



# Identification of bacterial strains using taxonomic microarrays

Hubert Rehrauer<sup>1</sup>, Susan Schönmann<sup>2</sup>, Ralph Schlapbach<sup>1</sup>, Leo Eberl<sup>2</sup>

<sup>1</sup> Functional Genomics Center Zurich, University/ETH Zurich. <sup>2</sup>Institute of Plant Biology, Department of Microbiology, University Zurich.  
E-mail: Hubert.Rehrauer@fgcz.ethz.ch

## Introduction

Microarray hybridizations are ideal tools to monitor microbial populations and identify individual strains, species or higher taxa. They fulfill the most frequently required constraints:

- fast
- high-throughput
- reliable
- cheap

Example applications for hybridizations using diagnostic microarrays are

- biodiversity assessments
- identification of potential pathogenic microbes

In our approach we start out from a microarray with a given set of discriminating probes. In a training phase, we experimentally validate the performance of the individual probes, exclude malfunctioning probes and define strain equivalence classes (i.e. groups of strains such that the members cannot be discriminated, but that different classes can well be discriminated). In the identification phase we compute the p-value for each strain class of being present in the sample (detection p-value).

## Strain Class Definition

The database of known 16S rRNA full length sequences contains more ~50k entries. The probes on the chip cover ~5k sequences including representative of all relevant *Burkholderia* species. However the set of probes cannot discriminate between all of the ~5k sequences, therefore we define sequence or strain classes  $c_i$  that group the sequences that cannot be discriminated. Each strain class  $c_i$  is uniquely characterized by a set of matching probes  $\Omega_i$  such that all probes within  $\Omega_i$  are matching probes to all 16S rRNA sequences in  $c_i$ .

## Probe Characterization

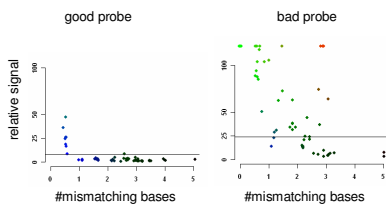
Despite the careful *in silico* design, probes may consistently malfunction and show false positive or false negative signals. Using hybridizations of a set of 39 different typestrains with known 16S rRNA sequence, we assessed the performance of each probe.

For each probe  $p_j$  a detection threshold  $d_j$  was defined as

$$d_j = \text{mean}(\text{background signal}_j) + 3 \text{sd}(\text{background signal}_j)$$

Probes must be present (signal above  $d_j$ ) when hybridizing a matching strain and absent when hybridizing a non-matching strain (>1 non-matching base). Non-matching bases are weighted according to the position of the probe.

Out of the 160 probes on the chip 8 were identified as bad and excluded, because their false positive rate was higher than 10%.



## Burkholderia and Pandoraea Phylochip

We demonstrate the methodology with an example application to a taxonomic array that discriminates bacterial strains based on the 16S rRNA sequence. It targets *Burkholderia* and *Pandoraea* that are in nature widely distributed  $\beta$ -Proteobacteria. They can act beneficial or pathogenic in the environment and in clinic e.g. as biocontrol agents or plant growth promoters, but also as human pathogens, especially in patients suffering from cystic fibrosis (CF).

The phylochip is composed of 160 probes covering the entire known diversity of the genera *Burkholderia* and *Pandoraea* and target 16S rRNA sequences at various taxonomic levels, ranging from higher taxa (*Bacteria*, *Archaea*, *Eucarya*, *Planctomycetes*, *Verrucomicrobia*,  $\beta$ -proteobacteria, *Burkholderia*) to species (*Burkholderia* sp.). It has been designed using the ARB software in combination with a database holding >50k 16S rRNA sequences. 18mer Probes were chosen for their predicted specificity and hybridization properties. Probes were spotted onto epoxy slides using a Perkin Elmer Piezorray spotter.

## Strain Class Identification

For each microarray hybridization, we compute the detection p-value for each strain class  $c_i$  by looking at the present status of all probes within the corresponding set of matching probes  $\Omega_i$ .

We define the prior probability  $\pi_j$  of probe  $p_j$  being present as

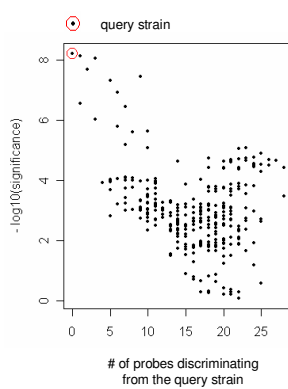
$$\pi_j = \frac{\# \text{ matching strain classes of probe } p_j}{\# \text{ all strain classes}}$$

and compute the detection p-value of class  $c_i$  as

$$\text{Prod}_j \begin{cases} \pi_j & \text{if } p_j \text{ is present} \\ 1 - \pi_j & \text{if } p_j \text{ is absent} \end{cases}$$

For the 39 typestrains with known sequence, we can assess the performance of the strain identification by comparing the detection p-value of the hybridized strain to the detection p-value of all strains.

Detection p-values of all strain classes



## Summary

We have applied our strain identification approach to the *Burkholderia* Phylochip. After experimental validation with 39 typestrains, we have identified 8 malfunctioning probes and ignored those in the subsequent identification steps. For all of the typestrains the corresponding strain class was identified as present with a significance threshold of 1e-3 and for 87% of the typestrains, the true strain class was also the one with the smallest detection p-value. In a mixture experiment of two typestrains, each strain was called detected as soon as its concentration was 5% or higher in the mixture. When hybridizing unknown strains, the tree visualization of the strain classes points to the strain family of the unknown strain.

Using microarray hybridizations we can successfully discriminate even between closely related strains and identify correct strain classes.

## Example: Identification and Visualization

Detection p-values for strain classes when hybridizing an unknown strain extracted from a forest soil sample. The 330 strain classes that can be discriminated by the chip are organized in a tree that reflects the similarities of the classes as seen by the chip.

