AN OPTIMIZED OLIGONUCLEOTIDE ARRAY DESIGN FOR CHIP-ON-CHIP

Fiona Nielsen[1], Stefan Gräf[2], Xinmin Zhang[3], Stefan Kurtz[4], Sergei Denissov[1], Roland Green[3], Ewan Birney[2], Paul Flicek[2], Martijn Huynen[1], Henk Stunnenberg[1]

[1] Nijmegen Centre for Molecular Life Sciences, the Netherlands
[2] EMBL-European Bioinformatics Institute, Hinxton, Cambridge
[3] NimbleGen Systems Inc., Madison, USA
[4] Center for Bioinformatics, University of Hamburg, Germany

The sequencing of whole genomes has allowed for custom-made genome-wide microarray assays such as the ChIP-on-chip. With this technology we can detect e.g. transcription factor binding sites over an entire genome. In principle, an accurate detection is only limited by the resolution of the chipdesign, i.e. the tiling density of the oligonucleotides. However, the inherent noise of the DNA hybridisation severely hampers the interpretation of the results.

We mined existing ChIP-on-chip datasets to identify the main sources of noise arising from the sequence selection. We found that limiting intervals must be imposed on 1)the melting temperature, 2)the lengths of the probes, 3)palindromic sequences and 4)the sequence uniqueness relative to the rest of the genome. Based on this knowledge we have developed an oligonucleotide array design algorithm to generate a signal-to-noise-optimised genome-wide array design for any given genome at a given tiling density. To obtain unique sequences we invented a novel approach for selecting the sequences. Using an augmented suffix-array implementation we score sequences by their content of sequence-unique subsequences and select preferentially the sequences with the highest content of unique subsequences.

We have tested our design using different parameter settings in a fractional factorial test setup, in effect testing eight different parameter combinations. The tests were designed for the mouse genome on the 2.18M feature array from Nimblegen and performed under true ChIP-on-chip experimental conditions using mouse TBP ChIP samples for the hybridisations. Test hybridisations were performed for three biological replicas, each hybridised three times, to estimate the variance across both biological and technical replicas.

From the test designs we deduced the effect of each parameter on the resulting signal and coverage of the design. We correlated the effects and interactions of the probe properties on the probe level (signal intensities) as well as on the design level (quality measures for the whole data set). From this analysis we quantify the effect of each parameter, thus allowing us to choose the design parameter settings that optimise the signal-to-noise ratio, while maintaining a high coverage of the genome. Using our design algorithm and the optimised parameter settings we can produce a genome-wide microarray design with low noise and high coverage for any sequenced genome. The first optimised design to be released will be based on the MM8 genome assembly of the mouse genome with high genome-wide coverage at ChIP-on-chip resolution.