Data mining the TopoChip

Hulsman, Marc; Unadkat, HV; Cornelissen, Kamiel; de Boer, Jan; Reinders, Marcel

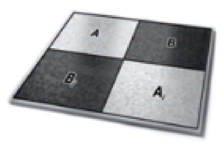
Publication date
2011

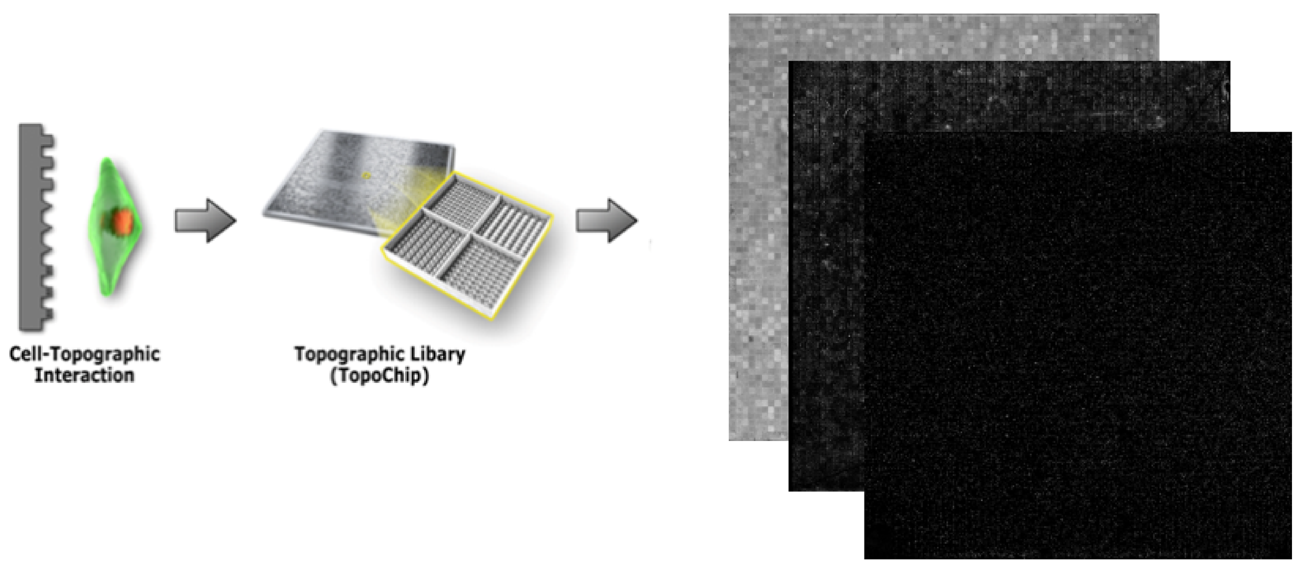
Data mining the TopoChip

Marc Hulsman, Hemant V. Unadkat, Kamiel Cornelissen, Jan de Boer, Marcel J.T. Reinders
Delft Bioinformatics Lab, Faculty of Electical Engineering, Mathematics and Computer Science, Delft University of Technology
MIRA Research Institute, Department of Tissue Regeneration, University of Twente

Overview pipeline

With a TopoChip, the response of cells to material surfaces is measured in high-througput. Each chip contains 2178 different material surfaces, replicated two times accross the chip.

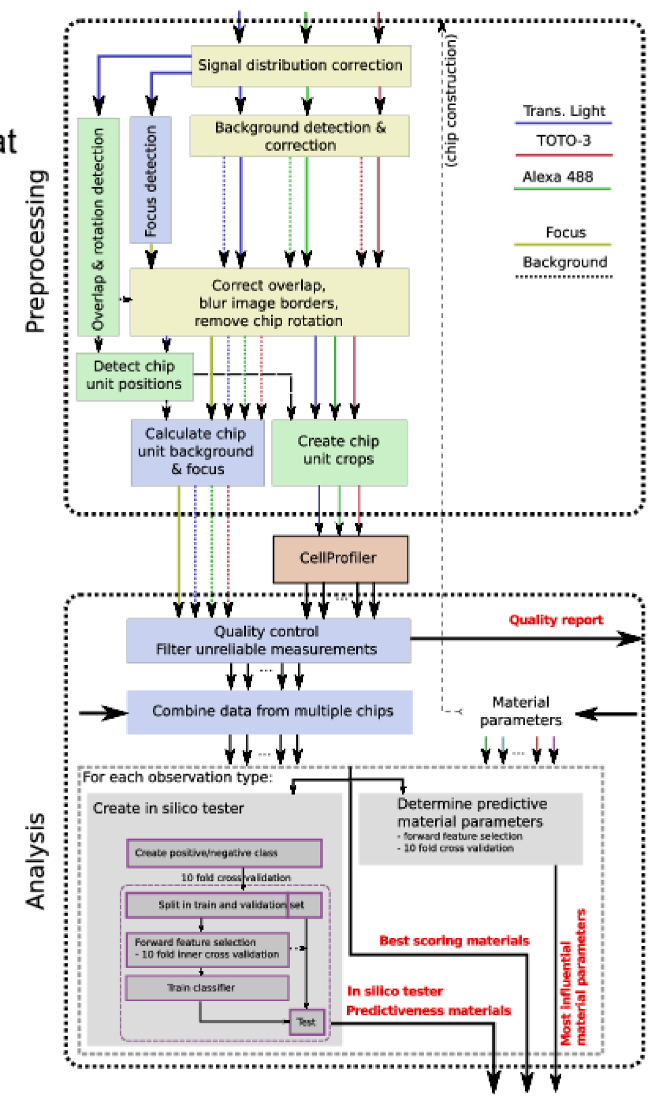


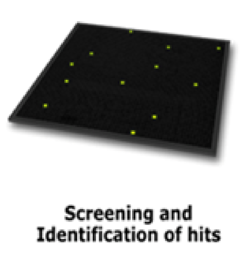


High-througput measurements require also high-througput analysis.

Therefore, a pipeline has been constructed that automatically processes the chips, returning:

- the best material hits
- the most influential material features
- a quality report
- an *in silico* predictor for material surfaces.



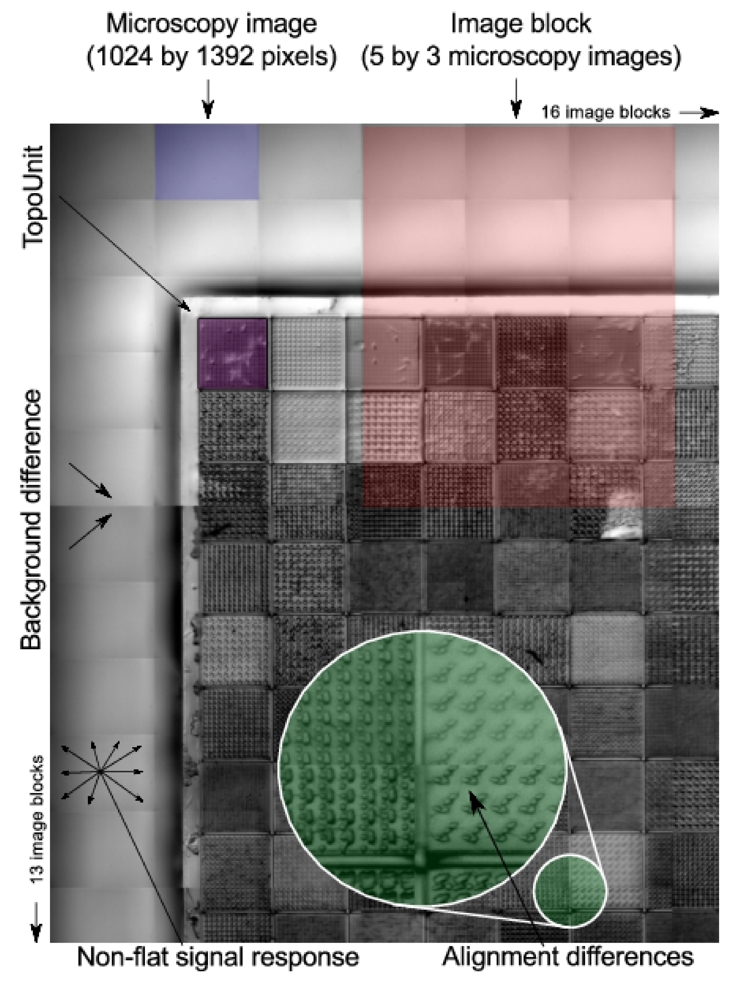


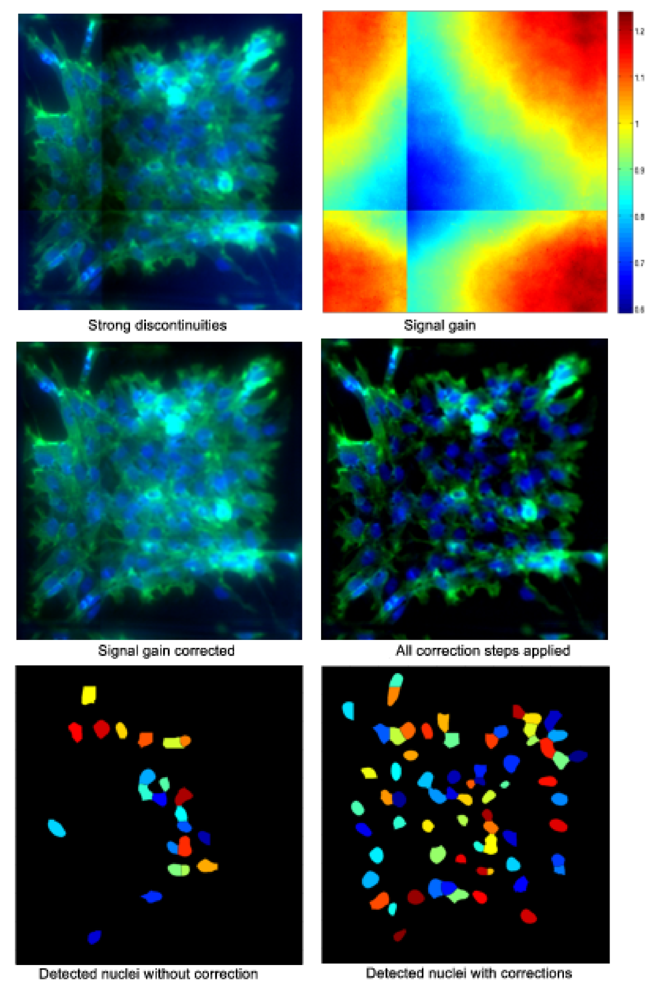
This is done in three separate steps:

- preprocessing: detecting the units, finding/correcting artifacts
- cell profiling (using CellProfiler)
- analysis: finding patterns in the data

Preprocessing

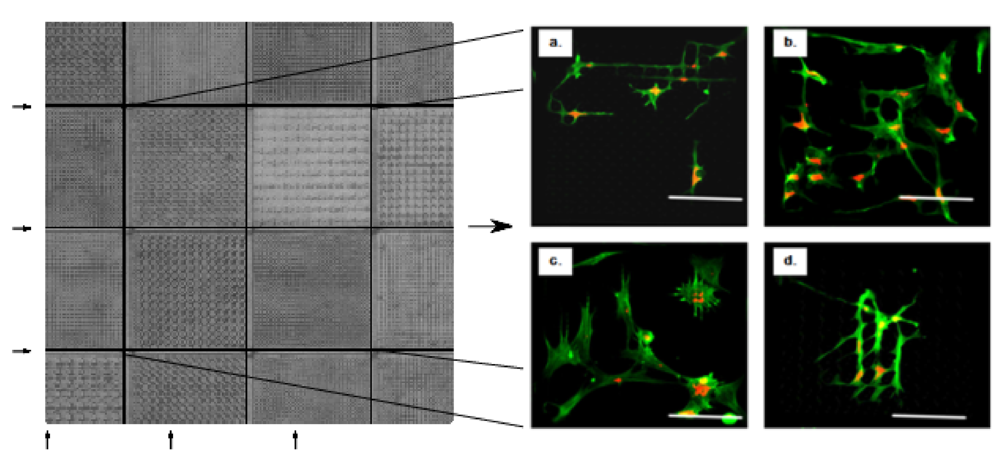
Each chip is imaged in high resolution, resulting in images of approximately 4 gigapixel. In the first step we remove border effects and determine overlap between microscopy images. Also, artifacts/unfocused areas are detected. This information is later used to flag outlier-units.





Unit detection and cell profiling

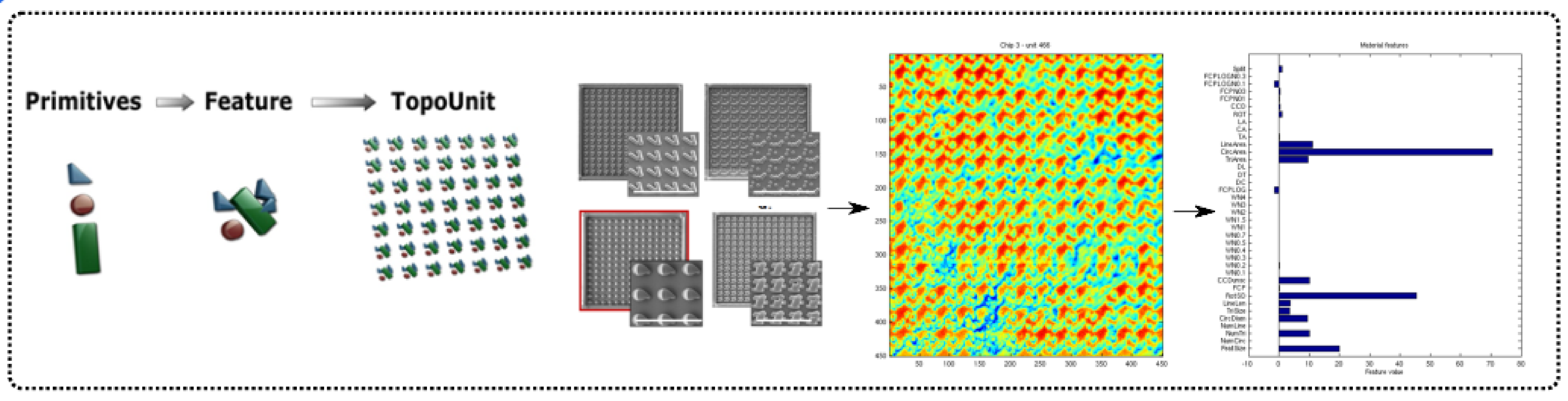
The location of individual units is detected. Possible rotation is removed, after which the figures are cropped, and stored for processing by CellProfiler.



To find the unit locations, an optimization problem is solved, which determines scale, rotation and translation of the units. Specifically, a score is minimized that takes into account border intensity and flatness, as well as the autocorrelation between border locations.

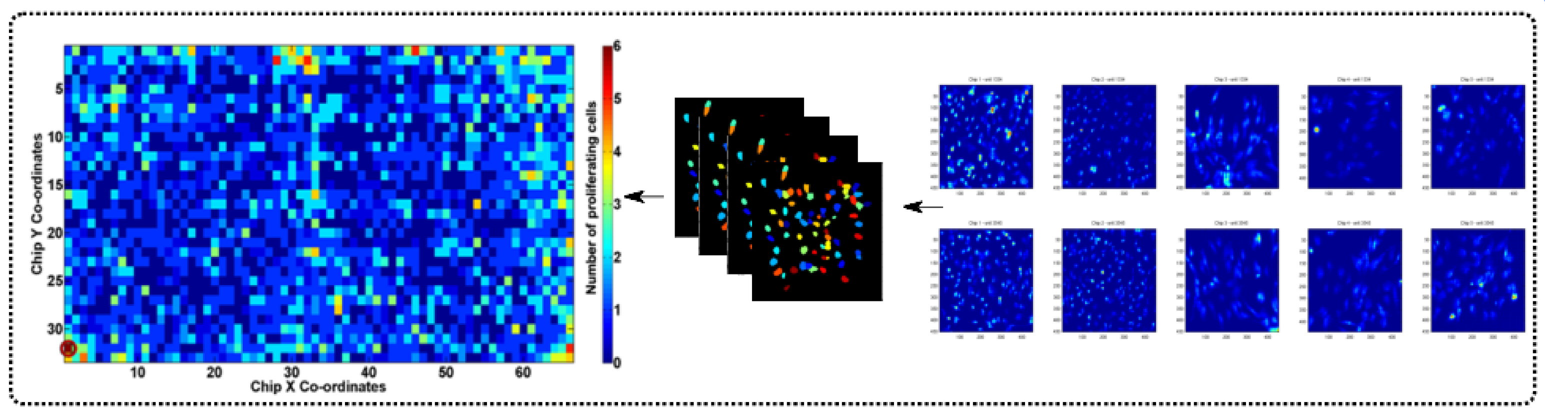
Each cropped unit is subsequently processed by a (customizable) workflow in CellProfiler, resulting in various cellular morpholocial measurements.

Analysis



Material properties

Find relations



Cellular attributes

One of the key points to learn from an TopoChip experiment is how material properties influence cell state.

In order to have a good description of these properties, the TopoChip material library has been built using a deterministic process, guided by a set of parameters. Additionally, extra descriptive parameters have been calculated afterwards (e.g. fourier descriptors).
During analysis, these parameters are correlated to cell attributes.

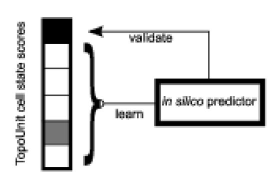
For each cell in a TopoUnit, various attributes are measured using CellProfiler, which together describe the cell state. From these attributes, a target label is determined, describing the cell state of interest.

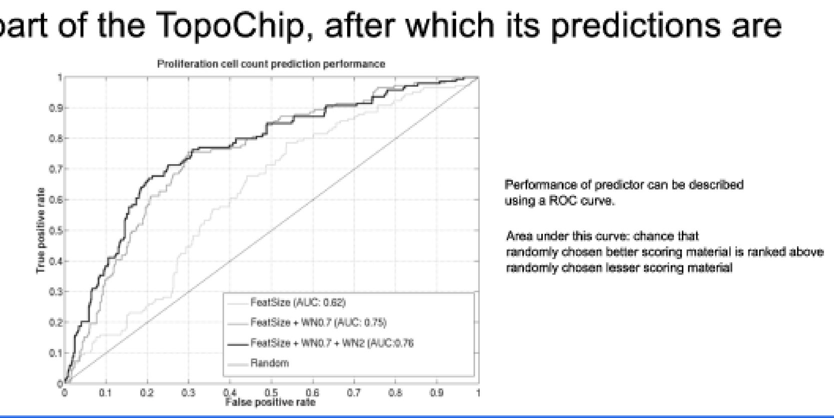
Relations between material properties and cell state are learned using a machine learning approach. The strength of these relations is determined by scoring the resulting *in silico* predictor in its capacity to predict cell state *using only material parameters*.

Influence of material properties

Designing improved materials requires knowledge about which material properties are important. The role of individual properties is determined using a feature selection algorithm. It returns a list of properties, ordered on maximum contribution to the performance of the predictor.

Predictor performance is determined by training it on part of the TopoChip, after which its predictions are validated using the measurements on rest of the chip.

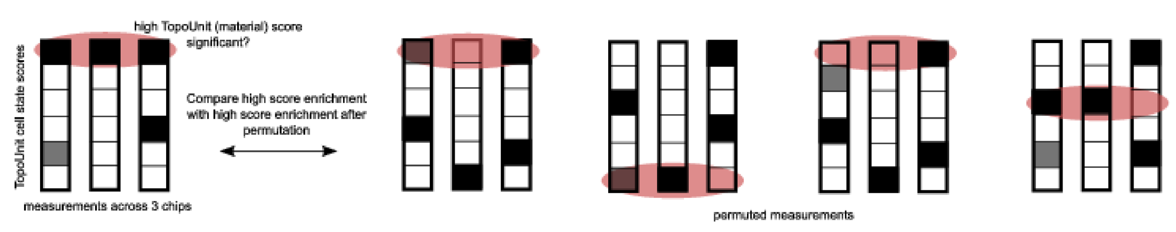




High scoring materials

The best materials can be obtained by ranking them w.r.t. to the scores of the cells that grow on them. However, to exclude the effect of random variations and artifacts, one can perform significance testing by comparing measurements accross multiple chips.

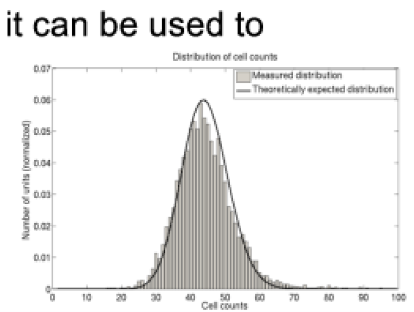
This is done using a permutation test, which estimates the chance that an observed high score will occur by chance.



Implementation

Image data for each chip measures 30-40 GB. Processing multiple chips takes considerable computational resources. For this reason, the pipeline has been made suitable for use on a computer cluster.

For validation purposes, a quality report is generated. Among other things, it can be used to check if unit locations were accurately found and artifacts were correctly detected and/or removed. Also, the cell count distribution is compared to theoretically expected distribution.



Discussion

Learning a model that relates material properties to cell state is an important step in interpreting a high-througput material testing experiment, enabling one to determine how (and if) cell state differences are induced by the material properties.

We found that relations between material properties and cell state are often non-linear. For this reason, we employ a nearest-neighbour predictor, which is capable of handling such relations.

In the future, predictors might be used to let the pipeline suggest new materials which do not occur on the TopoChip itself, but are predicted to have a high performance.

TU Delft Delft University of Technology