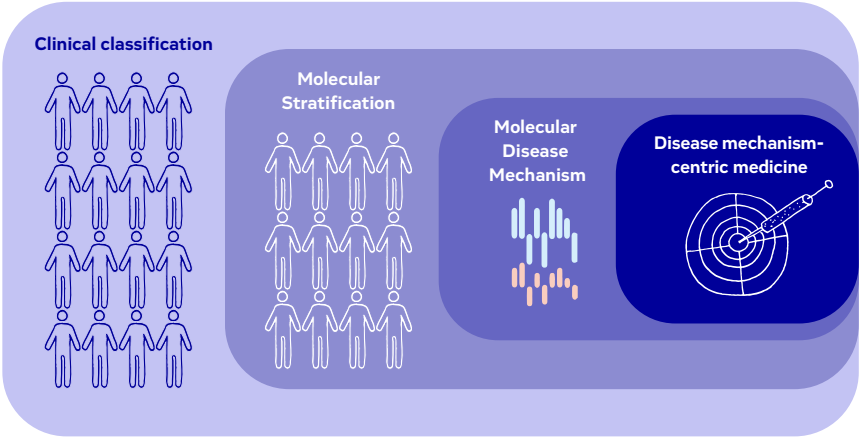
Mining omics data for drug discovery

Authors: Johannes Pospiech Ph.D. & Sven Sauer Ph.D.

This review addresses patient-centric Omics mining strategies to enable target identification for disease mechanism-centric medicine. The aim of the paper is to outline the process of identifying potential drug targets through detailed analysis of patient biopsy transcriptomes exemplified, in part, by data obtained from a chronic kidney disease cohort[1,2].

Most human diseases are multifactorial and present with a complexity that, until recently, could only be managed by basing a diagnosis and prognosis on certain accessible signs and symptoms. A typical example is the spectrum of disorders that are classified as chronic kidney disease (CKD) with five stages of disease based on estimated glomerular filtration rates (eGFR). While such classification facilitates a more precise understanding of the epidemiology of a disease, the fact that the disease outcomes are dependent on a multitude of factors such as, in the case of CKD, the variety of cell types and molecular pathways involved in renal function, there is a significant mechanistic gap between the clinical categorization of CKD and the presumed drivers of the disease. Such a situation means that there are critical restrictions on the ability to model the disease concerned and, in turn, on the ability to discover and develop effective therapeutic interventions.

State of the art systems biology can be used to integrate complex data sets arising from genomics, proteomics, transcriptomics and metabolomics and to create a verifiable model of a complex disease and the dynamics of its progression. This process represents a cornerstone of disease mechanism-centric medicine, a concept that considers individual variability when investigating disease complexity and therapeutic strategy.

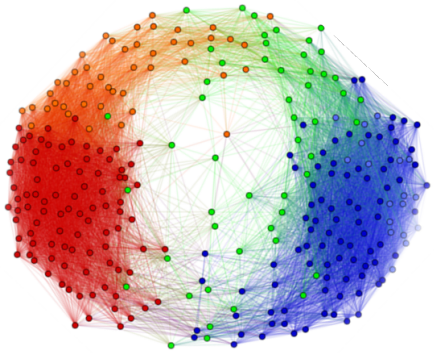


1 Bohnenpoll, T. et al. (2022): FC080: A Systems Nephrology Framework for the Molecular Classification of Chronic Kidney Disease. Nephrology Dialysis Transplantation 37, gfac114.004. 10.1093/ndt/gfac114.004.

2 Bohnenpoll, T. et al. (2022): Unsupervised Characterization of the NURTuRE Cohort Reveals Gene Expression and Tissue Remodeling Dynamics Along a Synthetic CKD Progression Axis [Abstract]. J Am Soc Nephrol 33, 2022: 883.

In the following paragraphs, we will be discussing the role of our technologies that have been used to mine molecular datasets obtained from CKD-related patient cohorts and allied biomedical annotations in order to generate target candidates for our PanOmics Target Identification Framework (*link to whitepaper*) that will enable target selection, transition to drug discovery and IND[1, 2]. Discussing the concepts behind this patient omics analysis strategy, we suggest that they are equally applicable to other patient cohorts and, with high probability, to other chronic diseases.

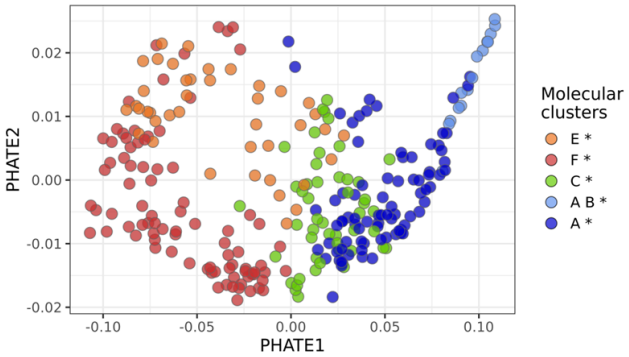
The first step in the process involves generating data clusters based on gene expression profiles in the patient-related, big-data set. These clusters, or molecular groups, are computed by a machine-learning process that is completely unbiased and data-driven, independent of any clinical analysis of the patient data. In other words, the applied algorithms structure transcriptome data into groups based on relationships – similarities and differences – between the individual data points. This unsupervised clustering allows the identification of polarized configurations that stratify the global molecular disease landscape in an unbiased manner (Figure 1a). Further characterization of this landscape is necessary before a mechanistic understanding of a disease can be fully appreciated and applied to the development of disease models.



Highly polarized structure defines global molecular landscape

Figure 1a: *Unsupervised clustering of kidney biopsy transcriptomes*

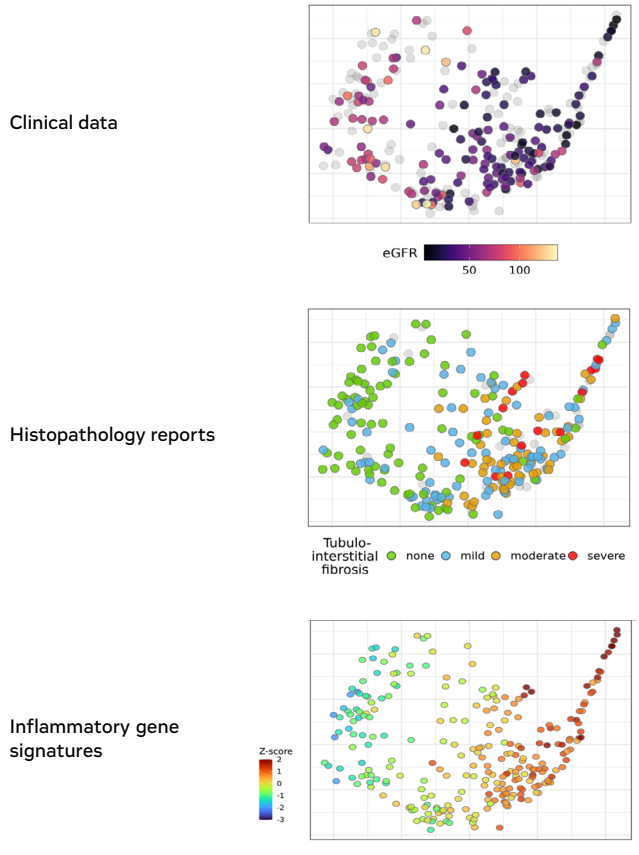
The interaction of multiple factors (e.g. age, disease stage, gender, genetic modifiers) with the actual disease-causing, molecular mechanism give the cluster landscape its exclusive level of complexity. The application of dimension-reducing algorithms lowers the level of complexity so that the relationships between the molecular groups can be better visualized, building a molecular map out of the disease landscape (Figure 1b).



Dimension reduction further structures the global molecular map

Figure 1b: *Non-linear dimension reduction of kidney biopsy transcriptomes*

An understanding of the relationships within the global molecular map can be obtained by overlaying it with clinical or morphological data originating from patient and sample annotations or the expression of disease-relevant gene profiles. In our example, the map of kidney biopsy transcriptomes has been overlayed with eGFR data, with scores of tubular fibrosis, and with immune response gene signatures (Figure 2).



Molecular changes align with clinical disease progression and inflammatory tissue status

Figure 2: *Integration of clinical and morphological patient and sample annotations*

The integration of these data can be associated with the polarized molecular structure from left to right, indicating an alignment of the molecular changes with clinical disease progression. In this way, the unbiased molecular stratification is contextualized by phenotypical data, integrating all sources of annotation within one harmonized framework. In our example of CKD, early disease states can be located to the left with a gradual shift towards a severely diseased state to the right.

Thus, by integrating all the available information from multiple sources, we are able to infer a disease progression axis which represents a holistic description of the course of a disease (Figure 3). By having access to this description of the actual disease process we are able to develop hypotheses about the mechanisms of the disease and conceptualize specific disease drivers.

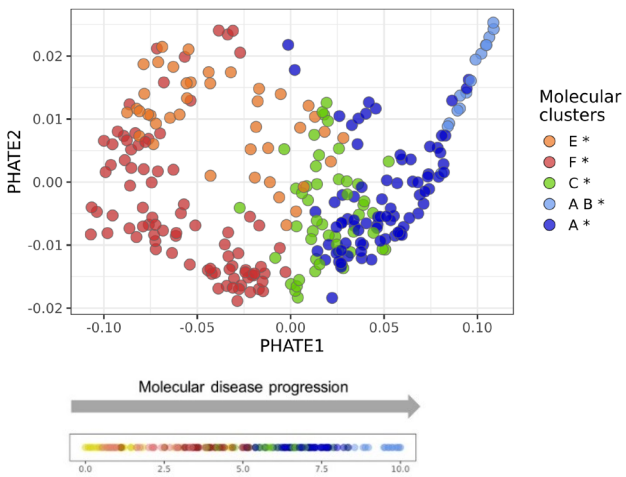
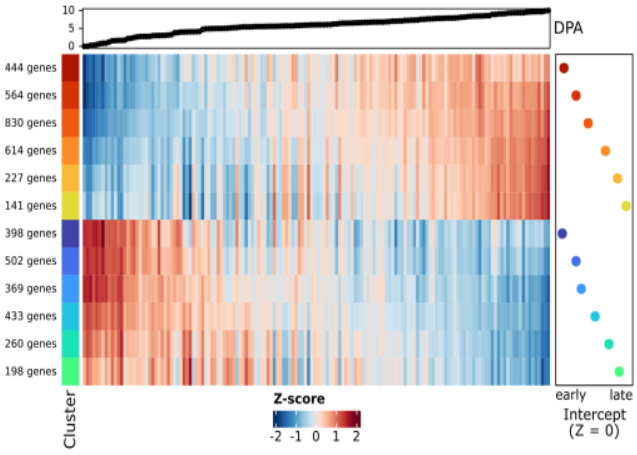
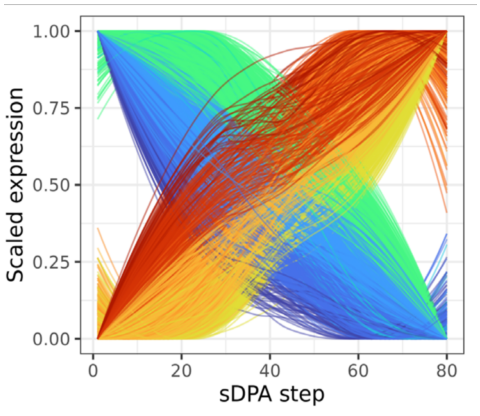


Figure 3

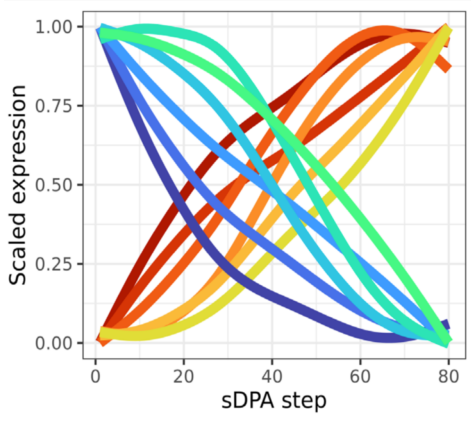
It is now possible to examine the expression dynamics within the transcriptomics data in order to identify genes that are regulated according to their position on the synthetic disease progression axis (sDPA). This pseudotemporal analysis results in a series of expression trajectories of genes associated with the disease progression axis (Figure 4). The ordering of the data in this way thus creates an synthetic axis for the progression of the disease that enables us to examine clusters of expression trajectories and how they behave over time, without the requirement for actual longitudinal data of the same patient or the same sample . By focusing on different sections of the disease progression axis we are able to separate early disease processes from later ones which facilitates mechanistic interpretation.



Gene highly correlated with the sDPA (-5000) were clustered hierarchically by their expression dynamics resulting in 12 clusters with different pseudotemporal change.



Expression trajectories of all genes that are correlated with sDPA



Clustered expression trajectories for pseudotemporal classification

Figure 4

Conclusion

In summary, we have shown how it is possible to create a pseudotemporal axis for the progression of a complex, multi-factorial disease and to use the resulting parameters to obtain a mechanistic analysis of patient-centric data and a representation of disease drivers that could lead to the identification of novel targets for therapeutic intervention. Furthermore, omics-based analysis of patient transcriptomes and an understanding of disease mechanisms will help to identify patients who will be expected to respond favorably to the targeted therapy and provide an important step in achieving the goals of disease mechanism-centric medicine.

Figure 5*: Workflow summary*