IDENTIFICATION OF CRC SUB-TYPES BY MULTI-OMIC DATA INTEGRATION

Authors
Giulia Brunelli[1,2]
Romina D’Aurizio[2,1]

Affiliations
1. University of Siena, IT;
2. ITT, CNR, Pisa, IT

Introduction

Colorectal cancer (CRC) is characterized by heterogeneous outcomes, variable drug responses, and often associated with a poor prognosis as consequence of the aberrant behavior of tumor cells [1]. Even if several attempts to classify CRC inter-tumor heterogeneity were made using single omics: genomic instability pathways (CIN and MS), CpG Island Methylator Phenotype (CIMP), Consensus Molecular Subtypes (CMS) and stromal transcriptional components (CRIS3) [2,3], none of them combined together the informations carried by genetics and epigenetics alterations and copy number variations

Methods

We applied Similarity network fusion [4,5] (SNF)-based integrative clustering combining gene expression, miRNA expression, copy number variation, DNA methylation and single nucleotide polymorfism of 326 CRC patients from the TCGA (READ and COAD) datasets downloaded from Xena Browser database. For each omic data we first remove features with more than 20% of missing data and sd = 0. Then , we identified the distance measure that was able to detect the most consistent partitioning of the samples according to the silhouette index and clusters size: we excluded distances for which the most part of the samples were assigned to just one cluster. We evaluated Spearman, Euclidean and Manhattan distances for continous data (CNV, RNA-seq, miRNA and DNA methylation) and Manhattan and Chi-squared distance for SNP. More precisely, we built a similarity network on each omic data on the base of the different distance measures, we applied the Spectral Clustering Algorithm on the respective network and evaluate the corresponding clustering as mentioned above. Once the distance has been chosen, we defined the similarity network on the base of that distance and applied the SNF algorithm. Finally, the Spectral Clustering algorithm was applied to the fused Network, selecting the number of clusters which maximized the Eigen Gap and minimized the Rotation Cost. To obtain a finer partition of patients, we applied the SNF algorithm to each group independently. In order to evaluate the biological meaning of the obtained clusterization, we compared the distribution of patients clinical features among the groups resulting by the SNF clusterization method using Chi-Squared or Fisher test and ANOVA or Kruskall-Wallis test according to the nature of our data. Likewise, we compared the CIMS and CRIS clusterization with the SNF partition. Subsequently, we performed a differential gene expression analysis between groups against the overall population, using quasi-likelihood (QL) F-test [6]. Finally, we evaluate which CNV were more likely to characterize each group fitting a multinomial model (for the discrete gstic value) to compare the probability to have a deletion or a duplication with the wild type phenotype in each group. To conclude, a surviva analysis has be done to asses the relationship between the SNF clusterization and OS and PFI.

Analysis & Results

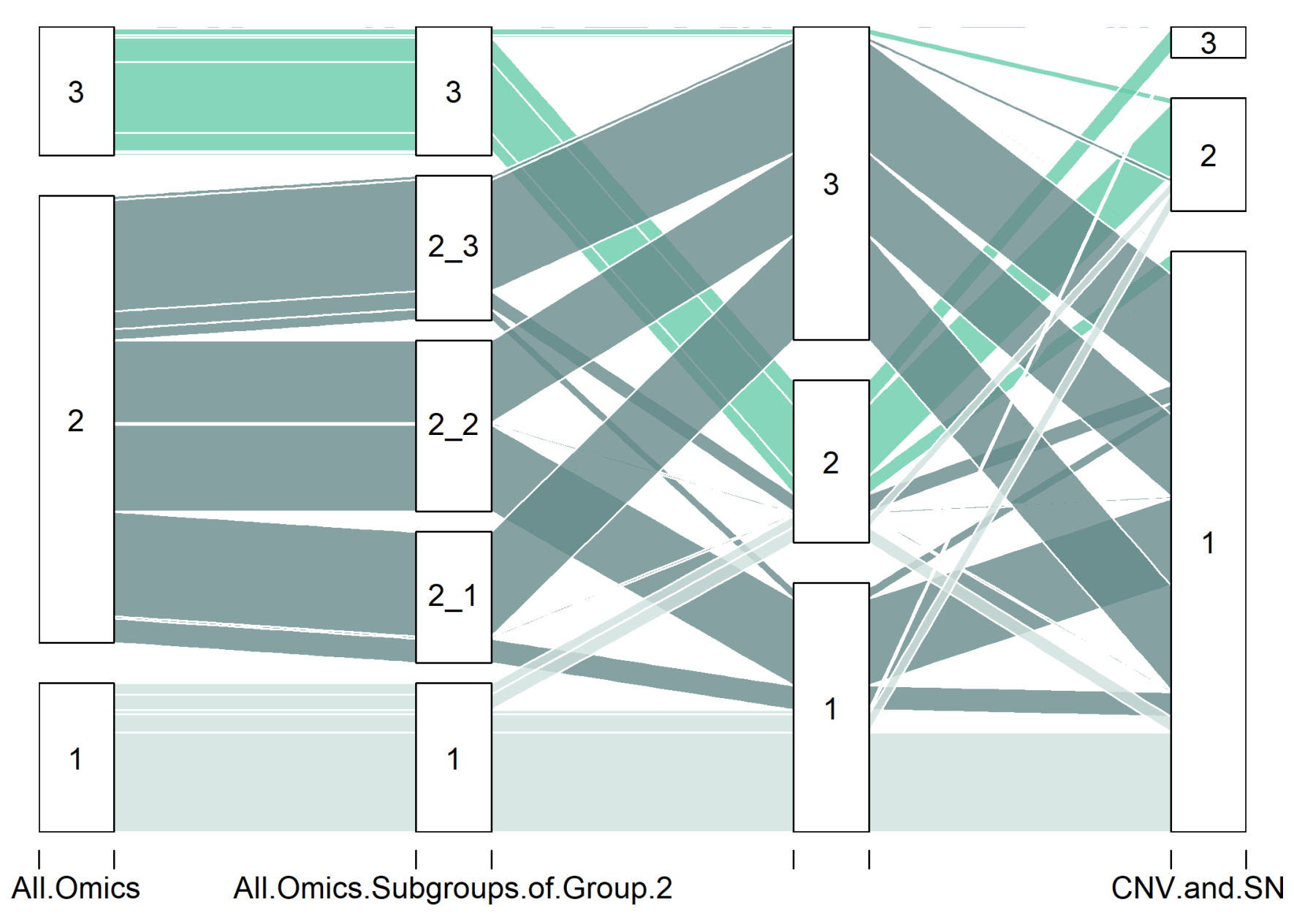


Fig.1. Alluvial plot to compare clusterizations on the base of different data types: we applied the SNF algorithm also on these pair of omic dataset separately. The SNF algorithm applied to all the omics data produced 3 groups having size equal to 67, 201, 58 respectively (Silhouette index = 0.362). Even if the silhouette index was higher for both the clusterization obtained from RNA-seq+miRNA and CNV+SNP (0.528 and 0.63 respectively), groups obtained from the fused network of CNV and SNP don’t exhibited any clinical differences. The groups derived from the spectral clustering applied on RNA-seq and miRNA where similar to those obtained from the all omics data integration.

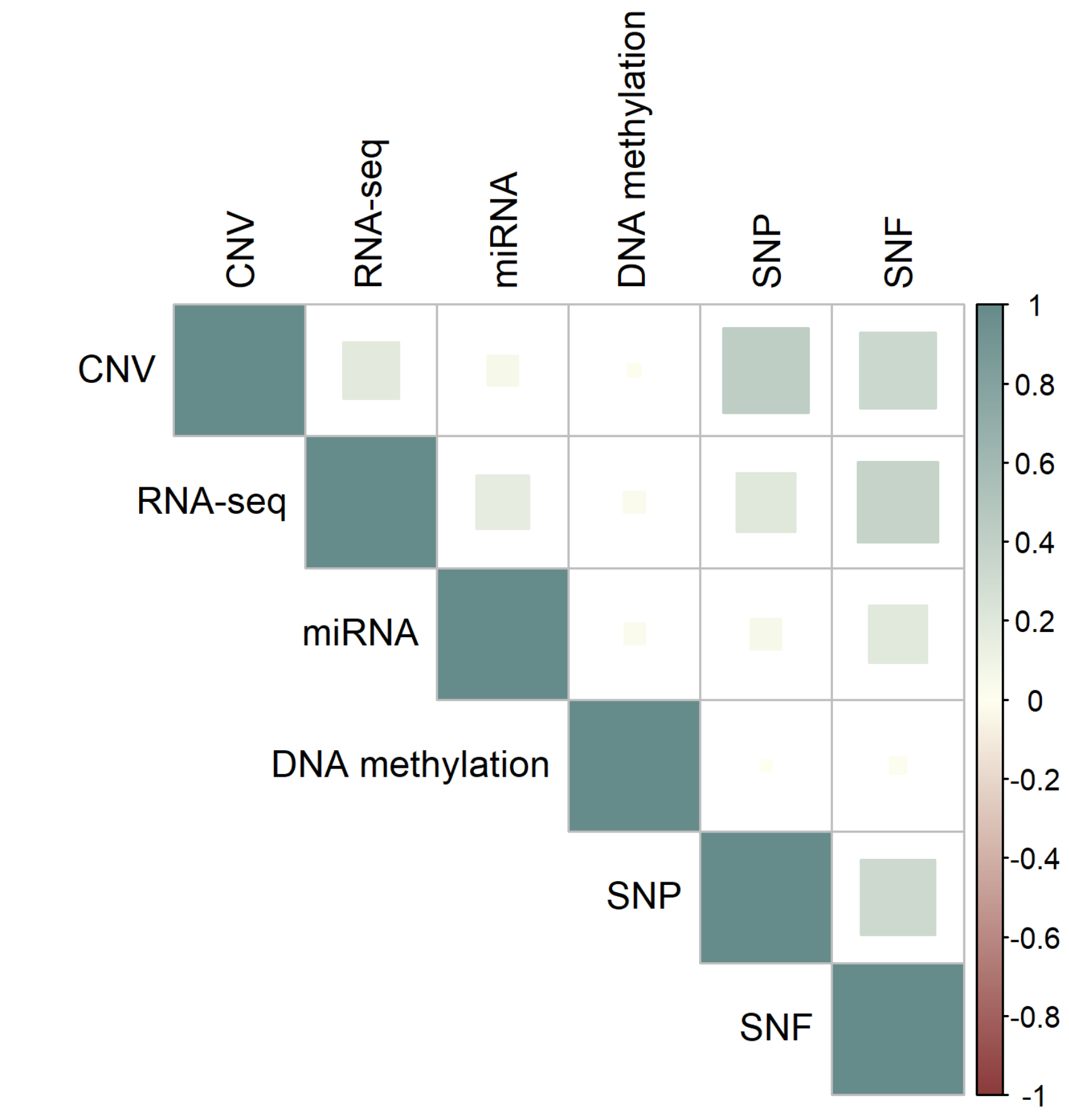


Fig.2. Pairwise NMI index between cluster assignments made with spectral clustering on all the similarity matrices obtained on each omic dataset and the fused network. The highest concordance with the fused network is obtained for RNA-seq, CNV and SNP data.

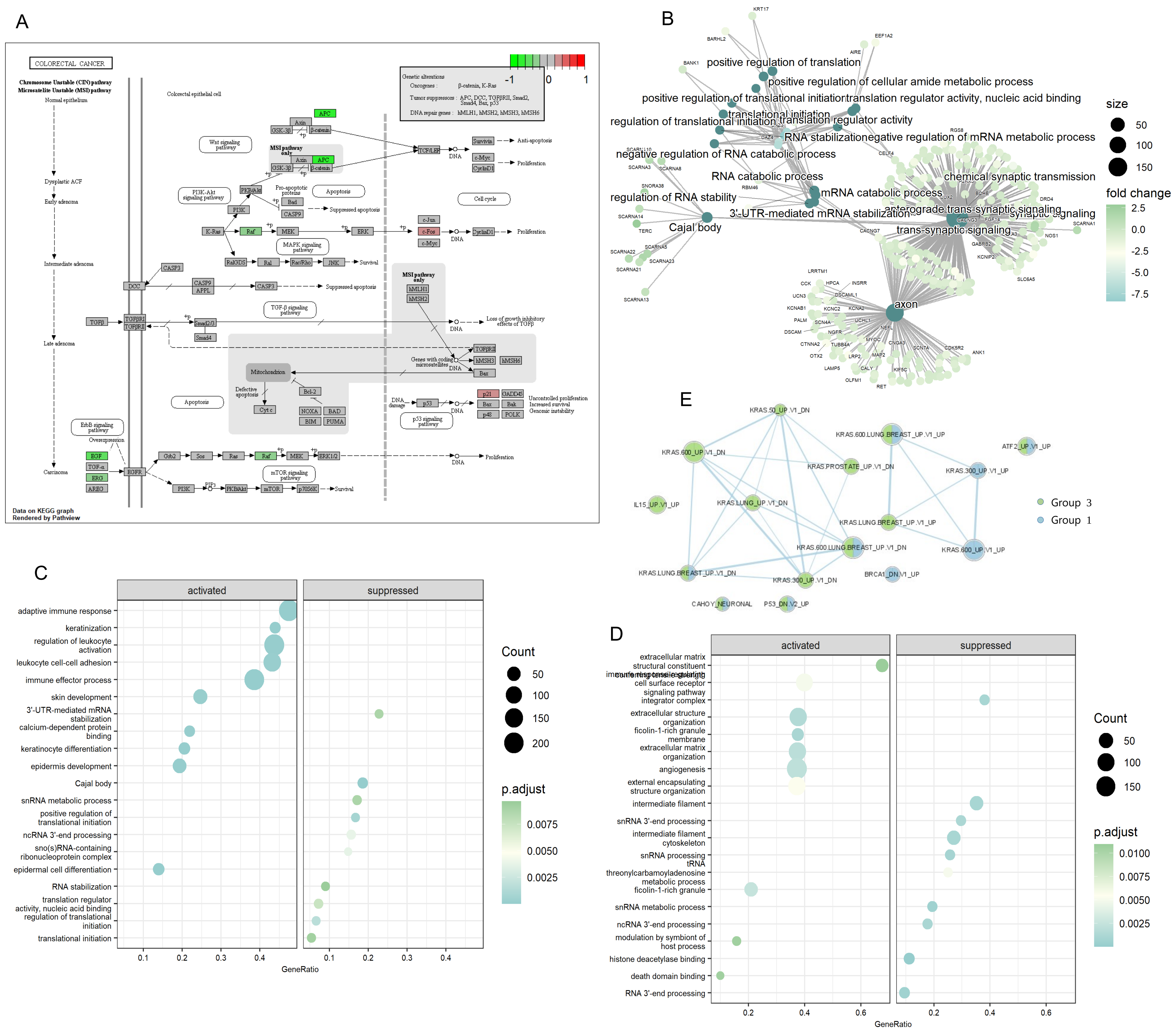


Fig.3. A. Group 1 KEGG Colorectal Cancer enriched pathway. B. Network of group 1 GO enriched pathways. C. Group 3 GO enriched pathways. D. Subgroup 2_2 GO enriched pathways. E. Group 3 and group 1 GSEA enriched pathways.
Group 3 is marked by the amplification of genes EGFR (adj.p-value < 0.001) and of the genes located on chr20 SRC (adj.p-value < 0.001) and AURKA (adj.p-value < 0.001), while it is affected by the deletion of the gene SKI (adj.p-value < 0.001). Amplification of EGFR and chromosome 20 have been shown to be associated with a better prognosis, while copy number increase of SKI is associated with a worse OS and DFS as compared to patients with no amplification or deletion of this gene [7]. Group 2.2 is characterized by patients with higher EGFR copy number and deletion of SKI. From the gene-set enrichment pathways analysis on the GO database, it can be noticed that in group 3 are mostly activated pathways related to the immune response. Group 1 shows differentially expressed genes involved in pathways related to the regulation of RNA translation. Group 2_2 depict the activation of pathways related to the extracellular activity organization and angiogenesis.

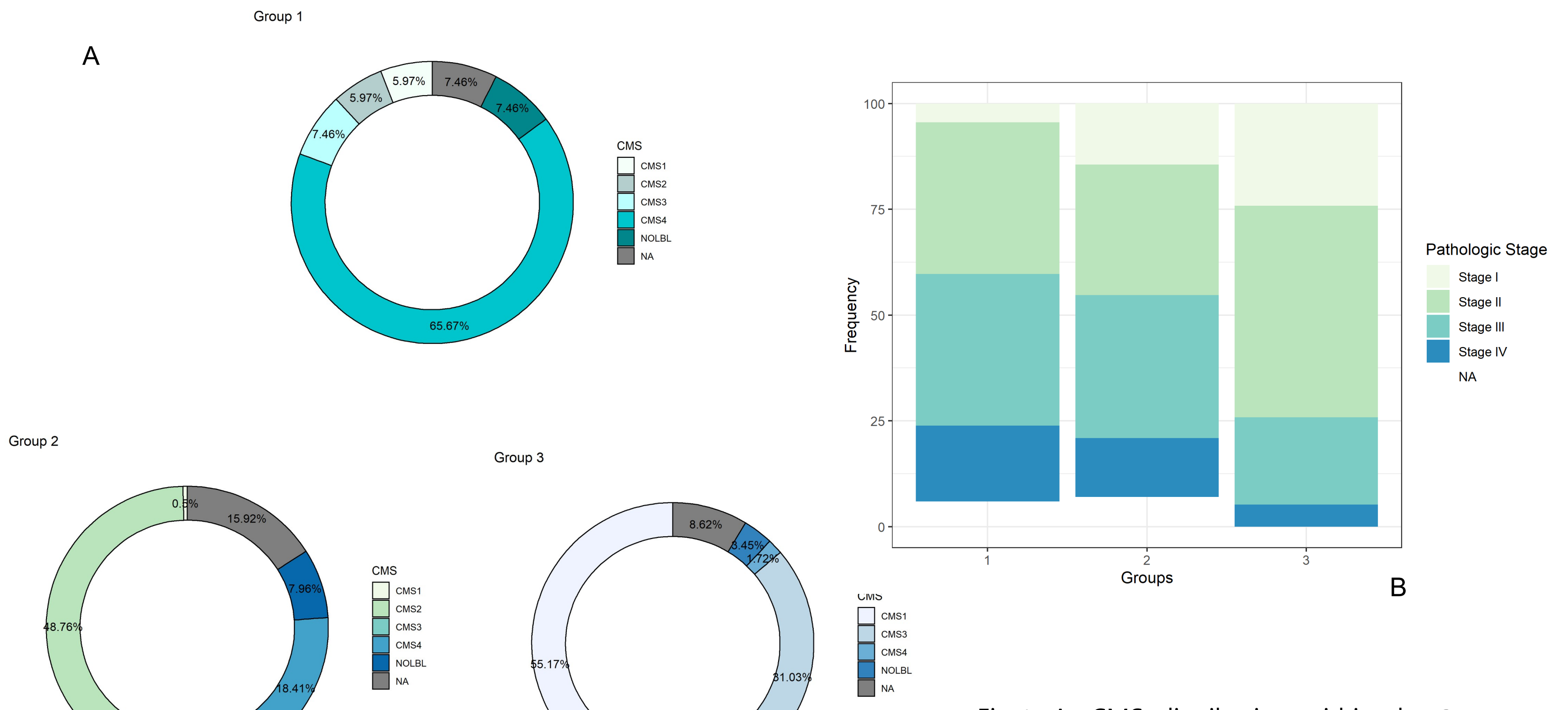
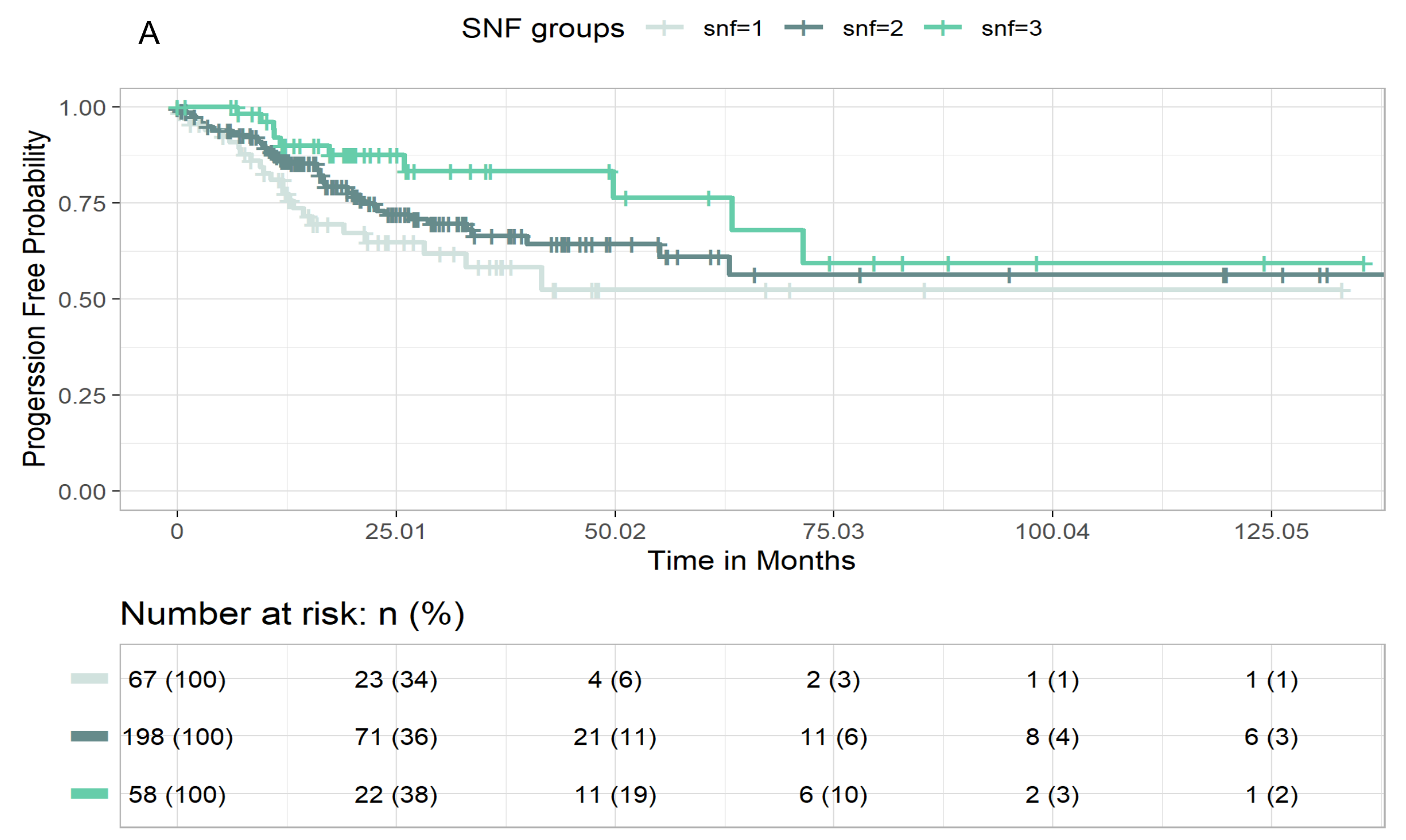


Fig.4. A. CMS distribution within the 3 groups obtained from the SNF algorithm.
B. Pathologic Stage distribution within the 3 groups obtained from the SNF algorithm.

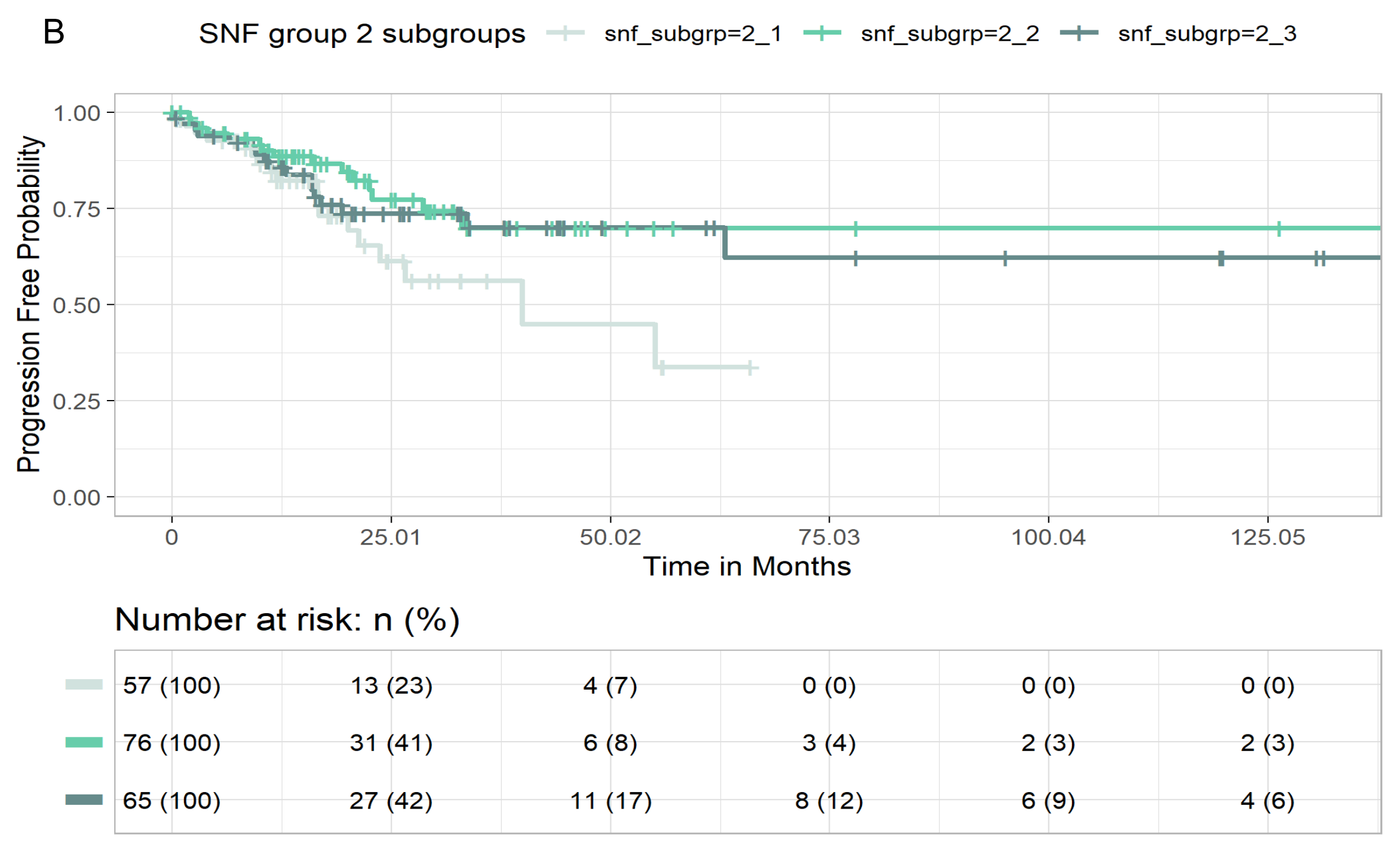
Patients in group 1 were more affected by lymphatic invasion (41.67% of patients, p-value 0.0309) if compared with groups 2 and 3 (27.06% and 20.07% respectively) and by venous invasion (32.76%, p-value 0.04) than second and third cluster (21.05.95% and 13.21% respectively). Group 3 was composed mostly by patients with N0 pathological stage N (75.86%, p-value 0.003), while the group 1 was the most enriched by patients with pathological stages N2 (28.36% against 19.2% in group 2 and 8.62% in group 3). Likewise, the percentage of patients with pathological stage I, was high in group 3 (24.14%, p-value 0.005) than in the other two clusters (4.76% group 1, 15.51% group 2). Ggroup 3 is composed for the 70.45% by patients assigned to the CRISA group and for the 25% to CRISB. Group 2 consists mostly of patients of CRISC (37.72%), CRISD (24.56%) and CRISE (24.56%), while Group 1 of patients of CRISA (21.28%), CRISB (31.91%) and CRISD (31.91%). Group 1 consist mostly of patients assigned to CMS4 (70.97%), group 2 of CMS2 (57.99%) and group 3 of CMS1 (60.38%)



Number at risk: n (%)

67 (100)	23 (34)	4 (6)	2 (3)	1 (1)	1 (1)
198 (100)	71 (36)	21 (11)	11 (6)	8 (4)	6 (3)
58 (100)	22 (38)	11 (19)	6 (10)	2 (3)	1 (2)

Fig.5. A.KM for Main Groups PFI estimation B.KM for subgroups of group 2 PFI estimation. Patients in group 3 have a lower risk to develop a progression if compared with those group 1 (Cox p-value: 0.022, HR: 0.42). These results are in line with that reported in [1]: subjects belonging to the CMS4 class (which characterizes group 1) report a lower proportion of event-free than those in CMS1 (which are more enriched in group 3). Further dividing group 2 using the SNF algorithm, we obtained 3 subgroups. Patients in subgroup 2_2 have a risk to develop a progression 50% less than those in group 2_1 (Cox p-value: 0.05).



Number at risk: n (%)

57 (100)	13 (23)	4 (7)	0 (0)	0 (0)	0 (0)
76 (100)	31 (41)	6 (8)	3 (4)	2 (3)	2 (3)
65 (100)	27 (42)	11 (17)	8 (12)	6 (9)	4 (6)

Conclusions

We applied Similarity Network Fusion Algorithm to 326 CRC patients aiming at define a samples clusterization and characterize inter-tumor heterogeneity. We saw that the integration of several omics sources produces groups that are differentially characterized in terms of both clinical and molecular properties.

Related Literature

[1] Guinney, J., Dienstmann, R., Wang, X. *et al. The consensus molecular subtypes of colorectal cancer*. Nat Med 21, 1350–1356 (2015). https://doi.org/10.1038/nm.3967
[2] Ried T, Meijer GA, Harrison DJ, Grech G, Franch-Expósito S, Briffa R, Carvalho B, Camps J. *The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome.* Mol Aspects Med. 2019 Oct;69:48-61. doi: 10.1016/j.mam.2019.07.007. Epub 2019 Aug 6. PMID: 31365882.
[3] Isella, C., Brundu, F., Bellomo, S. et al. *Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer.* Nat Commun 8, 15107 (2017). https://doi.org/10.1038/ncomms15107
[4] Wang, B., Mezlini, A., Demir, F. et al. *Similarity network fusion for aggregating data types on a genomic scale*. Nat Methods11, 333–337 (2014). https://doi.org/10.1038/nmeth.2810
[5] Chiu, A.M., Mitra, M., Boymoushakian, L. et al. *Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer.* Sci Rep 8, 11807 (2018). https://doi.org/10.1038/s41598-018-29992-5
[6] Robinson MD, McCarthy DJ, Smyth GK (2010). *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616
[7] Thomas Ried, Gerrit A. Meijer, David J. Harrison, Godfrey Grech, Sebastià Franch-Expósito, Romina Briffa, Beatriz Carvalho, Jordi Camps, The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome,Molecular Aspects of Medicine,Volume 69,2019, Pages 48-61, ISSN 0098-2997, https://doi.org/10.1016/j.mam.2019.07.007.