technische universität dortmund

Universitätsklinikum Essen

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

SFB 876 Providing Information by Resource-Constrained Data Analysis

Project C1

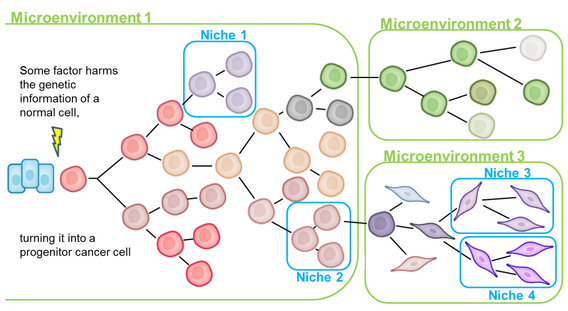
Feature selection in high dimensional data for risk prognosis in oncology

Dr. Sangkyun Lee, Prof. Dr. Alexander Schramm, Prof. Dr. Sven Rahmann

Problem

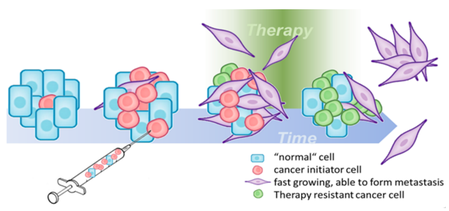
Cancer is an Evolutionary Process

Progression of cancer is driven by both natural selection and therapy, which requires adaptation to different niches and microenvironments.



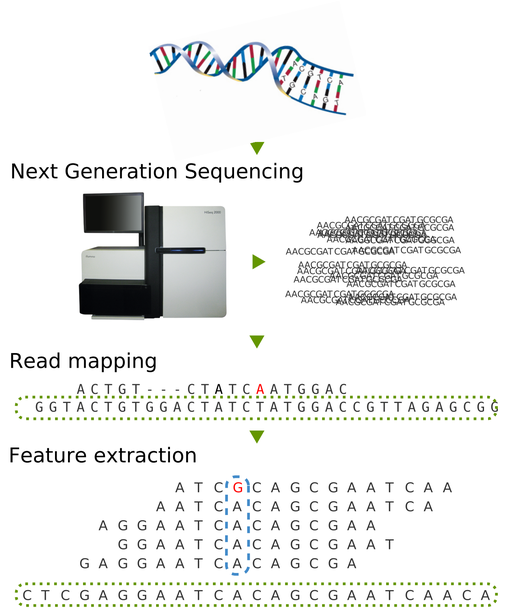
Tumours are Moving Targets

- Understanding inter- and intra-tumour heterogeneity
- Identifying genomic biomarkers for risk prognosis



[Schulte et al. 2018]

Feature Extraction and Selection from High-Throughput Sequence Data



Extremely high feature dimension vs. small number of samples:

- Up to millions of genetic variants
- How to find interpretable tumour-specific variants?

Analysis of Multiple Data Types C4

- Next-generation sequencing, microarrays, arrayCGH
- Gene expression, methylation, copy number variations

Computational Efficiency

- Resource-efficient read mapping and alignment (using GPUs) [Köster and Rahmann 2014]
- Resource-efficient variant filtering

Methodology

Resource-Efficient Feature Generation from DNA

Detailed Analysis of Variant-Tolerant Read Mapping by Min-Hashing

- Min-hashing to estimate the Jaccard similarity

$$J(K_Q, K_R) := |K_Q \cap K_R| / |K_Q \cup K_R|$$

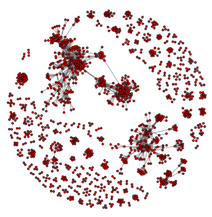
between k-mer sets K_Q, K_R

- Known variants yield additional k-mers which may be added to the reference K_R.
- We showed: beneficial only for variants with high population frequency. [Quedenfeld and Rahmann 2017]

Correspondence between specific SNVs and Methylation Level Changes

- Inter-individual methylation changes may often be traced to a single causative SNV.
- State (unmethylated, semi-methylated, fully methylated) requires discretisation of methylation level in $[0, 1]$.
- Achieved with beta mixture models; computed with a novel EM-type algorithm using moment estimators in the M step. [Schröder and Rahmann 2017]

Identification of Sparse Feature Graphs



A6

- Preliminary: Model similarity between protein complexes
- Focus: Dependencies between expression levels of genes
- Estimation of inverse covariance matrix $\Theta = \Sigma^{-1}$ with partial correlations

Penalized likelihood approach, assuming multivariate normal distribution:

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \; -\log\det\Theta + \mathrm{tr}(S\Theta) + R(\Theta)$$

LASSO:

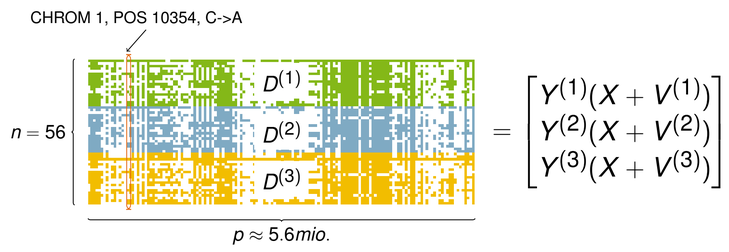
- $R(\Theta) = \lambda \|\Theta\|_1$
- Determine λ by controlling the family-wise error rate (restrictive) or by cross-validation (sample dependent)

SLOPE:

- $R(\Theta) = \sum_{i=1}^{p^2} \lambda_i |\Theta|_{(i)}$
- Determine λ_i by False Discovery Rate (FDR) control
- Efficient saddle point optimisation for the Dantzig selector formalisation

[Lee, Brzyski, and Bogdan 2016]

Tumour Subtype Identification



- Biclustering of patients and variants by a modified Boolean matrix factorization (BMF)

$$\min_{X,V,Y} \sum_{a=1}^{c} \|D^{(a)} - Y^{(a)} \cdot (X + V^{(a)})^T\|^2 + S(X, Y)$$

V indicates tissue-specific variants for each cluster, $S(X, Y)$ regulates specificity of V [Hess and Morik 2017]

- Solve NP-hard BMF by proximal alternating minimization of relaxed objective
- Determine the rank by FDR control

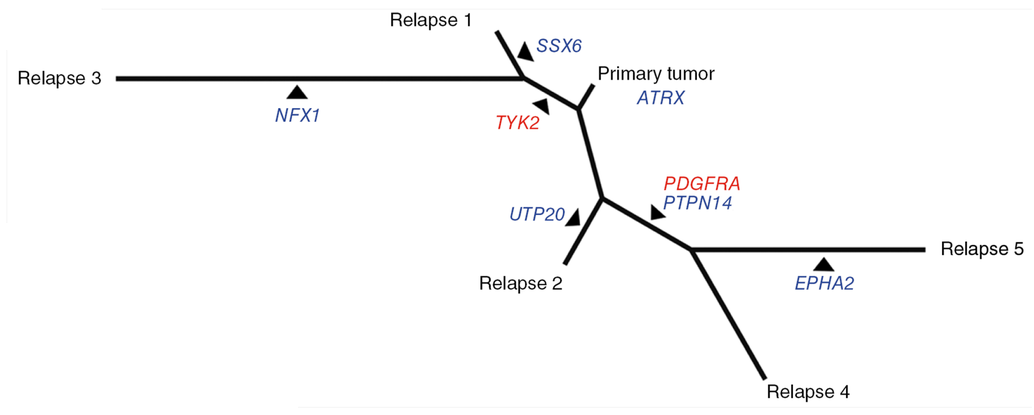
A1

[Hess, Morik, and Piatkowski 2017; Hess, Piatkowski, and Morik 2018]

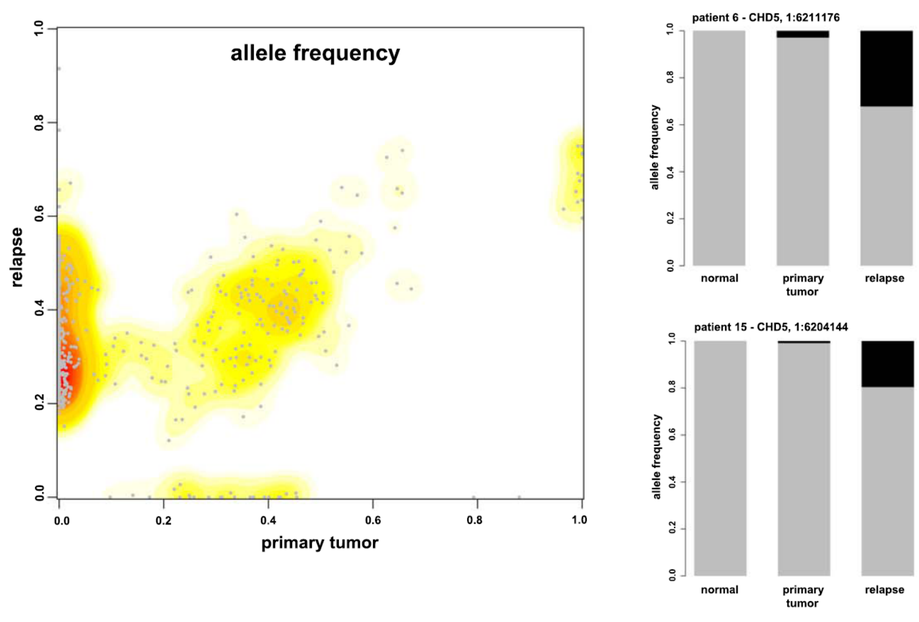
Results

Tumour Evolution

- Mutational dynamics were analyzed in one primary tumour and multiple relapse samples from a single neuroblastoma patient.
- A phylogenetic tree was built according to Hamming distances between the primary tumour and the relapse samples (Neighbor Joining).
- A stable pattern of bifurcation emerged, suggesting that at least two different subclones developed independently from the primary tumour.
- Every branch has 100% bootstrap support and is robust against various perturbations of the input data. [Schramm et al. 2015]



Intratumoural Heterogeneity



- Allele frequency shifts between matched pretreatment primary and relapse neuroblastomas.
- Most SNVs detected in relapse tumours are below the whole-exome sequencing detection limit in the primary tumour. [Schramm et al. 2015]

Artificial Intelligence Group

wtz westdeutsches tumorzentrum

Chair of Genome Informatics
Institute of Human Genetics
University Hospital Essen
University of Duisburg-Essen