Universitätsklinikum Hamburg-Eppendorf

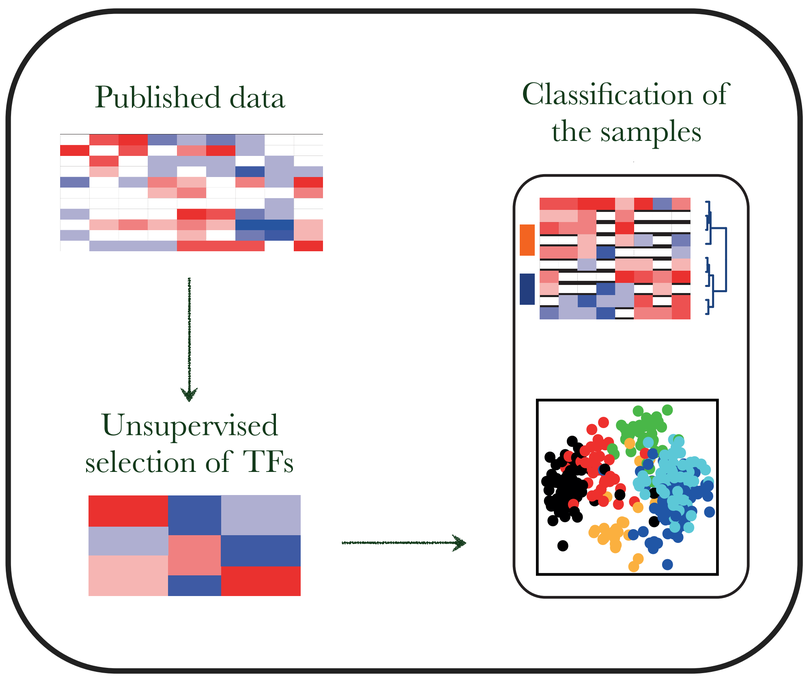
Center for Internal Medicine

Transcription factors to classify tumor types and subtypes

Benjamin Otto[1,2], Kristin Klätschke[2], Thomas Streichert[3], Christoph Wagener[2], Genrich Tolstonog[4]

1. Center for Internal Medicine, University Medical Center Hamburg-Eppendorf, Germany
2. Department of Clinical Chemistry, University Medical Center Hamburg-Eppendorf, Germany
3. Department of Clinical Chemistry, University Medical Center Cologne, Germany
4. Department of Otolaryngology – Head and Neck Surgery, CHUV, University of Lausanne, 1011 Lausanne, Switzerland

The development progress of whole genome analysis technologies such as microarrays or next generation sequencing, combined with bioinformatics, revealed the complexity of molecular biological regulatory mechanisms. The understanding of epigenetics, miRNAs or transcription factors and their relevance for control of the transcription process has since increased. Especially transcription factors play a central role in the regulatory network. Acting downstream of all signaling pathways they act similar to small interlocking systems receiving and bundling diverse incoming signals and translating them into the observed phenotype.



In a recent Publication we introduced a novel approach for the use of transcription factor coding genes to unveil breast tumor subtypes and to link murine phenotypes with potential human homologues. Two pathes can be followed, a supervised analysis to investigate mechanisms related to defined single factors or an unsupervised analysis to determine new key factors.

All genes → Filter all transcription factors

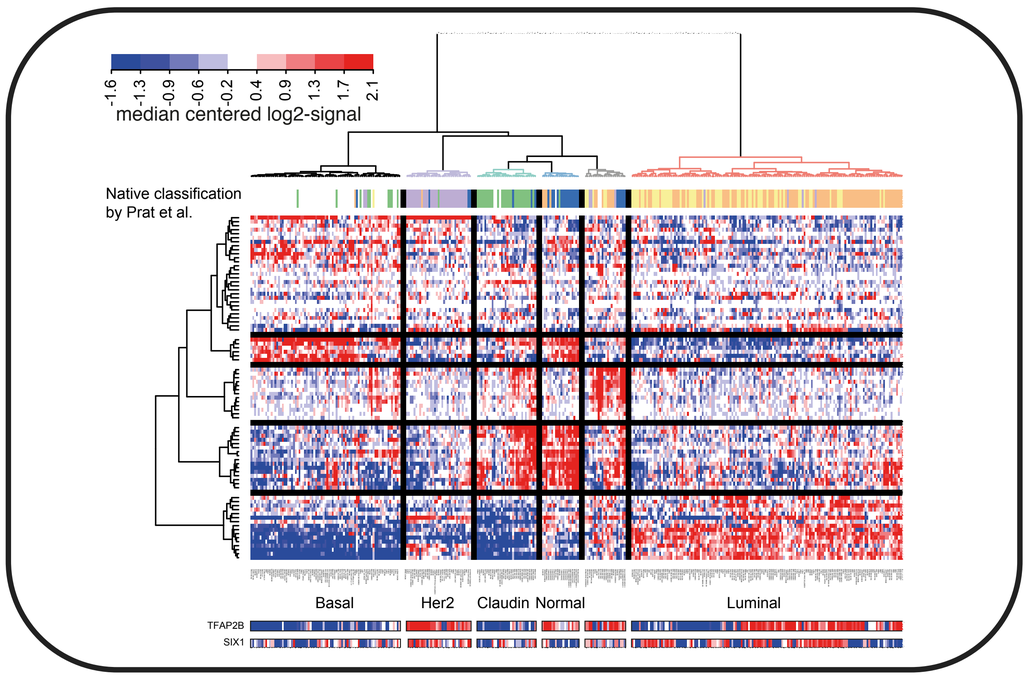
$$G_i = \begin{cases} F_i[z_{ij} \geq 1] \geq 0.8N \\ F_i[z_{ij} \geq 2] \geq 0.4N \\ \tilde{z}_{ij} \geq 0.8N \\ F_i[z_{ij} \geq 4] \geq p \end{cases} \Bigg| \; j \in N, i \in M, activeTFs\ G_i \in All\ TF genes$$

$$z_{ij} = |x_{ij} - \bar{x}_i| \qquad \tilde{z}_{ij} = \sum_j z_{ij} \Big| z_{ij} > Q_{i(1-0.8)}$$

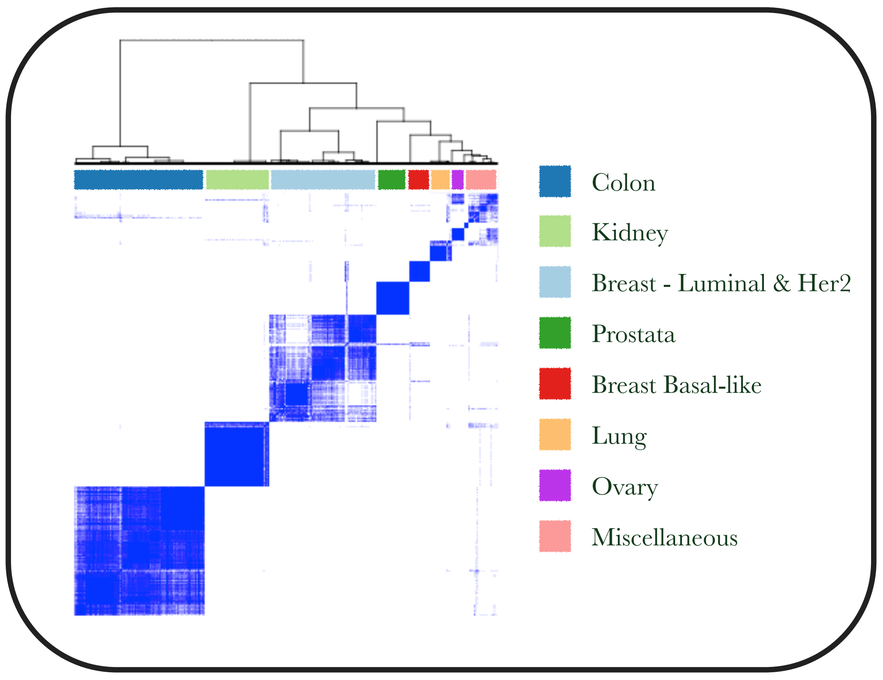
N number of samples. M number of genes.

z_{ij} absolute median centered gene expression of gene i.

Here we introduce the use of the unsupervised approach to identify transcription factors (TFs) that are specific for different tumor types. The approach identifies „active“ TFs in a dataset. TFs are scored as „active“ if they meet one of four criteria that resemble a clear differential expression in the majority of the samples, a very strong differential expression in a solid subset, a clearly heterogenous expression over the dataset or an extremly high expression in a small subset of the samples.



Application of the approach to a human breast cancer dataset (UNC337) in our publication revealed a set of 119 TFs that are capable of perfectly separating the classic breast cancer subtypes (luminal, basal-like, Her2, Claudin-low and Normal-like). Aside well known markers such as the estrogen receptor, a number of new potential key factors could be identified such as ELF5, SIX1 or TFAP2B.



In a second step we applied the method to a dataset of 1,000 tumors samples extracted from the expO collection (GSE2109) and used a consensus clustering analysis to cluster the samples. The resulting signature of TFs proved capable of 1) separating the different tumor types extremely well and 2) to identify even relatively small subsets (e.g. ovary cancer) as independent group.

Tumor type	**Associated TFs**
Colon	*PROX1, FOXA3, KLF5, FOXA2*
Kidney	*TCF2, HOXC10, NR1H4, PAX2*
Breast Luminal	*TFAP2B, ESR1, GATA3, FOXA1*
Breast Basal-like	*FOXC10, MYBL1, ELF5, SOX10*
Prostata	*ZBTB16, ESRRG, IRX4, RORB*
Lung	*TITF1, FOXA2*
Ovary	*EMX2, ESR1, ZBED2*

Third we identified TFs specifically associated with the single tumor types. The table above displays an excerpt of the identified transcription factors.

Conclusions:

Awareness of the role in tumor biology of epigenetics and transcription regulatory mechanisms, such as miRNAs, is increasing. We assume that transcription factors play a crucial role in that context, such that they might be capable of discriminating different tumor classes and yet unknown subgroups. Our approach aims at unsupervised identification of a) sample subsets within a dataset and b) the TFs associated with these subsets. The results show, that an unsupervised selection of TFs can clearly distinguish between the classic breast cancer subtypes and different tumor types such as breast, colon, kidney or prostata cancer. Therefore we assume that the method might help in future to identify new markers suitable for diagnosis of specific treatment related phenotypes or detection of tissues of origin.