

Eine molekulare Definition des Burkitt-Lymphoms aus Sicht der Bioinformatik

Rainer Spang

12.09.06 GMDS Leipzig 2006

Burkitt lymphoma

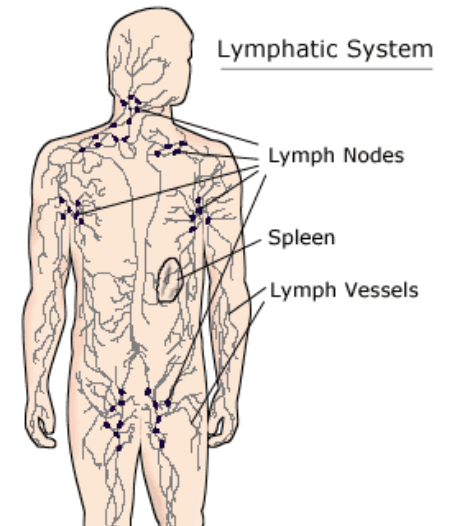
Difficult to diagnose ... DLBCL is very similar

Needs to be treated differently

Definition is still controversial

Goal:

A biological definition of Burkitt lymphoma using gene expression data (Microarrays)



The purely supervised approach

Set class labels BL/non BL without using the expression data

Use:

- panel morphology
- imunohistochemistry
- cytogenetics

Expert Class Label

→ apply statistical learning theory

Establishing a signature



**Split Data into
Training and
Test Data**

**Test data only:
Internal validation
Full quantitative
specification**

**External
Validations**

**Training data only:
Machine Learning**

- select genes
- find the optimal number of genes
- learn model parameters

What is this?

Expert class label: non-Burkitt

Gene expression signature: Burkitt

A: An error of the classification model

B: A hidden Burkitt lymphoma

The experts themselves: It could be B!

The statisticians: Then we were never facing
a classification problem

Reliable Disease Labels

Selecting informative genes needs ... reliable disease labels

Finding the optimal

number of genes needs

... reliable disease labels

Learning model parameters needs ... reliable disease labels

Internal evaluation needs

... reliable disease labels

External evaluation needs

... reliable disease labels

For Burkitt lymphomas

we did not have

... reliable disease labels

Molecular pathology & class finding

Idea: Use the expression profiles to form molecularly homogenous groups of patients. These groups are candidates for novel definitions of disease entities.

Goal: The expression based stratification of patients can be the basis of new clinical studies.

Do patients which display a certain expression signature respond differently to a certain drug or not?

This sounds like a clustering problem, but ...

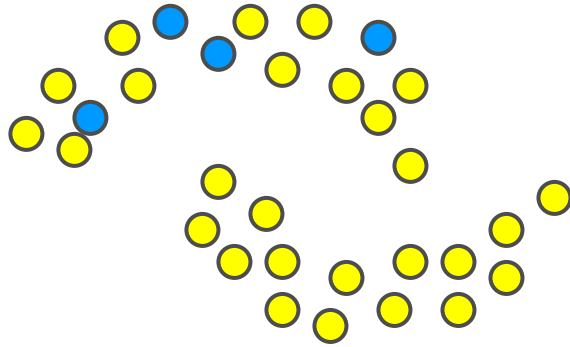
... you do not want any clustering of patients. You still want to gear the patient stratification towards the characterization of Burkitt lymphomas

→ semi supervised learning

... different sets of genes make different patients look Burkitt like

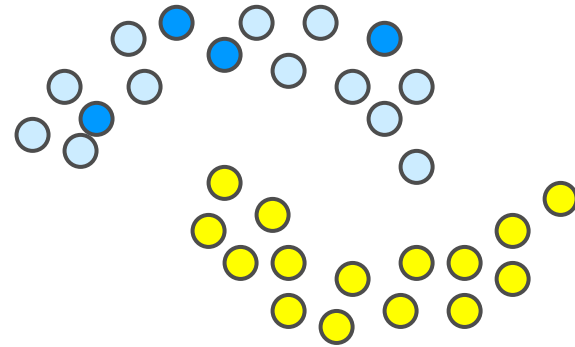
→ variable selection with unclear criteria

Semi supervised learning with known features



“Expert” diagnosis

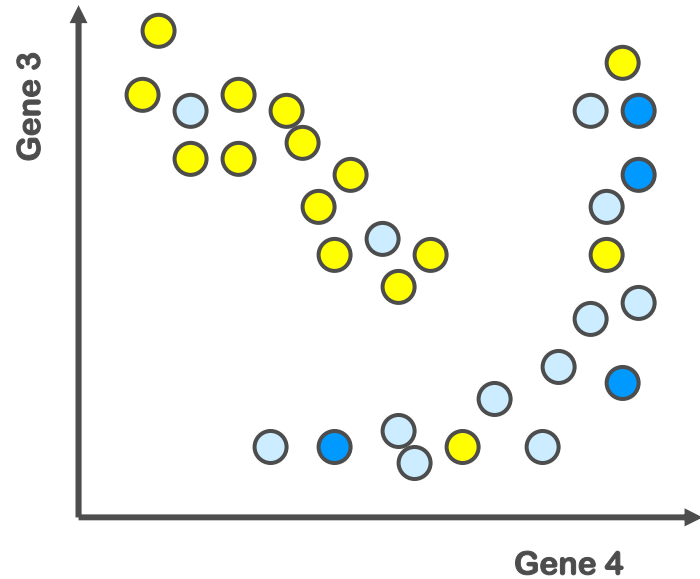
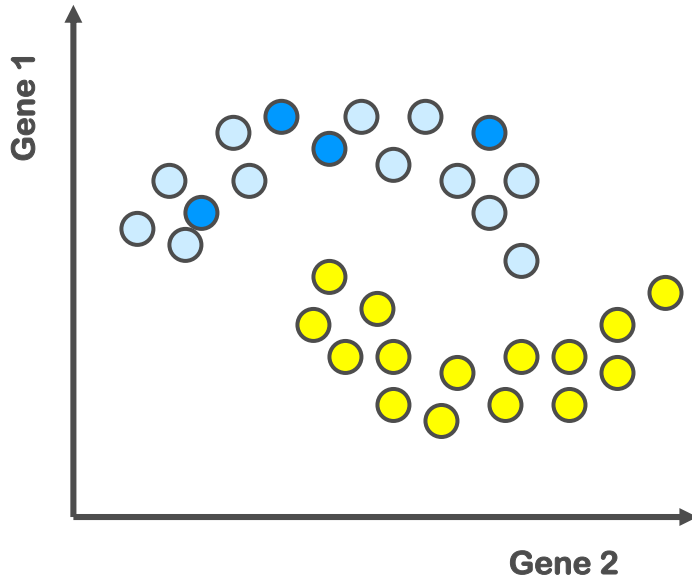
- Non-BL
- BL



Learned diagnosis

- Non-BL
- BL

Gene selection



Different genes lead to different disease definitions

→ which one is the better one?

Stability Analysis

Core group extension needs a validation

It does not make any misclassifications by definition

It does not predict a disease, it defines it

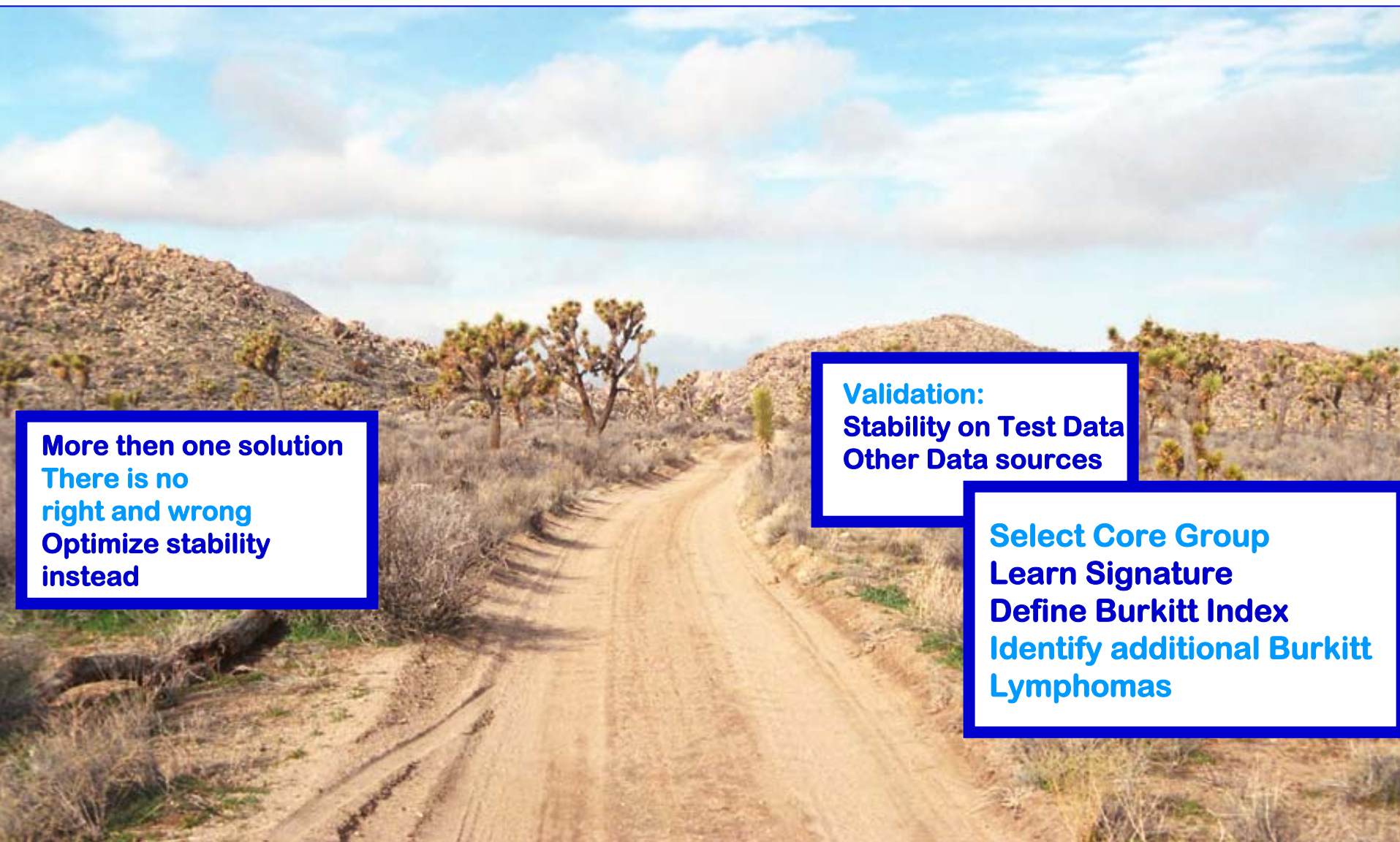
However, is this a good definition?

Ultimate answer: clinical study

For now: **stability of diagnosis**

use bootstrap samples from the core group, relearn a signature, diagnose all patients, look whether the diagnosis stays the same

Core Group Extension

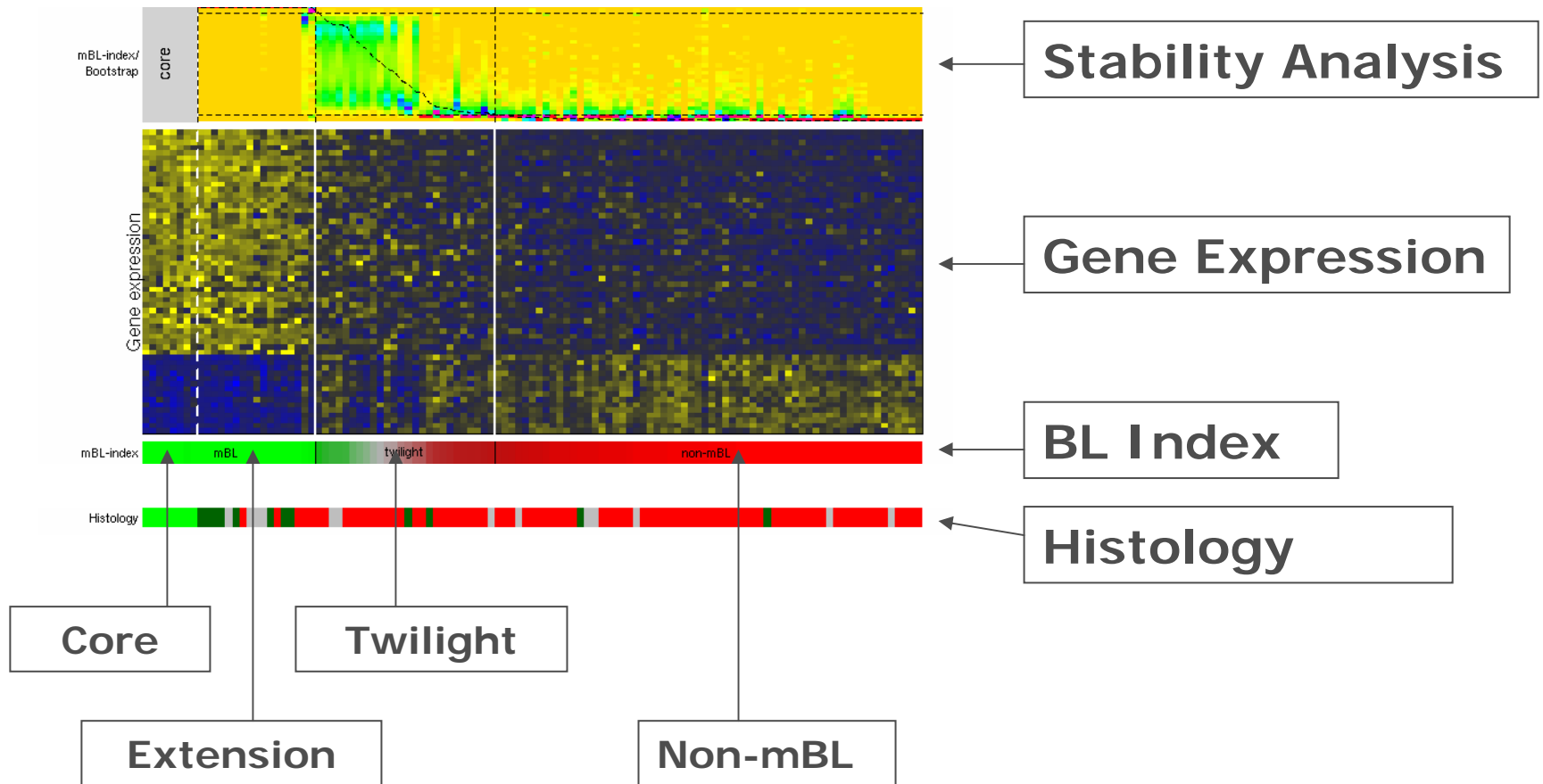


More than one solution
There is no
right and wrong
Optimize stability
instead

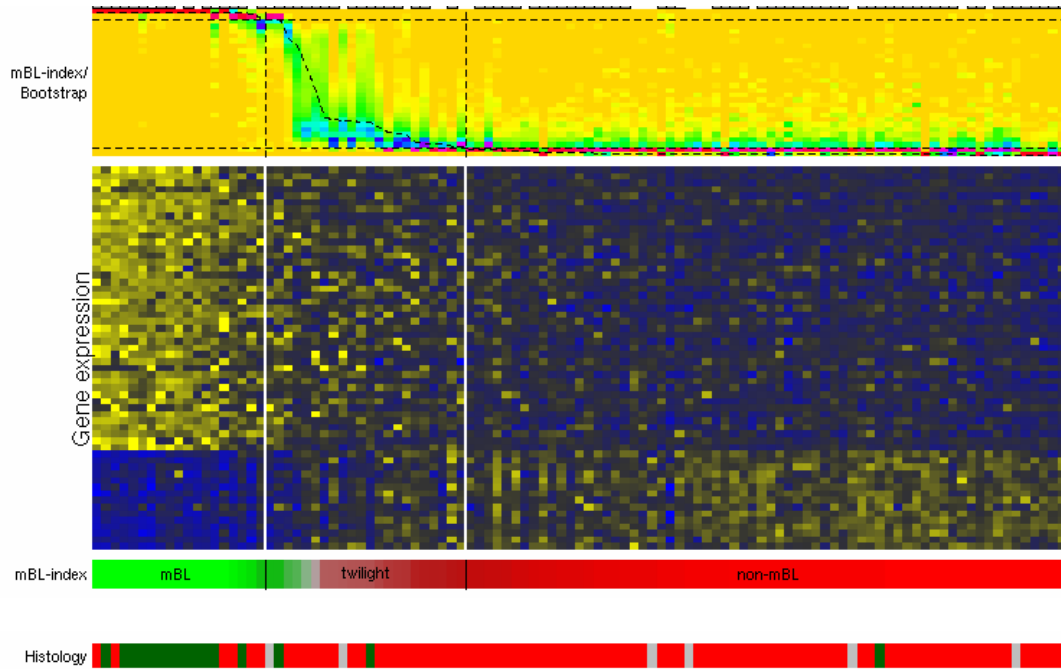
Validation:
Stability on Test Data
Other Data sources

Select Core Group
Learn Signature
Define Burkitt Index
Identify additional Burkitt
Lymphomas

The Definition of the molecular Burkitt Lymphoma (mBL)



Independent Test Set



Thank You