in single cells of the same isogenic population are strongly related. This allows us to measure variability of one type and infer them to the other types.

It is therefore fair to ask what regulates the weaker or stronger propensity of a gene to be regulated, both in terms of plasticity in different conditions and in terms of stochastic noise. It is widely recognized that specific promoter structures (TATA boxes) are found mainly in genes with functions related to the response to external stimuli, which are also genes that usually have widely fluctuating single-cell levels within populations [8–10]. The characteristics and dynamics of regulation are very likely related to the chromatin structure in the region of the promoter of the gene and, more specifically, to the nucleosome distribution [11].

This biological observation is reminiscent of a widely accepted concept in physics which goes under the name of 'fluctuation dissipation theorem' and states that quantities that are observed to stochastically fluctuate on a large scale are also likely to have large responses to a stimulus, whereas quantities that have limited stochastic fluctuations will have smaller responses to the same size stimulus [12]. The analogy with gene expression would suggest that when a gene needs to undergo large changes in its levels in response to signaling, for example, it will be easier to achieve that level if the gene already displays large stochastic fluctuations in the absence of the stimulus.

It is well known that tumours show increased heterogeneity compared to normal tissue [13–15].The presence of heterogeneity in tumours is furthermore known to affect aggressiveness and resistance to therapy[16, 17], but is traditionally investigated in solid malignancies, which can present a very diverse population of clones. However, even haematologic diseases, which are thought to arise from clonal populations, can display a degree of genetic and non-genetic heterogeneity [18].

In this work we will focus on gene expression variability between individuals. Variability of gene expression has been suggested to be an important parameter to be measured alongside the average levels of gene expression [19, 20]. However, comparing variability in two or more different groups of samples is not trivial. A single measure can easily be obtained for each group but, statistically, it would be desirable to produce multiple estimates of the measure per group and compare them across groups. We have developed a method that aggregates samples, allowing for estimations of fluctuations and comparisons across groups.

We focus on two datasets of chronic lymphocytic leukemia (CLL) - a B-cell neoplasm - in which gene expression was measured for large cohorts of patients in two independent datasets [21, 22] for which clinical data was also available. Two major subtypes in CLL are defined by the mutational status of the immunoglobulin heavy chain variable region (IgVH). The unmutated subtype (U-CLL) is characterized by being more aggressive than the mutated subtype (M-CLL), which shows better prognosis and overall survival. The Kulis study showed that although there are significant differences in the methylome of M-CLL and U-CLL, a strong correspondence with gene expression levels was not found [22]. A subsequent study [23], which extensively characterized the transcriptome of CLL using RNA sequencing, revealed two new subtypes of the disease, completely independent of the well characterized clinical subgroups based on the mutational status of the IgVH. This further demonstrated

that the two important clinical subtypes M-CLL and U-CLL do not seem to be reflected by gene expression levels.

Interestingly, applying our method to the data provided by these two previous studies, we find significant differences in the variability of gene expression between M-CLL and U-CLL. The more aggressive U-CLL subtype exhibits increased expression variability. Even more strikingly, we show that a classification based on gene expression variability correctly classifies the patients into the two subtypes of CLL.

In this work we demonstrate that there are strong relations between disease subtype and gene expression variability, and we suggest a non-trivial correlation with DNA methylation as a possible source of the observed expression variability.

Results and discussion

Gene expression levels and gene expression variability in CLL

To quantify the level of variability of tumour samples in a Chronic Lymphocytic Leukemia (CLL) patient cohort [22, 23] we study variability in terms of the Coefficient of Variation (CV). CV is defined as the ratio between the standard deviation of the variable measured across the patients and its mean. As gene expression variability is dependent on the gene expression levels, we analyze the dependence of the CV on the level of expression of the corresponding genes. With this purpose, we bin all genes into 10 equally sized classes based on gene expression levels (see Methods) and calculate for all genes in each bin the CV of expression across CLL patients (Figure 1). The relationship between CV and expression level is interesting and non-trivial. The highest levels of expression variability across patients is observed for genes with intermediate levels of expression and not for genes expressed at

high or low levels (Figure 1A). To understand the origin of this behavior is important to take into account the intrinsic stochasticity of biological processes, which will introduce fluctuations of a size inversely proportional to the number of elements involved in the system. This is a well established phenomenon observed in physical systems [24] and well characterized in biology [25]. Indeed, there is a component of CV that is given by the inverse of the mean of expression as a 1/x dependence on expression levels (Figure 1B). This dependence reflects the fact that introducing one additional element in a small collection (i.e. an extra copy of mRNA of a lowly expressed gene) will have dramatic consequences. In contrast, an extra copy of a transcript that is in high numbers will not produce a substantial change. Stochastic processes of this kind are likely to not be the sole determinants of the CV. The remaining component of the CV is given by the standard deviation of expression, which has a negative quadratic dependence on expression (Figure 1C) showing higher values for intermediate expression classes. That is, only the genes expressed at intermediate levels have a significantly high level of variability in their levels of expression. These observations highlight the importance of taking gene expression levels into account when evaluating gene expression variability in tumoral cells.

In the following and based on this initial observation, we suggest that expression variability can be subdivided into a 'noise' component, deriving from the inherent stochasticity of biological processes, and an additional component produced by a biological process that regulates variability. The 'noise' component of variability is expected to decrease as expression levels increase and to affect all genes and all individuals to the same extent. On the contrary, the additional biological process, which is regulated by the cell, should affect important genes to a lesser extent and will not dependent on the gene expression levels. This interpretation is in line with previous publications [26–29].

We reasoned, that the potential impact of this biological process on variability will be different for gene

classes with different functions. For example, we would expect housekeeping genes to be maintained under stricter control and hence to show lower 'regulated' variability compared to non-housekeeping genes, whereas the 'noise' component would be constant for all genes within a specific expression range. Indeed, we found that housekeeping genes have a generally lower CV than non-housekeeping genes, and that this difference is larger precisely in intermediate gene expression ranges (Figure 2). This result suggests a strong control of the expression of housekeeping genes, that might minimize variability to not compromise basic cellular processes. The lower variability of this set of genes which require steady expression throughout different conditions reinforces our hypothesis that the 'regulated' component of the variability may be of biological origin and it is probably influenced by natural selection.

Therefore, we propose that the expression variability in CLL patients can be explained by separating the variability into a 'noise' component, common to all genes, and a 'regulated' component of variability more affected by the processes in which the gene takes part. In the following we investigate whether differential gene expression variability could be behind the differential aggressiveness of the CLL clinical subtypes, as an example of a biomedically relevant cellular phenotype.

Inter-patient gene expression variability is different in two clinical subtypes of CLL with different prognosis

CLL's prognosis varies greatly depending on several factors, most importantly the maturation state of the leukemic cells, as monitored by the gene mutation status of the immunoglobulin variable-region heavy chain (IgVH). Patients showing fewer mutations in this region, defined as U-CLL (“unmutated”

CLL), have a worse prognosis compared to M-CLL ("mutated" CLL) patients, who show a larger number of mutations in the IgVH gene region. The presence of extended mutation is thought to indicate a leukemia originating from more mature lymphocytes [22]. The distinction between U-CLL and M-CLL divides CLL cases into groups of different aggressiveness, so we wondered whether there could be a difference in the gene expression variability of the two subtypes.

Consistent with our hypothesis, gene expression variability measured by the CV shows a clear difference between the two subtypes (Figure 3A) with higher variability associated to U-CLL, the more aggressive disease. On the contrary, the gene expression levels of U-CLL and M-CLL patients showed very little difference (Figure 3B), in agreement with previous reports [30]. These results suggest that expression variability across patients can be an important factor to discriminate the two disease sub-type, for which the general level of expression will not be discriminatory and very few differentially expressed genes have been identified with standard cutoffs on fold change [23]. In the context introduced above, the 'noise' component of variability is not expected to change between the two disease subtypes, as stochasticity due to low numbers effects and technical noise are a constant factor in the two patient types. In contrast, the 'regulated' biological variability component might be different in the two types. If this is the case, the prediction will be that major differences in CV would be observed for intermediate to highly expressed genes, as, indeed, is the case. Figure 3C shows a stronger difference in variability between U-CLL and M-CLL for genes that are more expressed. However, as shown in Figure 3A these most differentially variable genes have an intermediate level of expression.

Identification of genes that increase variability in U-CLL compared to M-CLL

As shown in Figure 3A, a substantial number of genes display higher variability across U-CLL patients compared to M-CLL patients. 2025 genes are found to have a higher variance in U-CLL whereas only 360 are significantly less variable (F-test, FDR=0.05, see Additional File 1, Supplementary Figure 1 and Supplementary Table 1). Repeating, the analysis with a different dataset by Fabris et al. [21], we confirm the increased variability in U-CLL patients (see Additional File 1, supplementary Table 1) and we see a very strong correlation between the CV of the CLL subtypes in the two patient cohorts(Pearson correlation: M-CLL r=0.67 and U-CLL, r=0.66, p-values$<10^{-16}$). Also the differences between CV values for genes in the Fabris et al. [21] and Kulis et al. [22] cohorts are significantly correlated (Pearson r=0.25, $p<10^{-16}$) as well as those for the standard deviation (Pearson 0.72, $p<10^{-16}$). When we took the top 500 genes that show highest changes of variability in U-CLL in each dataset, we found an overlap of 126 genes are found to be in both lists, a number which is significantly higher than expected at random (Hypergeometric test, p-value 5.3 10^{-67}). Therefore, our results are reproducible in the two datasets and cannot be attributable to batch effects.

Functional analysis of variable genes

We have stated that genes that show the highest increase in variability between U-CLL and M-CLL tend to be highly expressed but we haven't so far commented on whether specific functional categories are particularly affected. If we assume that most of the differential variability between the two subtypes is due to a biological process, we would expect specific functional classes of genes to be most affected. Looking at the 500 genes that increase their variability most in the U-CLL patients of the Kulis et al. [22] study (Additional File X1) we observe a very significant enrichment for processes related to the cell cycle, wounding, response to stimulus and multi-cellular processes and development and an enrichment for localization to the plasma membrane (Additional File X2). Performing the same analysis in the Fabris et al. Dataset [21] we recapitulate these results to a certain extent, finding enrichments in the immune system (Additional File X2), and - although not reaching statistical

significance with an FDR of 5% - haemopoiesis, differentiation and proliferation.

To investigate if there are functions that are reproducibly more variable in U-CLL, we performed an enrichment analysis of the genes that are in the top 500 genes with highest changes in CV between U-CLL and M-CLL in both studies (126 genes) and we confirmed enrichment in haemopoiesis, immune system, development and proliferation (Additional File X2). To further understand the functional context of these differentially variable genes we used a B-cell specific functional interaction network [31] and extracted the nearest neighbours of these 126 genes with increased variability in U-CLL. As a result, we identified a subnetwork of 329 genes with increased variability in UCLL and their neighbours. Figure 4 shows the network with nine highlighted subnetwork modules (Louvain method [32]. We then performed functional enrichment analysis of the different network modules identified (Additional data file X3).

Module 0 is enriched for localization to the plasma membrane and regulation of cell communication (Additional File X3), where the most connected genes are *PRKCA,* a kinase involved in cell adhesion, cell transformation, cell cycle checkpoint and cell volume control which is known to act as an anti-apoptotic agent by phosphorylating BCL2 in leukemic cells [33], and *CAV1* which is strongly related to signal transduction and able to affect cell function and cell fate [34, 35], and has furthermore been described to play a significant role in CLL progression [36]. Within module 1, which is strongly enriched for localization in the extra-cellular region (Additional File X3), the biggest hub is *MAF,* a transcription factor of the *AP1* family that has been shown to be involved in the development of multiple myeloma - a malignancy of mature B cells - and is also known to enhance pathological interactions between tumour cells and the stroma [37]. Module number 4 does not have genes with particularly high degree but is heavily enriched for functions related to the cell cycle, with genes like *CDK1, CDK2* and *CDK2AP1* (Cyclin-Dependent Kinases), *E2F4* (E2F Transcription Factor 4), *RB1*

(Retinoblastoma 1) and *TP53* (Tumor Protein P53). (Additional File X3). *MAP4K4*, a *MAP* kinase known to be deregulated in various lymphomas and particularly B cell leukemias [38] is heavily connected in the network within module 7, which is enriched in signal transduction through post-translational modifications and includes multiple members of the *MAPK* cascades (Additional File X3). Another hub identified in the network is the hypoxia inducible transcription factor *HIF1,* which is known to be frequently overexpressed in CLL [39, 40].

A possible interpretation of this data is that U-CLL patients show increased variability in proliferation rate (which would be also directly affected by the cell cycle regulation genes) and in their intercellular communication in the context of multicellular organisms and development. It is also possible that U-CLL samples would show increased variability in their developmental stage, indicating the likely presence of cells at different steps of differentiation. Increased proliferation rate heterogeneity in U-CLL compared to M-CLL could be impacting this disease subtype's aggressiveness and adaptability, possibly explaining U-CLL's worse clinical outcome.

Unsupervised clustering of the two CLL subtypes and patient classification using information on expression variability.

The previous results suggest that eliminating the stochastic component of variability should allow us to highlight the regulated component and, hence, to better distinguish the two clinical subtypes. As mentioned in Kulis et al., gene expression alone is not sufficient to cluster patients into the two classes (Figure 5A).

To exploit information about expression variability it would be ideal to have a significant number of

measurements for every patient. Instead of that, we propose an alternative procedure based on aggregating patients into groups and computing the variability and mean of gene expression within each group. We call these groups of patients 'superpatients', representing constructs that help to de-noise the data and to obtain surrogates of variability values, measured as the relative range of expression (see Methods). The results of clustering 'superpatients' are represented in Figure 5B,C.

We can successfully separate the U-CLL and M-CLL 'superpatients' by clustering, both using the average gene expression and the variability measure. These results are a substantial improvement on the separation between the two subtypes that was obtained using expression data in Kulis et al. [22]. We attribute this improvement in the classification to the de-noising produced on the expression data through averaging. As suggested by our results on differences of expression variability between the two clinical types, expression variability alone can classify the two subtypes surprisingly well and with a similar performance to the expression average.

Beyond the classification of 'superpatients' the ultimate challenge is the classification of single patients into the two clinical subtypes using expression variation data. We trained a Random Forest classifier [41] with the Kulis et al. data [22], first using expression levels for all genes as features and achieving an error rate of 5.74% (see Methods). Since we have observed that expression variability is an important feature in distinguishing the two patient types we then reduced the feature set to only the top 500 most differentially variable gene expression array probes (between U-CLL and M-CLL) and improved the classification error rate to 3.67%, corresponding to only 2 patients of each type being miss-classified.

Interestingly, we can use our classifier trained on the Kulis et al. dataset [22] to predict the CLL subtypes of patients in the Fabris et al. dataset [21], achieving an AUC of 0.84 (See Additional File 1, Supplementary Figure 2).

Finally, inspired by our results on the importance of the variability of gene expression as a defining characteristic of the two CLL subtypes, we decided to define a measure for each patient and each gene that could be a proxy for the CV of expression which we would only be able to measure for a group of patients. To this end we defined the absolute distance of a gene's expression value from the median of expression of that gene across all patients (see Methods). Using these new features we improve the classification of the Fabris et al. patients based on the Kulis trained classifier to an AUC to 0.91 (see Additional File 1, Supplementary Figure 2).

DNA methylation and the origin of expression variability

Although gene expression and promoter methylation of different genes in the same individual are known to be inversely correlated, only (very) low levels of correlation are observed when comparing the same gene across individuals [42]. Moreover, recent research has highlighted the importance of methylation heterogeneity and its evolution in the progression of CLL [43]. We therefore investigated the relationship between methylation and expression CV.

To start with, few genes show substantial differences between levels of methylation in the two subtypes as shown by hierarchical clustering of the Kulis et al. [22] methylation data (Figure 6A) and by more direct comparison of the promoter methylation values (Figure 6B). As shown in Figure 6C, we recover a negative dependence between expression and promoter methylation, but only for reasonably low levels of expression. This finding suggests that methylation can affect the turning on and off of gene expression, and it could be potentially related with the observed differences in gene expression variability between individuals, while it plays little or no role in regulating the levels of expression.

Indeed, an interesting non-linear relationship is observed between CV and methylation (Figure 6D). Again, at this level we can separate the two components of expression variability. The CV generally tends to be higher in genes whose promoters are non-methylated, therefore more highly expressed, possibly reflecting the 'regulated' component. The CV increases again for highly methylated promoters, probably due to the 'noise' component which dominates for lowly expressed genes. (Gene body methylation shows a similar tendency, but in this case only a few genes are found in the bins that show high expression, see Additional File 1, Supplementary Figure 3 and few genes show differences in methylation between the two subtypes, see Additional File 1, Supplementary Figure 4). When looking more specifically at the most variable genes, we find the promoter methylation to be directly proportional to the level of variability, something which is lost when these most variable genes are grouped into very large bins (see Additional File 1, Supplementary Figure 5). Since we are interested in the regulated component of variability, and not in the noise that is present for low levels of gene expression, we will focus on the left-most part of the plot (Figure 6D), in which a clear negative relationship between CV and promoter methylation is observed. This result suggests that the regulated gene expression variability component can potentially be due to the methylation of gene promoters, particularly when methylation levels are low and expression is high.

Conclusions

We found that U-CLL, which is a more aggressive type of CLL, is characterized by higher variability in gene expression across patients compared to M-CLL, which might impact the aggressiveness, adaptability and hence resilience to drugs of these tumours.

The genes that display differences in gene expression variability between the two subtypes are generally moderately to highly expressed and enriched for functions related to proliferation, growth, development and apoptosis, reinforcing the possible link between increased expression heterogeneity

and clinical subtypes. Other functions that seem to have increased variability in the U-CLL patients are related to membrane components and inter-cellular communication.

Recent research has shown the presence of more genetic heterogeneity in U-CLL samples [18], which could suggest that the increased variability that we are observing is a reflection of the genetic composition at the gene expression level. Moreover, we have used variability across patients in the hope that this measure would work as a proxy for tumour heterogeneity in single patients. Currently, this hypothesis allows us to relate the across-patient variability to the worse prognosis observed for U-CLL patients, which can be attributed to the presence of heterogeneity and hence adaptability and resilience to drugs in the patients.

In the final part of the paper we investigate the potential control of the differences in gene expression between individuals by their DNA methylation levels. Epigentic control in general, and DNA methylation in particular, are obvious candidates to be the regulators of expression variation. Indeed, Kulis et al. (Kulis et al., 2012) showed that methylation profiles were enough to approximately cluster U-CLL and M-CLL patients (a good classification can also be obtained with our Random Forest classifier, (data not shown), demonstrating the relation between methylation and disease aggresivity. Indeed, it has been recently sugested that the heterogeneity at the epigenetic level might be related with the U-CLL phenotype [44]. Overall, these observations suggest an interesting connection between the epigenotype, in this case DNA methylation, and the heterogeneity of phenotypes, where epigenetics factors may exert the control of inter-individual expression variability by switching specific genes on or off.

Furthermore, our observations across patients could relate to single-cell heterogeneity in each patient. To verify this hypothesis, larger datasets and single-cell genomics data would be an invaluable new

source of complementary information and they could contribute to classifying patients in a clinical setting.

Methods

All analyses were performed using R version 3.0.2 (x86_64-pc-linux-gnu) [45]

Gene expression and methylation datasets

We used the ICGC CLL microarray datasets previously published by Kulis et al. [22] and Ferreira et al. [23]. Gene expression measurements were obtained by Affymetrix Human Genome U219 Array Plates. 48,786 features of the microarray had passed quality controls and filtering as described previously [23]. The dataset comprises 122 CLL samples (70 M-CLL and 52 U-CLL) and 20 control samples of different healthy B-cells (5 naive B cells, 3 IgM^+ and IgD^+ memory B cells, 4 IgA^+ and IgG^+ memory B cells and 8 $CD19^+$ B cells). DNA methylation was measured by Infinium HumanMethylation450K BeadChips. 48,786 probes had passed quality control and filtering procedures as described previously [22], as well as an an additional filter removing probes containing SNPs.

Furthermore we included an additional gene expression dataset of CLL published by Fabris et al. [21] GEO accession number GSE9992, containing 60 samples (24 M-CLL and 36 U-CLL) in our analyses. The data was quality assessed and preprocessed as described in [23].

In order to be able to investigate the relationship between gene expression and DNA methylation, we mapped the microarray probe identifiers to Ensembl identifiers and used the average of the measurements for each gene. DNA methylation features were mapped to genomic regions (especially promoters and gene bodies) as described previously [22].

Generating bins in figures

We established 10 equally large classes of gene expression levels from lowly to highly expressed genes

(the overall expression values range from 3.6 to 14.4, see Additional File 1, Supplementary Methods, Supplementary Table 2) and ten equally large classes of DNA methylation levels given in beta values from unmethylated (0) to fully methylated (1) (see Additional File 1, Supplementary Methods, Supplementary Table 3).

Hierarchical clustering

Hierarchical agglomerative clustering was performed on the Spearman correlation matrix of the gene expression and the methylation dataset from Kulis et al [22] using the function Agnes from package cluster (version 1.14.4) in R using the 'average' method and default settings for the distance metric. Heatmaps were then generated using heatmap.2 from the package gplots (version 2.12.1).

Aggregating patients to estimate gene expression averages and fluctuations in single cells

We describe the procedure for creating variables for aggregate 'superpatients'. We extracted 5 random U-CLLpatients and another 5 M-CLL patients and repeated the operation until sets of 5 could be made without repeating the patients in the groups. For each group of patient we calculate the mean expression, mean methylation and relative range of expression (difference between maximum and minimum of expression divided by the mean expression). This is chosen over the CV as a better statistical estimate for such low numbers of patients. We thus produce a new cohort of 'superpatients' of which half are U-CLL and the other half M-CLL.

F-test for differential variance

We performed gene-wise F-tests comparing M-CLL with U-CLL using R. Multiple hypotheses testing correction was performed using the Benjamini-Hochberg algorithm [46]. To test if the results of the F-test were reproducible between the two datasets analysed [21, 22], we performed a hypergeometric test

on the overlap of the top genes showing differential variability based on the significance levels given by the F-test.

Functional Analysis

Functional analyses were performed on the top 500 genes showing increased variability in U-CLL based on differences in CV. To test for enrichment of biological functions and pathways we used DAVID [47]. We uploaded the list of top 500 genes of the ICGC CLL dataset [22], the top 500 genes of the Fabris CLL dataset [21] and the 126 genes which are in common in the list of the top 500 genes in both studies and used as the background set the corresponding set of genes analysed in the dataset respectively common in the two datasets. We tested for the following functional annotation: GOTERM_BP_ALL, GOTERM_CC_ALL, GOTERM_MF_ALL, KEGG_PATHWAY and REACTOME_PATHWAY and set the threshold of Counts to a minimum of 3 genes. We consider terms and pathways as significantly enriched when the corresponding p-value adjusted by the Benjamini-Hochberg algorithm [46] for multiple hypotheses correction is smaller than 0.05.
To further confirm our results we performed the same analysis using R. We employed the packages Category and Gostats [48] from R/Bioconductor to search for enriched Gene Ontology biological processes and overrepresented KEGG pathways. We used conditional testing and fetched all results without setting a p-value cutoff, removed all gene sets containing fewer than three genes and performed multiple hypothesis testing correction on the entire data set using the Benjamini-Hochberg algorithm [46].

Network Construction

We used the B cell specific functional interaction network of Lefebvre et al., 2010 [31] containing 5,748 nodes (genes) and 64,600 unique edges (interactions) based on Entrez gene identifiers. We

selected the 126 genes which are in common within the top 500 genes with differential variability between the two CLL datasets analyzed [21, 22] and mapped them to Entrez gene identifiers resulting in 127 unique Entrez gene identifiers. We then selected these 127 genes and their direct neighbors in the network which led to a network of 329 genes connected by 1,108 edges. This network of 329 genes is the one we investigated further.

We performed a functional enrichment analysis of this network compared to all the genes contained in the whole B cell network which are also present in both of our gene expression datasets (used as background) in DAVID as described before. This analysis results in 475 significantly enriched GO terms and pathways.

We then identified 9 network communities in the network of 329 genes by using Gephi [49] and Louvain's method [32].

Single patient classification features

Random Forest models [41] were trained on the Kulis et al. dataset [22] (randomForest package in R, 1000 trees). Expression values for each probe were used as features for the first classifier. We further created models using only expression of the top 500 most variable probes and genes. Finally we defined a new feature as the distance from one gene expression value to the median of that gene over the population (abs(x-median(x))). All models were first evaluated with the error rate estimated from the Objects Out of the Bag (OOB) and confusion matrices. The ROCR package was used to generate Receiving Operator Characteristic curves, which were used to evaluate the prediction of the Fabris et al. patients, our independent test set [21].

List of abbreviations

CLL Chronic Lymphocytic Leukemia

FDR False Discovery Rate

AUC Area Under the Curve

ROC Receiver Operating Characteristic

Competing interests

The author(s) declare that they have no competing interests

Authors' contribution

S. E. performed the analysis and wrote the paper. V. P. conceived the study, performed the analysis and wrote the paper. D. R. wrote the paper and guided the analysis. A. V. wrote the paper and supervised the work.

Description of additional data files

The following additional data are available with the online version of this paper.

Additional_file_1.pdf: **Main supplementary file.** Supplementary figures, tables and methods.

Additional_file_X1.xls: **Differentially variable genes**. Excel table of top differentially variable genes with overlaps of the two datasets.

Additional_file_X2.xls: **Functional enrichments of variable genes**. Excel table of enrichment analysis of top differentially variable genes in the two datasets and their overlap.

Additional_file_X3.xls: **Functional enrichments of network modules.** Excel table of functional enrichments of the different modules in the highly differentially variable gene network.

Acknowledgements

This work was funded by the Spanish Ministry of Economy and Competitivity (MINECO, BIO2012-40205) and by BLUEPRINT Consortium (FP7/2007-2013, under grant agreement number 282510) and the CLL Genome project (http://www.cllgenome.es) of the International Cancer Genome Consortium (ICGC), which is funded by MINECO through the Instituto de Salud Carlos III (ISCiii). The authors thank everyone in the CLL Genome project of the International Cancer Genome Consortium, in particular Elias Campos, Jose Ignacio Martin-Subero and their teams for stimulating discussions on the topic of this project. They also gratefully thank David Juan, Victor de la Torre and all members of the Structural Biology and BioComputing programme for interesting discussions. SE is supported by a fellowship from La Caixa and VP by a FEBS long-term fellowship.

References

1. Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell.** *Science* 2002, **297**:1183–6.

2. Carey LB, van Dijk D, Sloot PMA, Kaandorp JA, Segal E: **Promoter sequence determines the relationship between expression level and noise.** *PLoS Biol* 2013, **11**:e1001528.

3. Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, Dekel E, Yarnitzky T, Liron Y, Polak P, Lahav G, Alon U: **Oscillations and variability in the p53 system.** *Mol Syst Biol* 2006, **2**:2006.0033.

4. Ashall L, Horton CA, Nelson DE, Paszek P, Harper C V, Sillitoe K, Ryan S, Spiller DG, Unitt JF, Broomhead DS, Kell DB, Rand DA, Sée V, White MRH: **Pulsatile stimulation determines timing and specificity of NF-kappaB-dependent transcription.** *Science* 2009, **324**:242–6.

5. Paszek P, Ryan S, Ashall L, Sillitoe K, Harper C V, Spiller DG, Rand DA, White MRH: **Population robustness arising from cellular heterogeneity.** *Proc Natl Acad Sci U S A* 2010, **107**:11644–9.

6. Lehner B: **Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast.** *PLoS One* 2010, **5**:e9035.

7. Tirosh I, Reikhav S, Levy AA, Barkai N: **A yeast hybrid provides insight into the evolution of gene expression regulation.** *Science* 2009, **324**:659–62.

8. Basehoar AD, Zanton SJ, Pugh BF: **Identification and distinct regulation of yeast TATA box-containing genes.** *Cell* 2004, **116**:699–709.

9. Choi JK, Kim Y-J: **Intrinsic variability of gene expression encoded in nucleosome positioning sequences.** *Nat Genet* 2009, **41**:498–503.

10. Salari R, Wojtowicz D, Zheng J, Levens D, Pilpel Y, Przytycka TM: **Teasing Apart Translational and Transcriptional Components of Stochastic Variations in Eukaryotic Gene Expression**. *PLoS Comput Biol* 2012, **8**:e1002644.

11. Dong D, Shao X, Deng N, Zhang Z: **Gene expression variations are predictive for stochastic noise.** *Nucleic Acids Res* 2011, **39**:403–13.

12. Lehner B, Kaneko K: **Fluctuation and response in biology.** *Cell Mol Life Sci* 2011, **68**:1005–10.

13. Marusyk A, Polyak K: **Tumor heterogeneity: causes and consequences.** *Biochim Biophys Acta* 2010, **1805**:105–17.

14. Meacham CE, Morrison SJ: **Tumour heterogeneity and cancer cell plasticity.** *Nature* 2013, **501**:328–37.

15. Brock A, Chang H, Huang S: **Non-genetic heterogeneity--a mutation-independent driving force for the somatic evolution of tumours.** *Nat Rev Genet* 2009, **10**:336–42.

16. Almendro V, Cheng Y-K, Randles A, Itzkovitz S, Marusyk A, Ametller E, Gonzalez-Farre X, Muñoz M, Russnes HG, Helland A, Rye IH, Borresen-Dale A-L, Maruyama R, van Oudenaarden A, Dowsett M, Jones RL, Reis-Filho J, Gascon P, Gönen M, Michor F, Polyak K: **Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity.** *Cell Rep* 2014, **6**:514–27.

17. Marusyk A, Almendro V, Polyak K: **Intra-tumour heterogeneity: a looking glass for cancer?** *Nat Rev Cancer* 2012, **12**:323–34.

18. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ: **Evolution and impact of subclonal mutations in chronic lymphocytic leukemia.** *Cell* 2013, **152**:714–26.

19. Ho JWK, Stefani M, dos Remedios CG, Charleston MA: **Differential variability analysis of gene expression and its application to human diseases.** *Bioinformatics* 2008, **24**:i390–8.

20. Hulse AM, Cai JJ: **Genetic variants contribute to gene expression variability in humans.** *Genetics* 2013, **193**:95–108.

21. Fabris S, Mosca L, Todoerti K, Cutrona G, Lionetti M, Intini D, Matis S, Colombo M, Agnelli L, Gentile M, Spriano M, Callea V, Festini G, Molica S, Lambertenghi Deliliers G, Morabito F, Ferrarini M, Neri A: **Molecular and transcriptional characterization of 17p loss in B-cell chronic lymphocytic leukemia.** *Genes Chromosomes Cancer* 2008, **47**:781–93.

22. Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, Martínez-Trillos A, Castellano G, Brun-Heath I, Pinyol M, Barberán-Soler S, Papasaikas P, Jares P, Beà S, Rico D, Ecker S, Rubio M,

Royo R, Ho V, Klotzle B, Hernández L, Conde L, López-Guerra M, Colomer D, Villamor N, Aymerich M, Rozman M, Bayes M, Gut M, Gelpí JL, et al.: **Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia.** *Nat Genet* 2012, **44**:1236–1242.

23. Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Djebali S, Papasaikas P, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, Clot G, Aymerich M, Rozman M, Kulis M, Tamborero D, Gouin A, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, et al.: **Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia.** *Genome Res* 2013.

24. Kampen NG Van: *Stochastic Processes in Physics and Chemistry*. North Holland; 2007.

25. Kaern M, Elston TC, Blake WJ, Collins JJ: **Stochasticity in gene expression: from theories to phenotypes.** *Nat Rev Genet* 2005, **6**:451–64.

26. Raj A, van Oudenaarden A: **Nature, nurture, or chance: stochastic gene expression and its consequences.** *Cell* 2008, **135**:216–26.

27. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise.** *Nature* 2006, **441**:840–6.

28. Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR, Collins JJ: **Phenotypic consequences of promoter-mediated transcriptional noise.** *Mol Cell* 2006, **24**:853–65.

29. Alemu EY, Carl JW, Corrada Bravo H, Hannenhalli S: **Determinants of expression variability.** *Nucleic Acids Res* 2014, **42**:3503–14.

30. Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, Neri A, Califano A, Dalla-Favera R: **Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells.** *J Exp Med* 2001, **194**:1625–38.

31. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A: **A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers.** *Mol Syst Biol* 2010, **6**:377.

32. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks**. *J Stat Mech Theory Exp* 2008, **2008**:P10008.

33. Ruvolo PP, Deng X, Carr BK, May WS: **A functional role for mitochondrial protein kinase Calpha in Bcl2 phosphorylation and suppression of apoptosis.** *J Biol Chem* 1998, **273**:25436–42.

34. Shatz M, Liscovitch M: **Caveolin-1 and cancer multidrug resistance: coordinate regulation of pro-survival proteins?** *Leuk Res* 2004, **28**:907–8.

35. Engelman JA, Zhang X, Galbiati F, Volonte D, Sotgia F, Pestell RG, Minetti C, Scherer PE, Okamoto T, Lisanti MP: **Molecular genetics of the caveolin gene family: implications for human cancers, diabetes, Alzheimer disease, and muscular dystrophy.** *Am J Hum Genet* 1998, **63**:1578–87.

36. Gilling CE, Mittal AK, Chaturvedi NK, Iqbal J, Aoun P, Bierman PJ, Bociek RG, Weisenburger DD, Joshi SS: **Lymph node-induced immune tolerance in chronic lymphocytic leukaemia: a role for caveolin-1.** *Br J Haematol* 2012, **158**:216–31.

37. Eychène A, Rocques N, Pouponnot C: **A new MAFia in cancer.** *Nat Rev Cancer* 2008, **8**:683–93.

38. Del Giudice I, Osuji N, Dexter T, Brito-Babapulle V, Parry-Jones N, Chiaretti S, Messina M, Morgan G, Catovsky D, Matutes E: **B-cell prolymphocytic leukemia and chronic lymphocytic leukemia have distinctive gene expression signatures.** *Leukemia* 2009, **23**:2160–7.

39. Shachar I, Cohen S, Marom A, Becker-Herman S: **Regulation of CLL survival by hypoxia-inducible factor and its target genes.** *FEBS Lett* 2012, **586**:2906–10.

40. Lee YK, Strege AK, Bone ND, Wellik LE, Chan DA, Giaccia AJ, Mukhopadhyay D, Kay NE: **Hypoxia Inducible Factor-1{alpha} Is over Expressed in CLL B Cells Because of an Impaired Proteasome Pathway Associated with Defective Interaction with von Hippel-Landau Protein.** *ASH Annu Meet Abstr* 2005, **106**:2115.

41. Breiman L: **Random Forests**. *Mach Learn* 2001, **45**:5–32.

42. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET: **Passive and active DNA methylation and the interplay with genetic variation in gene regulation.** *Elife* 2013, **2**:e00523.

43. Oakes CC, Claus R, Gu L, Assenov Y, Hüllein J, Zucknick M, Bieg M, Brocks D, Bogatyrova O, Schmidt CR, Rassenti L, Kipps TJ, Mertens D, Lichter P, Döhner H, Stilgenbauer S, Byrd JC, Zenz T, Plass C: **Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia.** *Cancer Discov* 2014, **4**:348–61.

44. Oakes CC, Claus R, Gu L, Assenov Y, Hüllein J, Zucknick M, Bieg M, Brocks D, Bogatyrova O, Schmidt CR, Rassenti L, Kipps TJ, Mertens D, Lichter P, Döhner H, Stilgenbauer S, Byrd JC, Zenz T, Plass C: **Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia.** *Cancer Discov* 2014, **4**:348–61.

45. R Core Team: **R: A Language and Environment for Statistical Computing**. 2013.

46. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289 – 300.

47. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.

48. Falcon S, Gentleman R: **Using {GOstats} to test gene lists for {GO} term association**. *Bioinformatics* 2007, **23**:257.

49. Bastian M, Heymann S, Jacomy M: **Gephi: An Open Source Software for Exploring and Manipulating Networks**. 2009.

Figure Legends

Figure 1: Definition of Coefficient of Variation (CV) and its dependence on gene expression levels. A) CV versus expression of genes in bins of increasing expression level (see Methods). B) Relationship between reciprocal of mean expression and expression in bins of increasing expression level. C) Dependence of the standard deviation of expression across patients on the level of expression.

Figure 2: Comparison of CV of housekeeping genes to non housekeeping genes for equal levels of gene expression. Non-housekeeping genes (green) show a more marked increase in CV for intermediate levels of expression compared to housekeeping genes (blue).

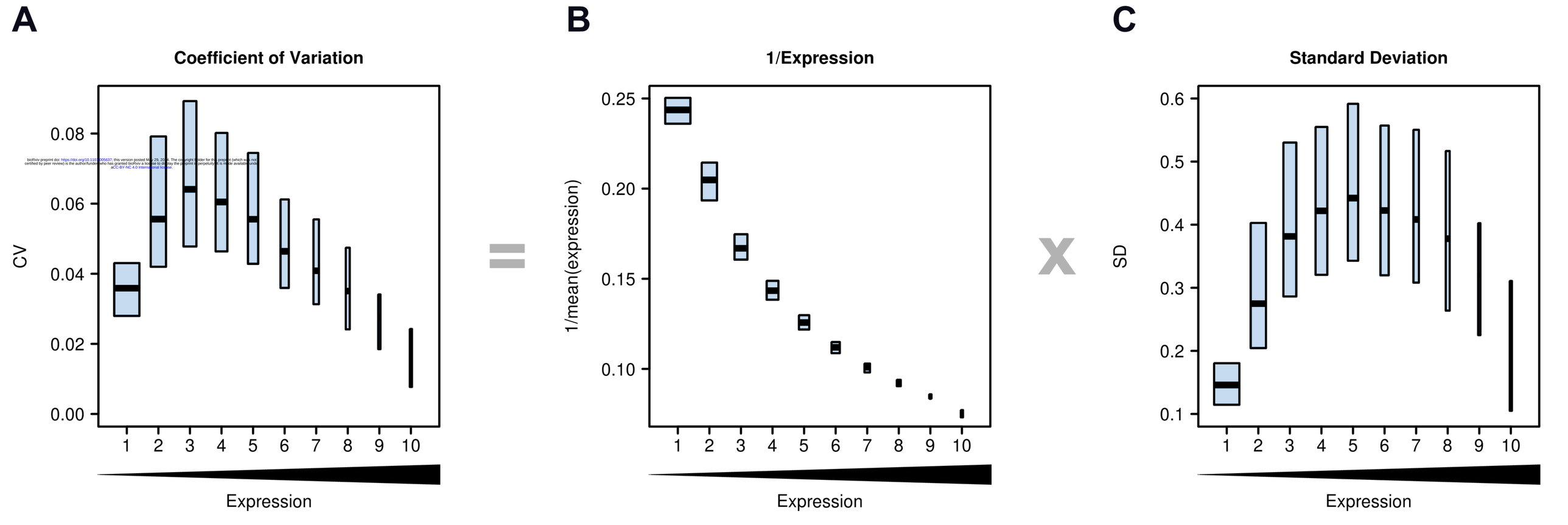
Figure 3: Analysis of expression data comparing U-CLL and M-CLL patients. A) Scatter plot of CV of expression across patients in the two disease subtypes. B) Scatter plot of mean of expression across patients in the two disease subtypes. C) Log-ratio of expression CV between M-CLL and U-CLL and its dependence on values of gene expression.

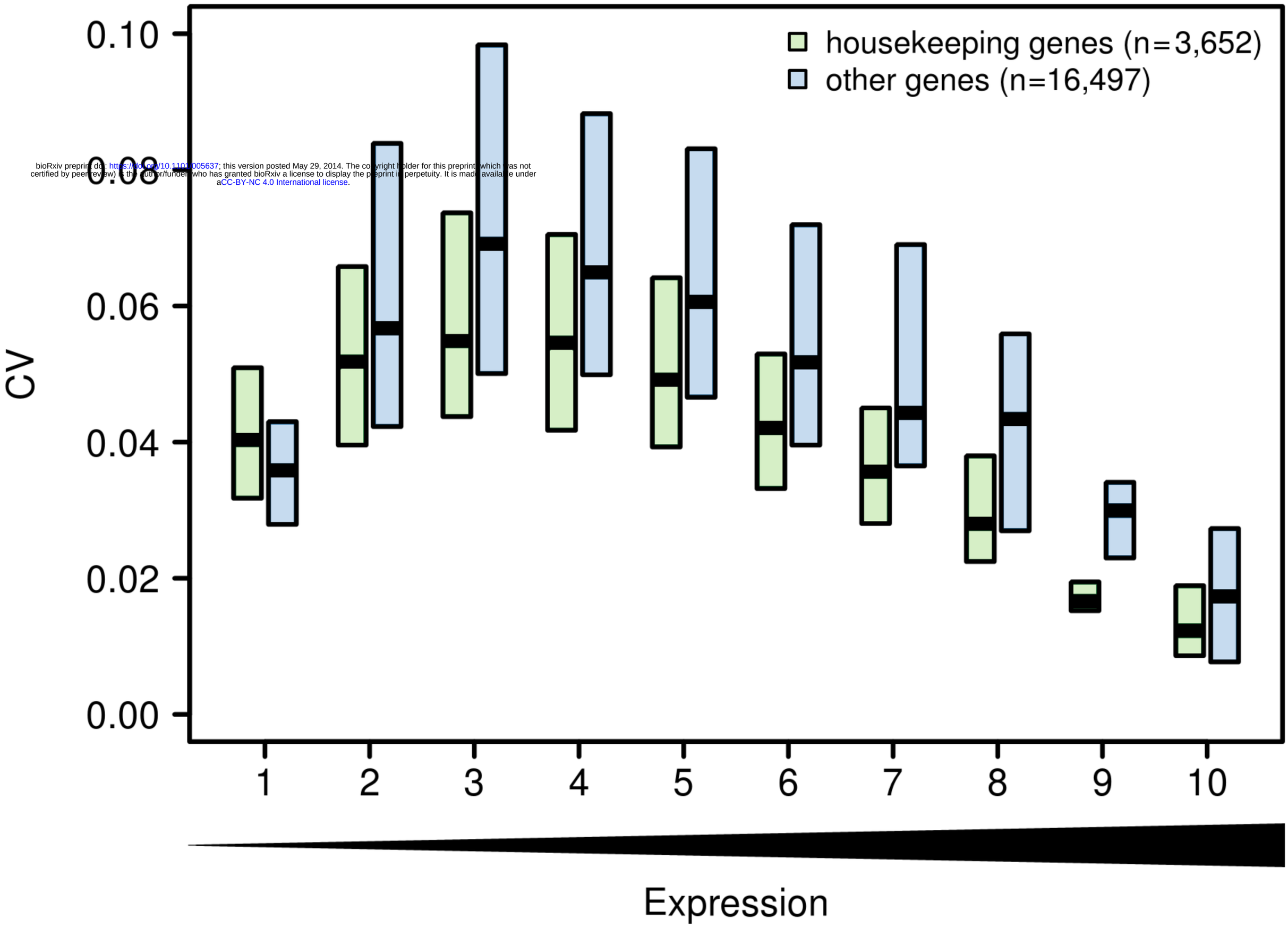
Figure 4: Network representation of genes with increased variability in U-CLL patients in the context of a B-cell specific network [31]**.** Node sizes are relative to the degrees of the nodes, i.e. big nodes represent highly connected genes. Different modules are highlighted in different colours.

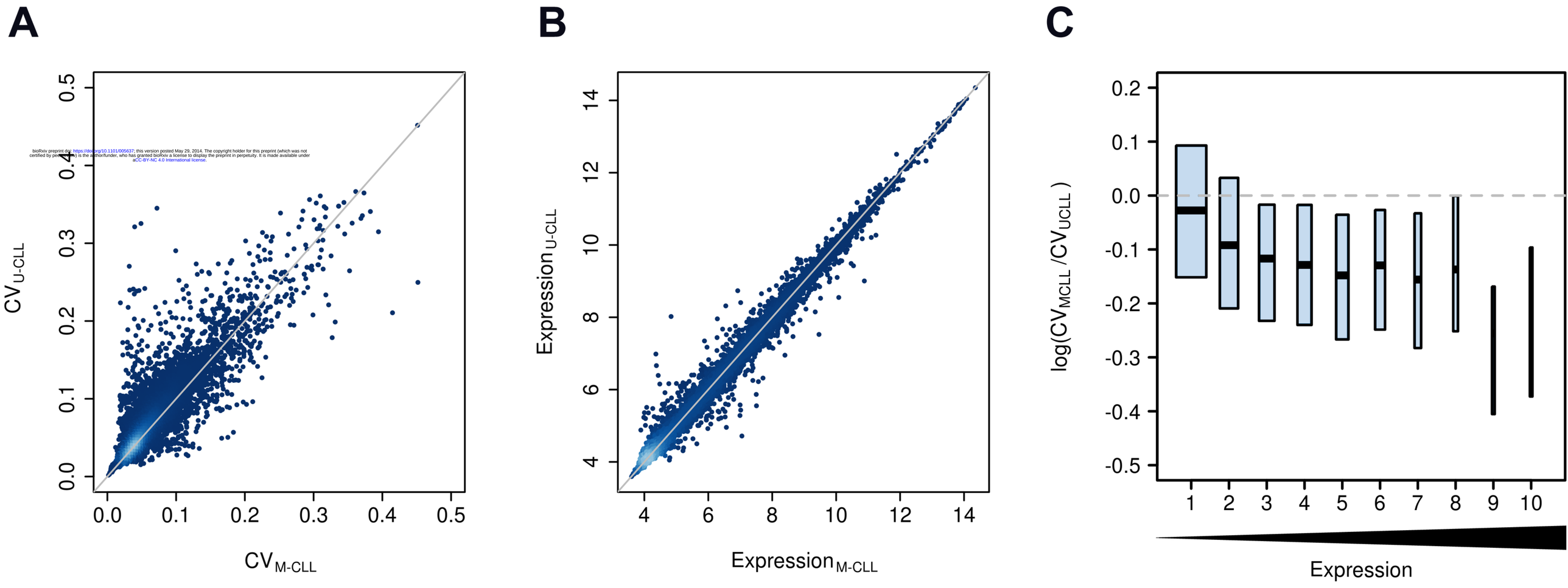
Figure 5: Hierarchical clustering of expression and methylation data. A) Heatmap representing clustering of patients into U-CLL (orange), M-CLL (magenta) and healthy cells (see methods) based on gene expression data. B) Heatmap representing clustering of 'superpatients' into U-CLL and M-CLL

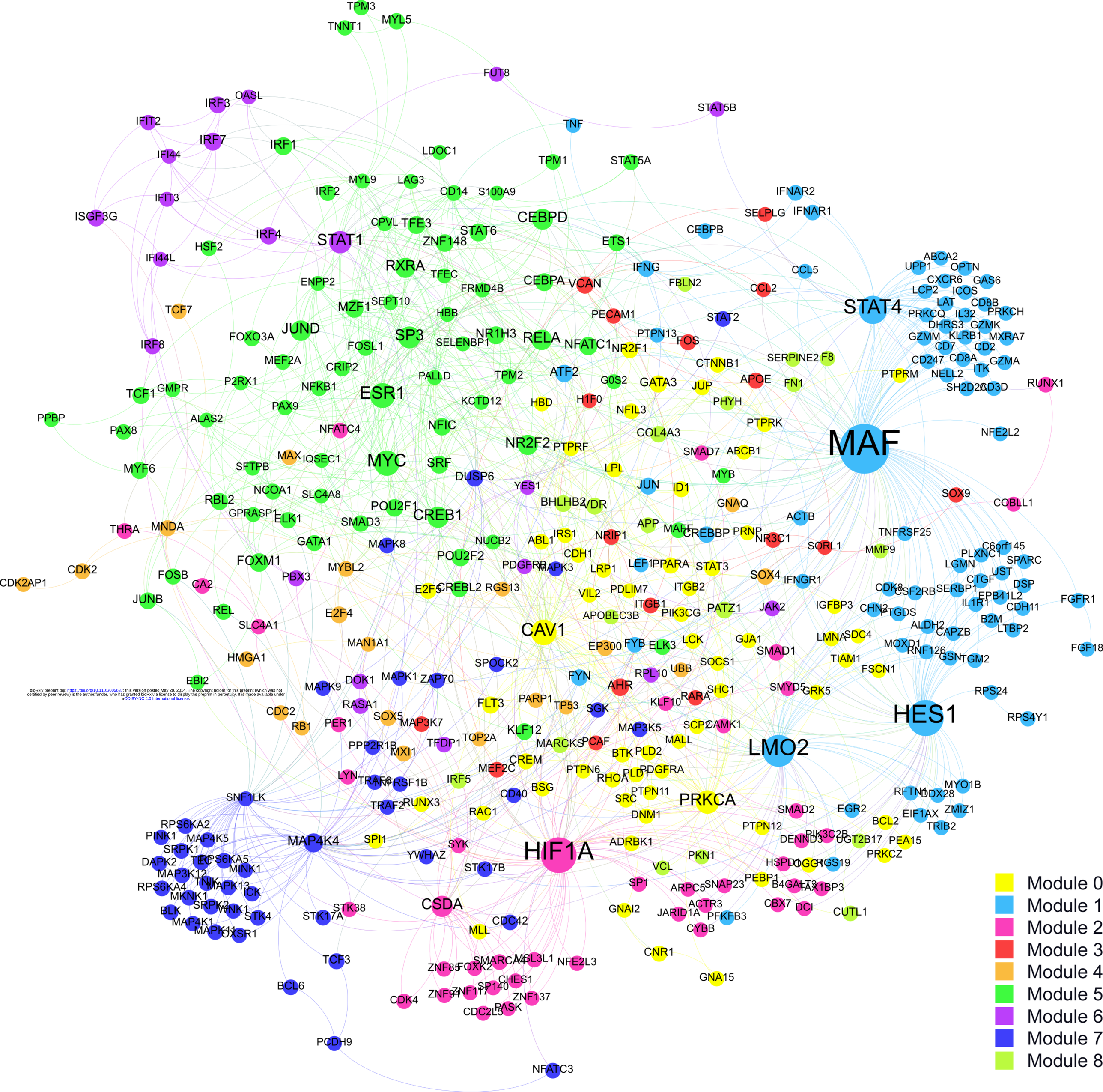
based on aggregated expression mean of the superpatients. D) Heatmap representing clustering of 'superpatients' into U-CLL and M-CLL based on variability of gene expression (relative range) in the superpatients.

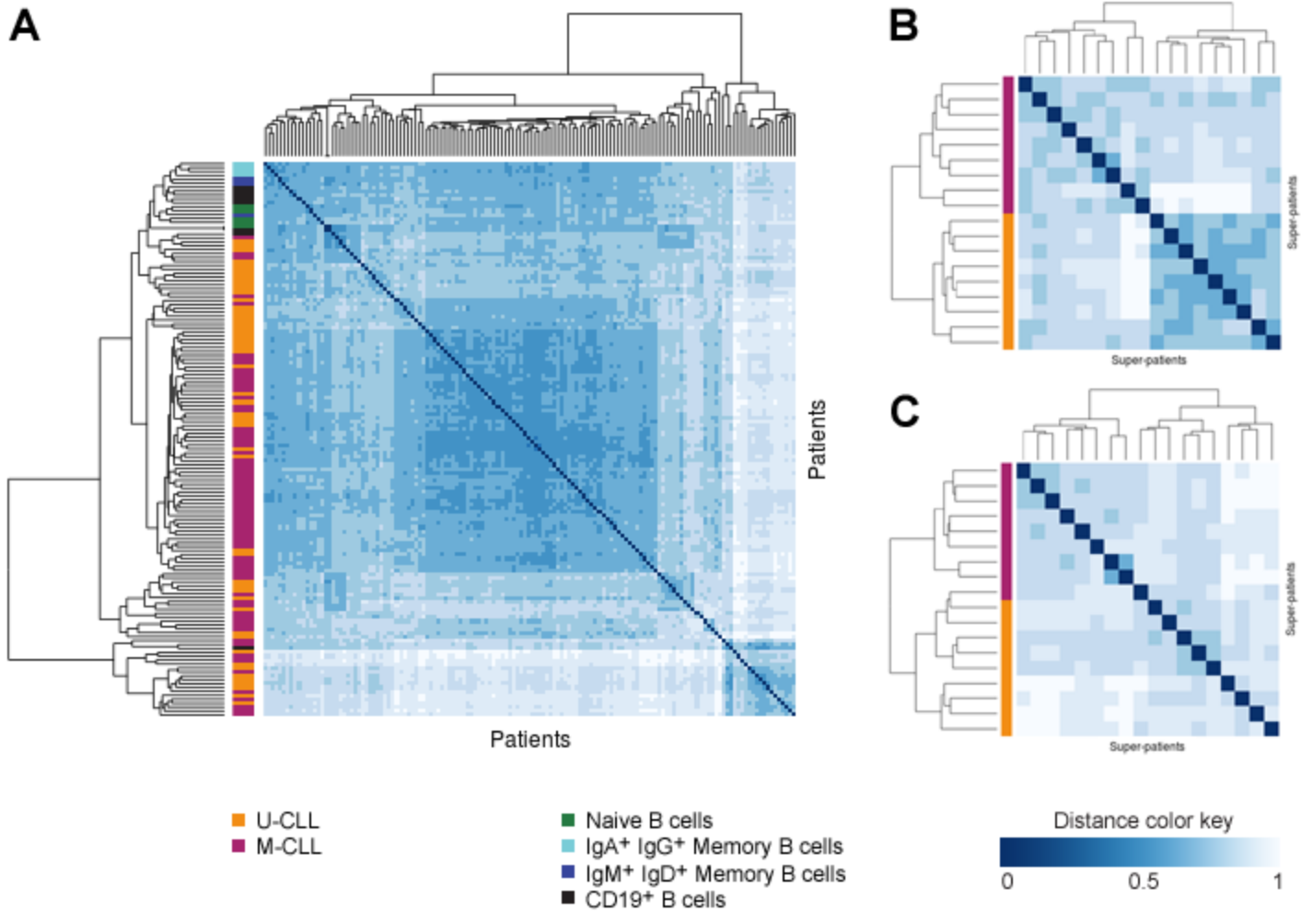
Figure 6: Analysis of methylation data comparing U-CLL and M-CLL patients. A) Heatmap representing clustering of Kulis et al. patients into U-CLL (orange), M-CLL (magenta) and healthy cells based on promoter methylation data. B) Scatter plot of mean promoter methylation across patients in the two disease subtypes. C) Relationship between gene expression and methylation for bins of increasing methylation D) Relationship between expression CV and different levels of promoter methylation.



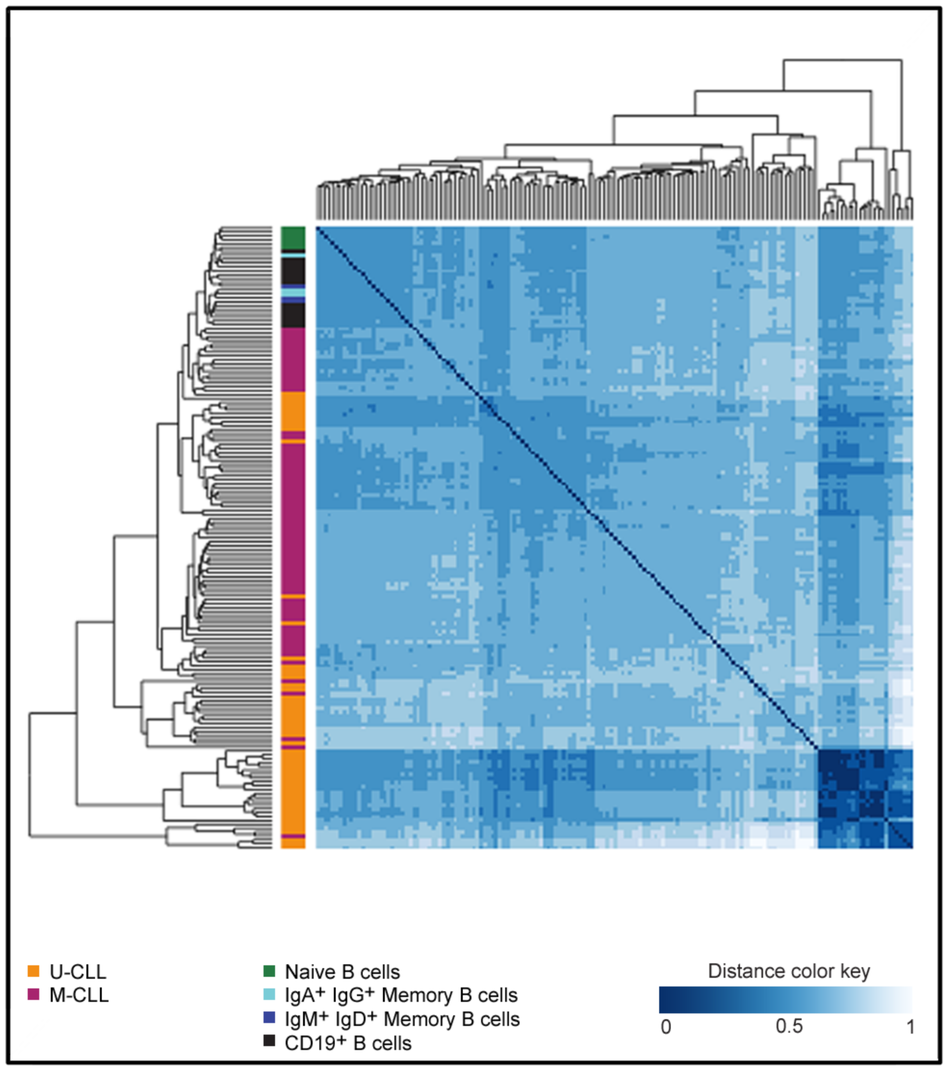




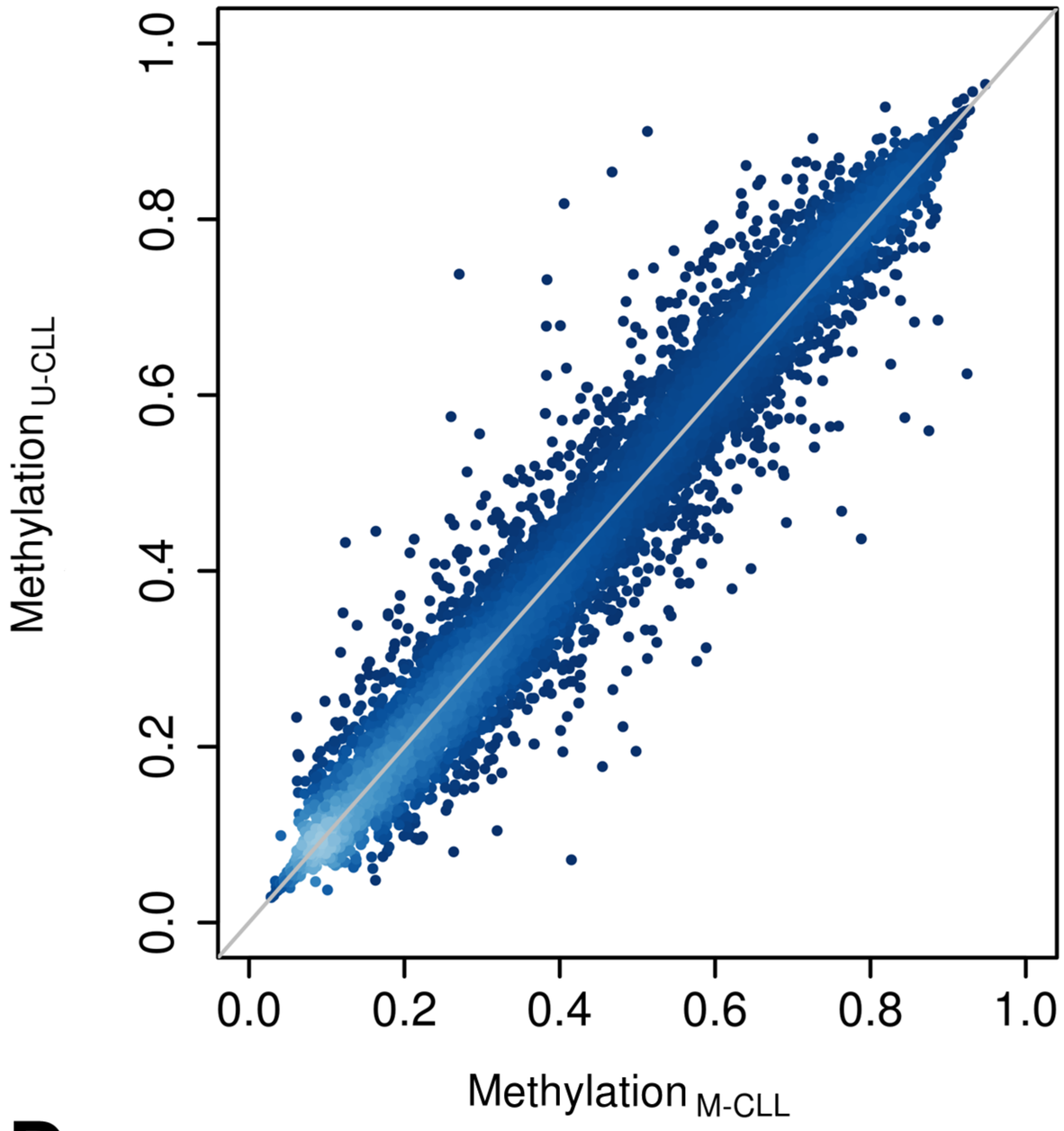



A
B
C
Patients
Patients
Super-patients
Super-patients
Super-patients
Super-patients
U-CLL
M-CLL
Naive B cells
IgA+ IgG+ Memory B cells
IgM+ IgD+ Memory B cells
CD19+ B cells
Distance color key
0
0.5
1

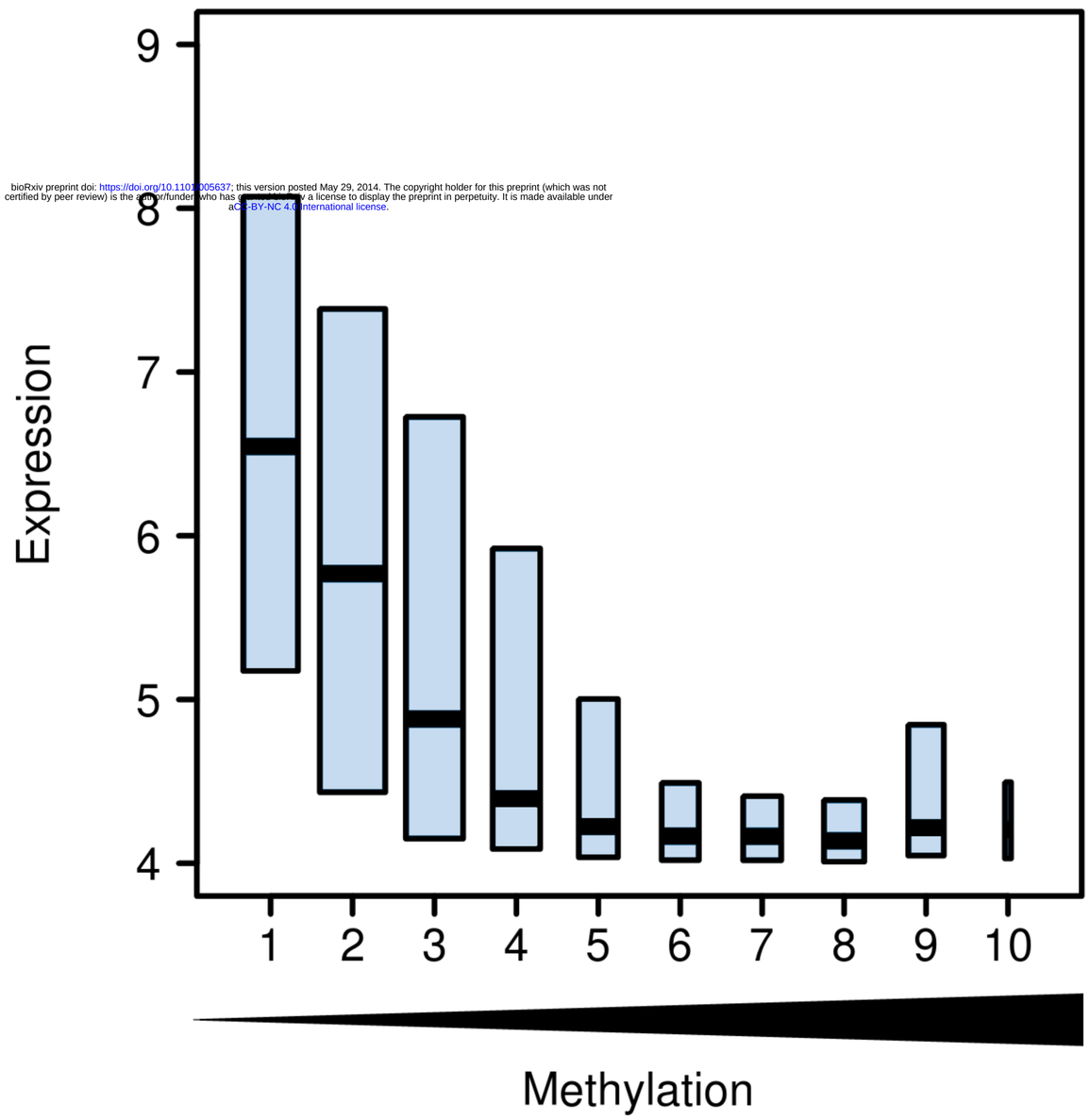
A



B



C



D