Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies

[Article published on Scientific Reports 7, Article number: 43946 (2017), doi:10.1038/srep43946, IF: 5.578]

Sara Ciucci[1,2,†], Yan Ge[1,†], Claudio Durán[1,†], Alessandra Palladini[1,2,3,†], Víctor Jiménez Jiménez[4], Luisa María Martínez Sánchez[1], Yuting Wang[5,6], Susanne Sales[5], Andrej Shevchenko[5], Steven W. Poser[7] , Maik Herbig[8], Oliver Otto[8], Andreas Androutsellis-Theotokis[6,7,9], Jochen Guck[8], Mathias J. Gerl[2] and Carlo Vittorio Cannistraci [1,*]

[1]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2]Lipotype GmbH, Tatzberg 47, 01307 Dresden, Germany
[3]Membrane Biochemistry Group, DZD Paul Langerhans Institute, Technische Universitaet Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[4]Integrin Signalling Group, Fundación Centro Nacional de Investigaciones Cardiovasculares Carlos III, Melchor Fernández Almagro 3, 28029 Madrid, Spain
[5]MPI of Molecular Cell Biology and Genetics, Pfotenhauerstrstraße 108, 01307 Dresden, Germany
[6]Center for Regenerative Therapies Dresden (CRTD), Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, Fetscherstraße 105, 01307 Dresden, Germany
[7]Department of Internal Medicine III, University Hospital Carl Gustav Carus at the Technische Universität Dresden, Fetscherstr.74, 01307 Dresden, Germany
[8]Cellular Machines Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[9]Department of Stem Cell Biology, Centre for Biomolecular Sciences, Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham NG7 2RD, U.K.
†These authors contributed equally to this work.
*Correspondence should be addressed to: kalokagathos.agon@gmail.com

Keywords: Principal component analysis - Network-inference - Discriminative functional network modules

Omic science is rapidly growing and one of the most employed techniques to explore differential patterns in omic datasets is principal component analysis (PCA). However, a method to enlighten the network of omic features that mostly contribute to the sample separation obtained by PCA is missing. An alternative is to build correlation networks between univariately-selected significant omic features, but this neglects the multivariate unsupervised feature compression responsible for the PCA sample segregation. Biologists and medical researchers often prefer effective methods that offer an immediate interpretation to complicated algorithms that in principle promise an improvement but in practice are difficult to be applied and interpreted. Here we present PC-corr: a simple algorithm that associates to any PCA segregation a discriminative network of features. Such network can be inspected in search of functional modules useful in the definition of combinatorial and multiscale biomarkers from multifaceted omic data in systems and precision biomedicine. We offer proofs of PC-corr efficacy on lipidomic, metagenomic, developmental genomic, population genetic, cancer promoteromic and cancer stem-cell mechanomic data. Finally, PC-corr is a general functional network inference approach that can be easily adopted for big data exploration in computer science and analysis of complex systems in physics.

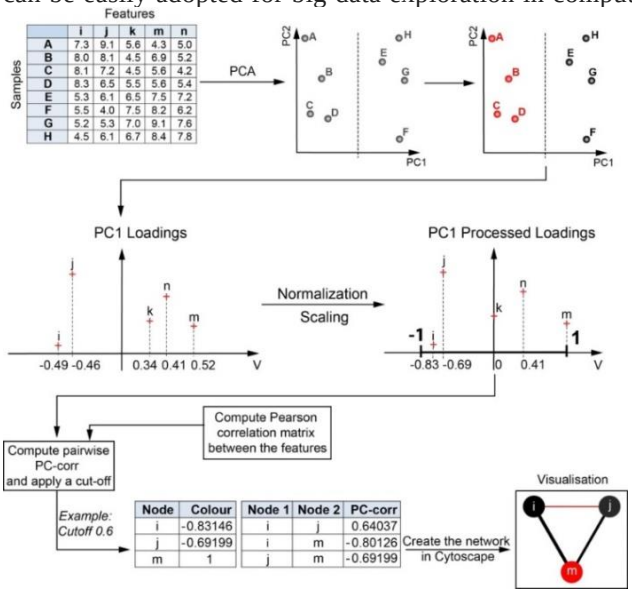

Figure 1. PC-corr network. The omic dataset is analysed unsupervisedly by PCA, which discriminates the samples in two groups (red and black) along PC1. The discriminative correlation network is constructed thereof. Indeed the features' loadings of the most discriminative dimension (PC1 in the example) are normalized with the heuristic function and scaled, then they are combined together with the Pearson correlations on the features, for the calculation of the edge values $\boldsymbol{PC\_corr_{i,j}}$. Finally, after applying a cut-off on the edge weights $\boldsymbol{PC\_corr_{i,j}}$, two tables (Edge weight/colour table and Node weight/colour table) are obtained in order to visualize the related network by a network visualizer such as Cytoscape.

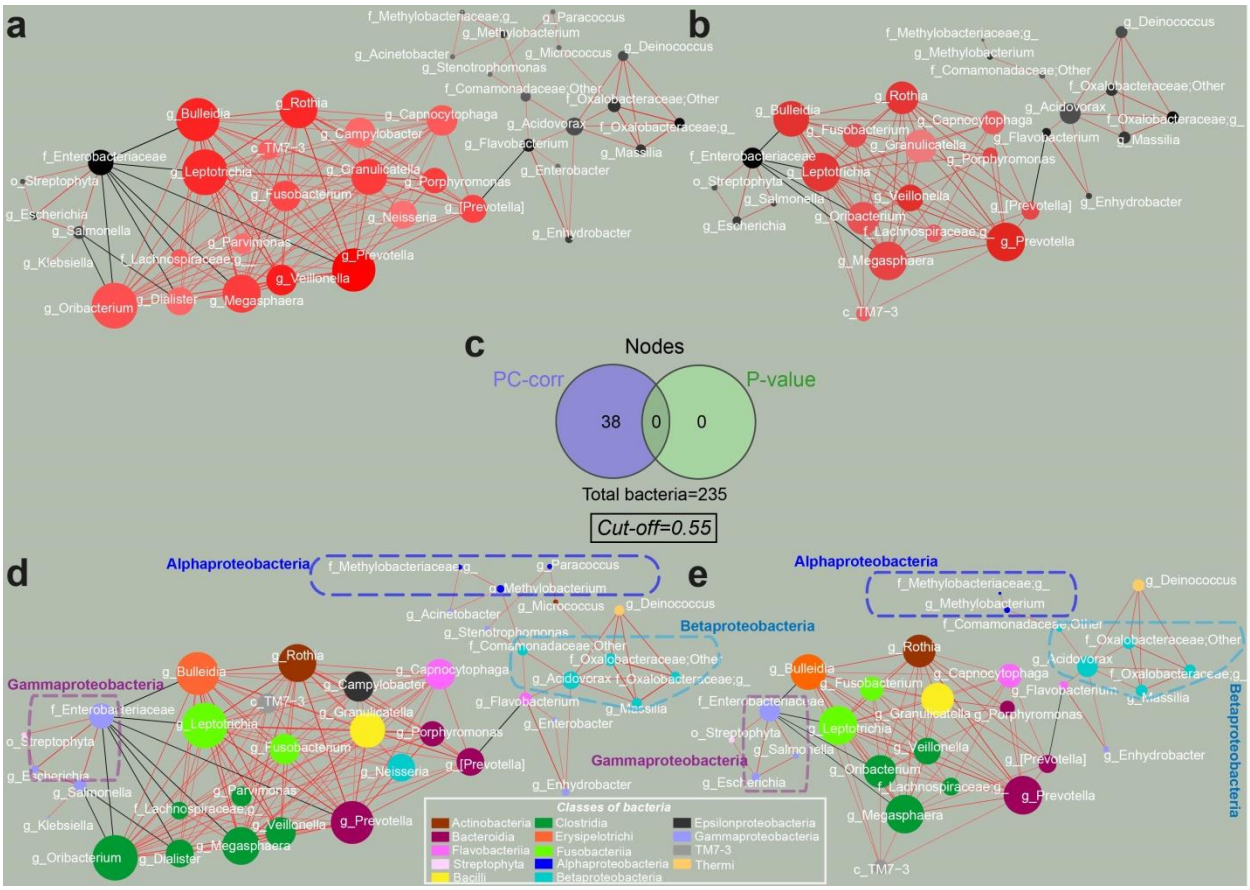

Figure 2. PC-corr networks at two different cut-offs. (**a**) The PC-corr network was constructed at cut-off 0.55 from PC2 loadings since the discrimination of samples before and after PPI treatment was significant along PC2 of PCA. (**b**) As in (a), the PC-corr network was also constructed at cut-off 0.6 from PC2 loadings. (**c**) The PC-corr network can find interactions between the bacteria while the P-value cannot, as illustrated in the Venn diagram for the number of nodes. (**d**) PC-corr network in (a) coloured according to class-level taxonomy. The coloured dashed lines highlight the classes of bacteria that constitute the black submodules. (**e**) PC-corr network in (b) coloured according to class-level taxonomy. The coloured dashed lines highlight the classes of bacteria that constitute the black submodules.

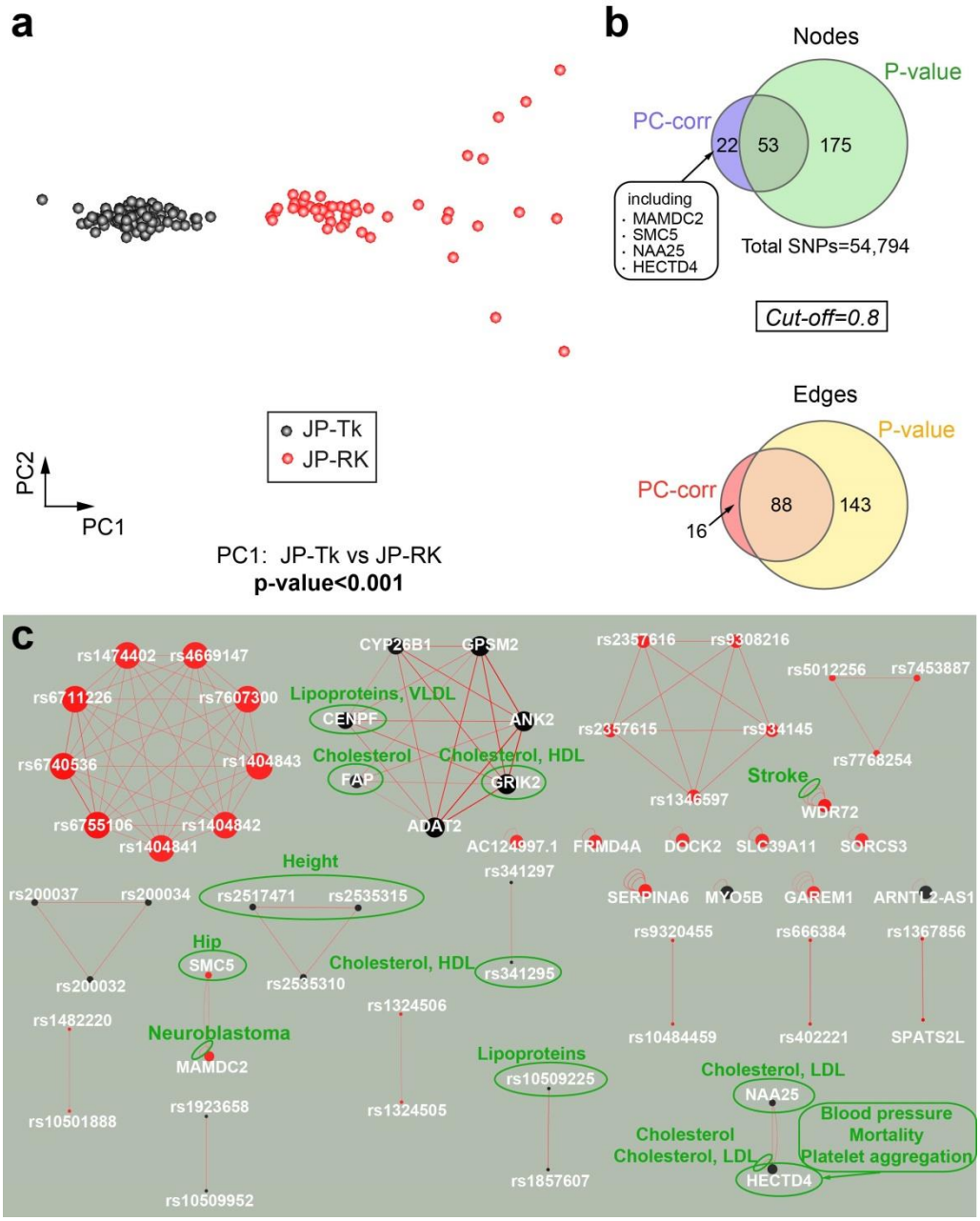

Figure 3. Genetic structure of the Japanese population based on autosomal SNP genotyping. (**a**) PCA discriminated the Japanese population into Tokyoites (JP-Tk) and Okinawans (JP-RK). (**b**) For cut-off = 0.8, the SNP difference between the PC-corr and P-value networks in the nodes and edges is illustrated by the proportional Venn diagrams. Some SNPs are only present in the PC-corr network, like the ones indicated by the genes that contain the mutation (box on the right). Some interactions among the SNPs are also missing in the P-value network. (**c**) As shown in (a), PCA detects the difference between JP-Tk and JP-RK along PC1, thus its loadings are employed for the construction of SNP PC-corr network. The SNP name (rs#, in white colour) is changed to the name of the gene that contains the mutation when it is an intron, missense or 3′ UTR variant. Otherwise, it is left unchanged. Some SNPs were significantly associated to phenotypes (green circles and text).

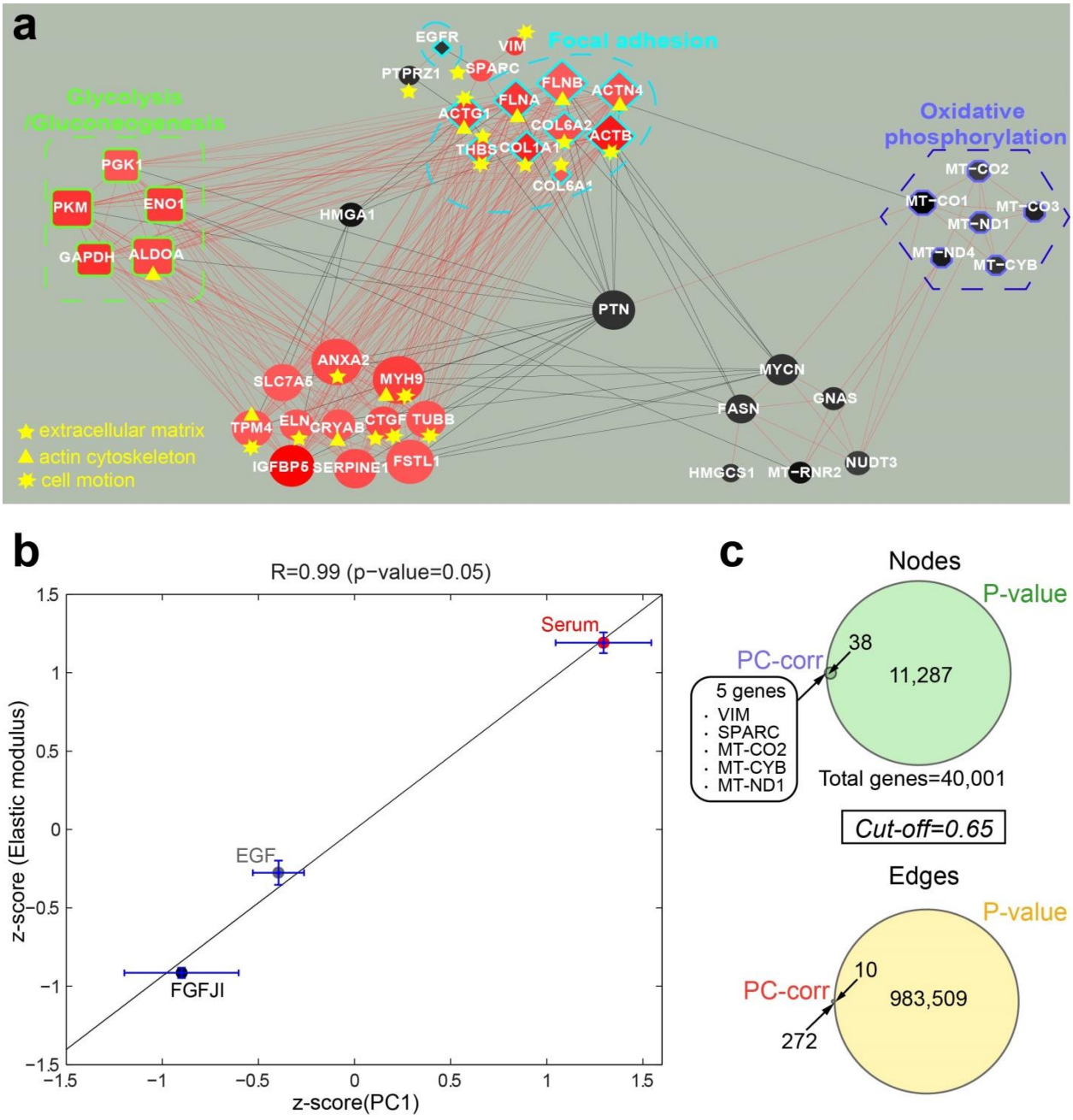

Figure 4. Cancer stem cells gene expression pattern and correlation with cell mechanics. (**a**) PC-corr gene network constructed using cut-off 0.65. The three looped-dashed lines indicate three significant pathways, while yellow different symbols highlight the enriched genes classified as GO terms. (**b**) Linear regression plot of z-score of PC1 and z-score of elastic modulus for each cell culture condition. Each point represent the PC1-score-average vs the elastic-modulus-average of all the samples in the given cell culture condition. The standard error for each computed average is reported as an error bar. (**c**) For cut-off = 0.65, the gene difference between the PC-corr and P-value networks in the nodes and edges are illustrated in the proportional Venn diagrams. Some genes are only present in the PC-corr network, like the genes reported in the box on the left. Only 10 gene interactions present in the PC-corr are present in the P-value network too.