CANCER PROTEOMICS

Prognostic hallmarks in AML

A data-clustering method that incorporates prior knowledge of biological context reveals prognostic signatures of proteomic expression in patients with acute myeloid leukaemia.

Matthew A. Clarke and Jasmin Fisher

The prognosis of acute myeloid leukaemia (AML) currently relies on markers of cytogenetic abnormalities, on the patients' ability to perform specific tasks, and on age, white-blood-cell counts and other biomarkers[1]. These evaluations are sufficient to stratify patients into three groups: favourable prognosis, intermediate prognosis, and unfavourable prognosis[2]. New data sources (such as the number of mutations in the genes *FLT3*, *CEBPA* and *NPM1*; ref. [3]) are being explored to improve prognosis accuracy and to enable patient-specific treatment decisions. For example, recent strategies enable the stratification of patients with normal cytogenetics yet diverse survival rates[3]. Advances in gene-expression profiling are also better capturing the inter-patient and intra-patient variabilities in the mutations driving cancer[3,4]; however, these techniques cannot capture mechanisms of post-translational modification and regulation[5], as the expression of a gene is not a reliable indicator of changes at the protein level.

Resolving the structure of biological systems, especially in single-cell transcriptomics, remains a challenge[6], despite the availability of methods for dimensionality reduction and feature selection. Clusters based on similar gene-expression profiles may be associated with different phenotypes, as small changes in expression, protein abundance and protein splicing that may be missed by clustering tools may nevertheless have disproportionally large effect sizes. Furthermore, finding a small set of predictive factors from many hundreds of candidates is statistically difficult (although there have been attempts to solve this problem[7]). All these challenges associated with big-data analyses are exacerbated by the increased complexity inherent of proteomics data. Amina Qutub, Steven Kornblau and colleagues now report in *Nature Biomedical Engineering* the quantification of the heterogeneous nature of protein expression in 205 samples from patients with AML and in 111 leukaemia cell lines, as well as the identification of proteomic hallmarks for patient stratification[8].

Qutub and co-authors' approach — termed MetaGalaxy analysis[9] — pre-groups proteins that, on the basis of prior knowledge (from the expert-curated Kyoto Encyclopedia of Genes and Genomes; KEGG; refs [10,11]), are known to be related, and clusters the protein groups via unsupervised methods on the basis of protein-expression patterns (Fig. 1). The authors show how the clusters obtained can be used to stratify patients with specific proteomic signatures that are predictive of overall survival and of the duration of remission, and that these predictions remain significant when adjusting for existing prognostic factors. The MetaGalaxy analysis clustered 228 proteins into 31 functional groups by first taking into account known associations and similar functions in the KEGG database. The authors then identified 154 recurring functional patterns of expression (which they termed 'constellations') within the predefined functional groups by Progeny Clustering[12] (a bootstrapping-based method). Some functional patterns were sufficient for the stratification of patients into groups with significantly different prognosis; for example, differences in the functional patterns present in the 'hypoxia' functional group could be used to stratify patients (Fig. 2), even patients belonging to an unfavourable cytogenetics group, and in the absence of other clinical data that so far has been necessary for the stratification of AML patients. Interestingly, these functional groups did not seem to serve as proxies for known markers, as they did not match existing groupings (such as the French–American–British class) nor genetic mutations (such as *FLT3*) known to be prognostic markers[3]. Altogether, the 205 patients in the cohort exhibited 11 clusters of functional patterns that enabled their allocation into 13 groups with common signatures (defined by the recurrence of the clusters of functional patterns) that were prognostic for survival and remission duration. The clustering remained significantly different even when adjusting for associations with existing clinical factors (age, cytogenetics, and white-blood-cell count). Furthermore, MetaGalaxy enhanced the stratification of patients with respect to the analysis of the same data by k-means clustering or by hierarchical clustering. This implies that the authors' approach may be identifying novel protein sub-groups with distinct prognosis, and therefore likely to represent distinct biological factors and requirements for treatment.

Proteomics analyses are more complex than genomics and transcriptomics analyses, because proteomics analyses need to account for alternate splicing, post-translational modifications and other variables[13]. Yet the increased complexity may hold useful answers in the understanding of disease mechanisms. For example, Qutub and colleagues found that changes in post-translational modifications do not match protein-expression levels, and thereby that they would not be predicted by transcriptomics data. Despite these advantages, high-throughput gene-expression data is likely to remain cheaper to acquire, and will cover a wider range of genes for the foreseeable future[14] (in fact, the Reverse Phase Protein Array panel used in Qutub and colleagues' study covered 228 proteins, whereas RNA-Seq is routinely used to profile all expressed genes in a tissue or cell). It thus remains to be seen whether proteomics data can provide more accurate information for AML prognosis than transcriptomic methods[3,4].

Qutub and co-authors analysed the similarity (and therefore, appropriateness) of cell-line models for the study of AML. Notably, of the functional patterns carrying prognostic value identified in patient-derived samples, a majority (52.6%) were unrepresented in the panel of 111 common leukaemia cell lines profiled, uncovering further limitations of the use of cell lines in AML research. The results of these comparisons were placed in a publicly accessible database (the AML Proteome Atlas; https://www.leukemiaatlas.org/code), and should serve as a support tool for researchers interested in choosing the most appropriate cell line for the pathway, process or type of AML of interest. The tool might enable further optimization of the MetaGalaxy approach, for example via the use of alternative prior-knowledge sources to inform the selection of functional groups.

Matthew A. Clarke[1] and Jasmin Fisher[1,2]*
[1]Department of Biochemistry, University of Cambridge, UK
[2]UCL Cancer Institute, University College London, UK
*e-mail: jasmin.fisher@ucl.ac.uk

References

1. Liersch, R., Müller-Tidow, C., Berdel, W. E. & Krug, U. *Br. J. Haematol.* **165**, 17–38 (2014).
2. Grimwade, D. *et al. Blood* **92**, 2322–2333 (1998).
3. Mrózek, K., Marcucci, G., Paschka, P., Whitman, S. P. & Bloomfield, C. D. *Blood* **109**, 431–448 (2007).
4. Bullinger, L. *et al. N. Engl. J. Med.* **350**, 1605–1616 (2004).
5. Guhaniyogi, J. & Brewer, G. *Gene* **265**, 11–23 (2001).
6. Ding, J., Condon, A. & Shah, S. P. *Nat. Commun.* **9**, (2018).
7. Fan, J. & Li, R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. in *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006* 595–622 (European Mathematical Society Publishing House, 2006). doi:10.4171/022-3/31
8. Hu, C. W. *et al. Nat. Biomed. Eng.* **3**, XXX–YYY (2019).
9. Hoff, F. W. *et al. Proteomics Clin. Appl.* **13**, e1800133 (2019).
10. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. *Nucleic Acids Res.* **47**, D590–D595 (2019).
11. Kanehisa, M. & Goto, S. *Nucleic Acids Res.* **28**, 27–30 (2000).
12. Hu, C. W., Kornblau, S. M., Slater, J. H. & Qutub, A. A. *Sci. Rep.* **5**, 1–12 (2015).
13. Manzoni, C. *et al. Brief. Bioinform.* **19**, 286–302 (2018).
14. Hegde, P. S., White, I. R. & Debouck, C. *Curr. Opin. Biotechnol.* **14**, 647–651 (2003).

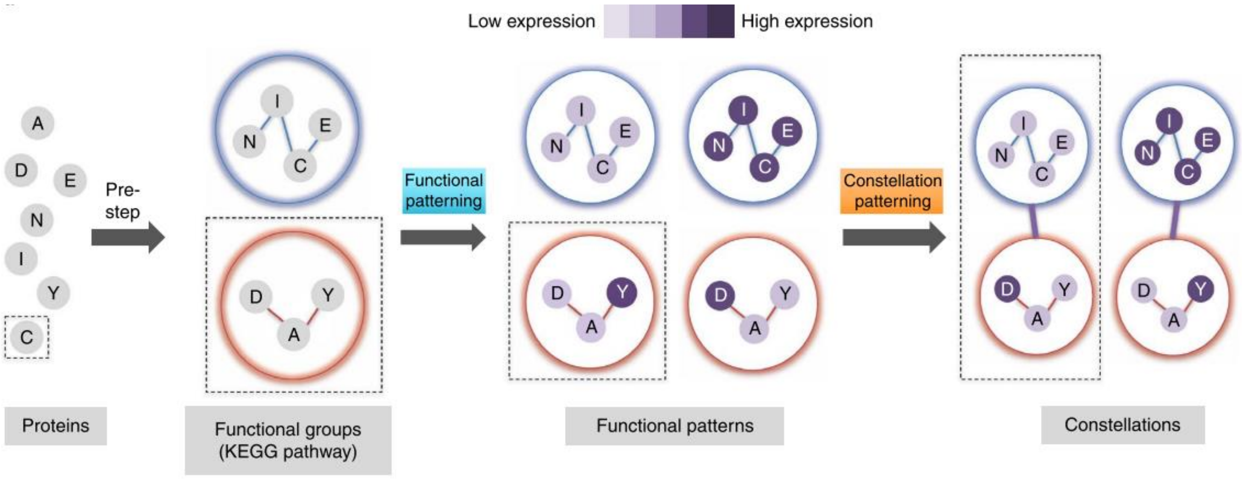

Fig. 1 | Overview of the MetaGalaxy analysis. Proteins are first gathered into functional groups based on known associations and similar functions according the KEGG database[10,11]. Functional patterns of expression within the groups are identified by k-means clustering and by Progeny Clustering[12] (two types of unsupervised clustering algorithms). Recurring groups of functional patterns (‘constellations’) are identified in patients using the same combination of k-means clustering and Progeny Clustering. Figure reproduced from ref. [8].

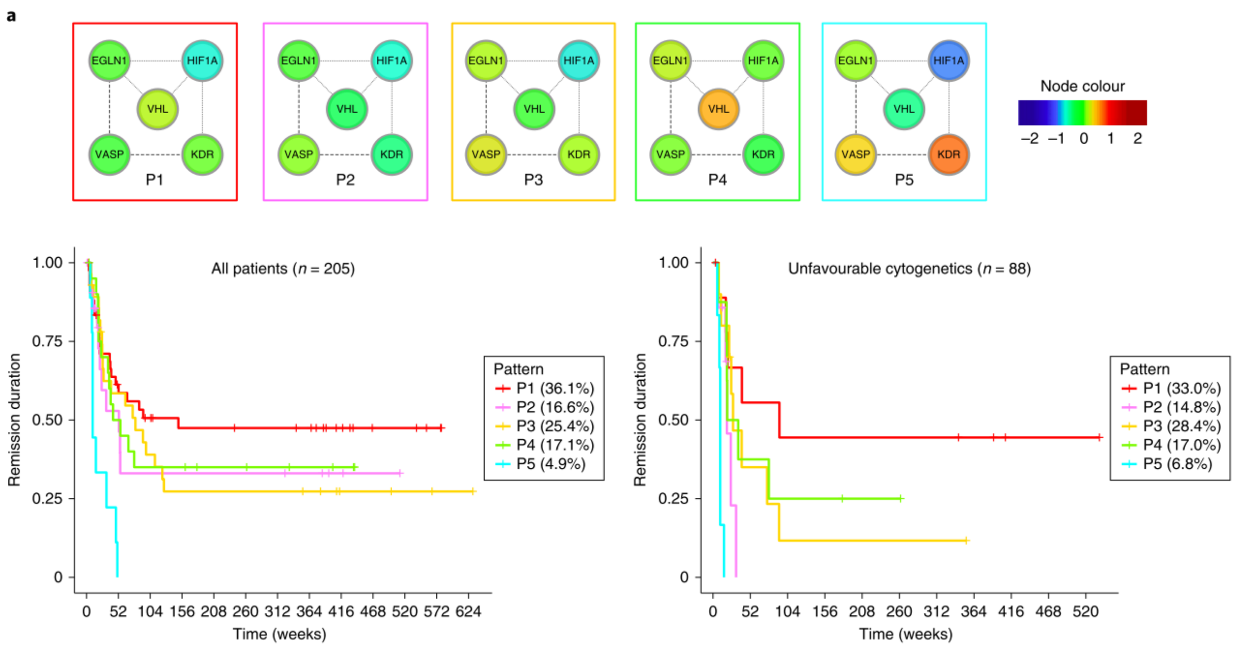

Fig. 2 | Functional patterns observed in the hypoxia functional group. a, Functional patterns (P1 to P5) found in the hypoxia functional group of proteins (each protein is indicated as a node in the protein network). The colour of a node indicates the expression levels of the gene with respect to the expression-level average in control samples. **b**, The functional patterns stratify all patients (left), including the subgroup of patients with unfavourable cytogenetics (right), according to the duration of remission. The percentage figures indicate the fraction of patients in each functional pattern. Figure reproduced from ref. [8].