A statistical method to detect expressed segments in Tiling Array data

Christian Otto

Bioinformatics, Leipzig

February 2010

Table of content

Why tiling arrays?

- unbiased view of transcription

Why tiling arrays?

- unbiased view of transcription
- no mRNA enrichment necessary

Why tiling arrays?

- unbiased view of transcription
- no mRNA enrichment necessary
- detection of rare transcripts and different splice variants

Why tiling arrays?

- unbiased view of transcription
- no mRNA enrichment necessary
- detection of rare transcripts and different splice variants
- can capture expression on large genomic regions

## Design of genomic tiling arrays
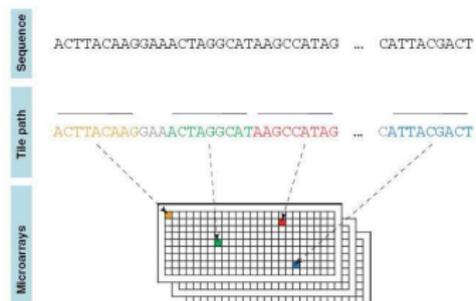
General design:

## Design of genomic tiling arrays

General design:
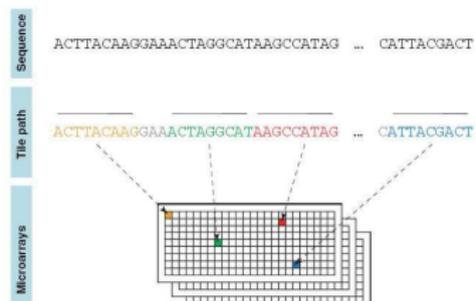
Tile path selections:
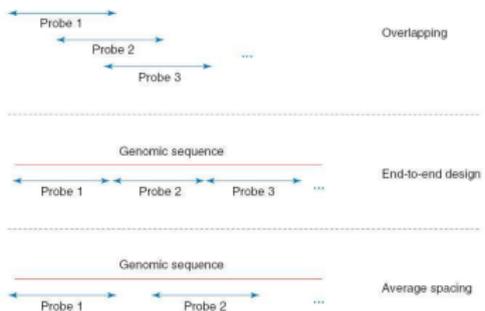
# Transcriptome mapping with tiling arrays
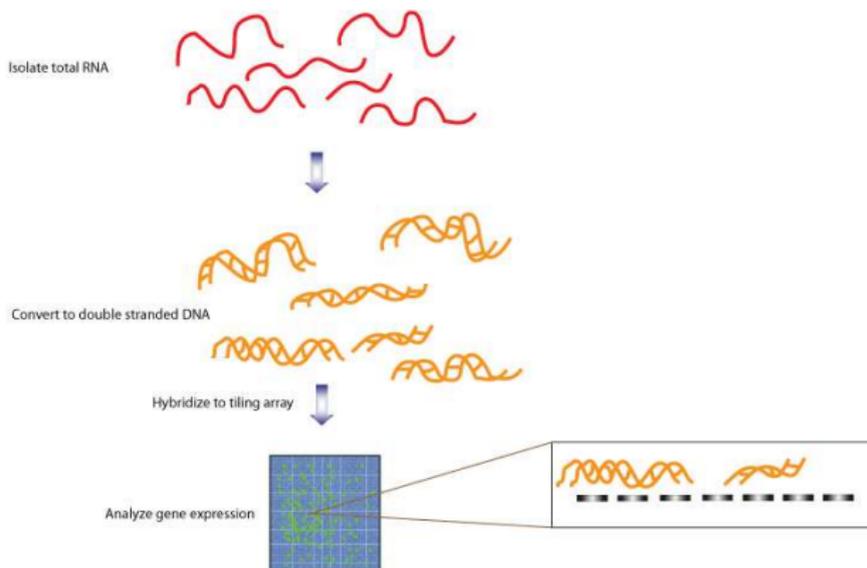
Cross-hybridization effects in tiling array data

= hybridization to sequence that is similar or identical to the target

Cross-hybridization effects in tiling array data

= hybridization to sequence that is similar or identical to the target

- use of mismatch probes in Affy Tiling Arrays
- exclude repeat regions annotated by Repeat Masker

Cross-hybridization effects in tiling array data

- $=$ hybridization to sequence that is similar or identical to the target
- use of mismatch probes in Affy Tiling Arrays
- exclude repeat regions annotated by Repeat Masker

$\Rightarrow$ rough intensity signal
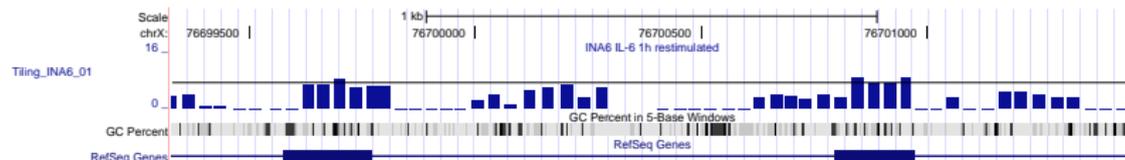
Cross-hybridization effects in tiling array data

$=$ hybridization to sequence that is similar or identical to the target

- use of mismatch probes in Affy Tiling Arrays
- exclude repeat regions annotated by Repeat Masker

$\Rightarrow$ rough intensity signal
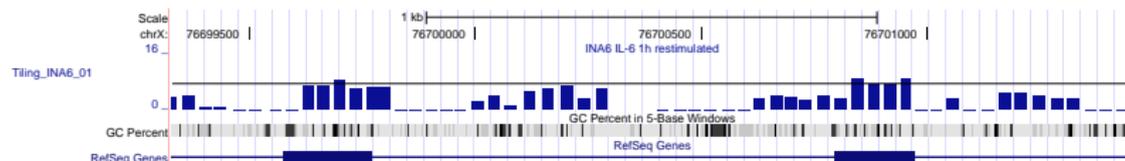

$\Rightarrow$ task of data segmentation

Previous work

- simple intensity thresholding
- proximity-based heuristics (e.g. Kampa et al. 2004)
- dynamic programming
- hidden Markov models

Analysis of Kampa et al. (2004)
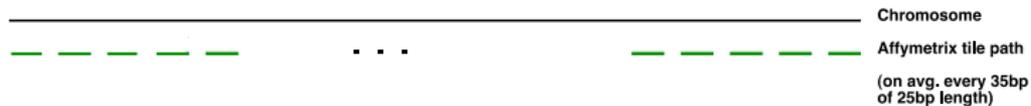
Chromosome

Affymetrix tile path

(on avg. every 35bp
of 25bp length)

Analysis of Kampa et al. (2004)

Analysis of Kampa et al. (2004)

Analysis of Kampa et al. (2004)

Chromosome

Affymetrix tile path

(on avg. every 35bp
of 25bp length)

P−BW   P   P+BW

Pseudo−median

Pseudo−median > INTENSITY
= positive probe

< MAXGAP

> MINRUN

Transfrags
(contiguously transcribed
elemenets)

2.9% false positive rate: BW = 35bp   MAXGAP = 40bp

INTENSITY = 150   MINRUN = 90bp
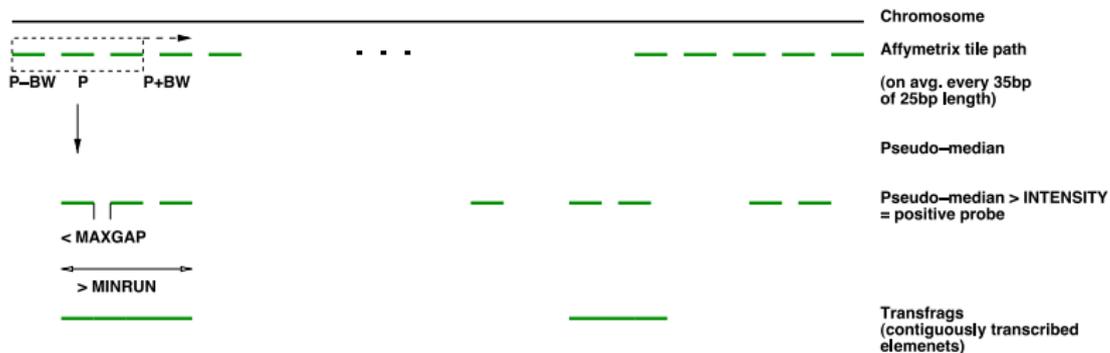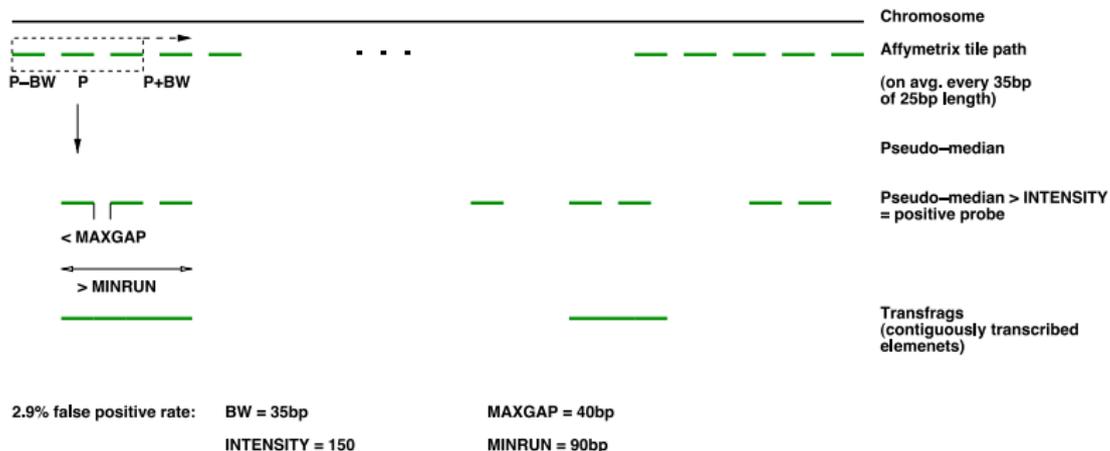
Statistical segmentation method
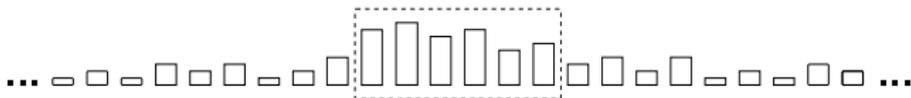
Signal intensity and GC content
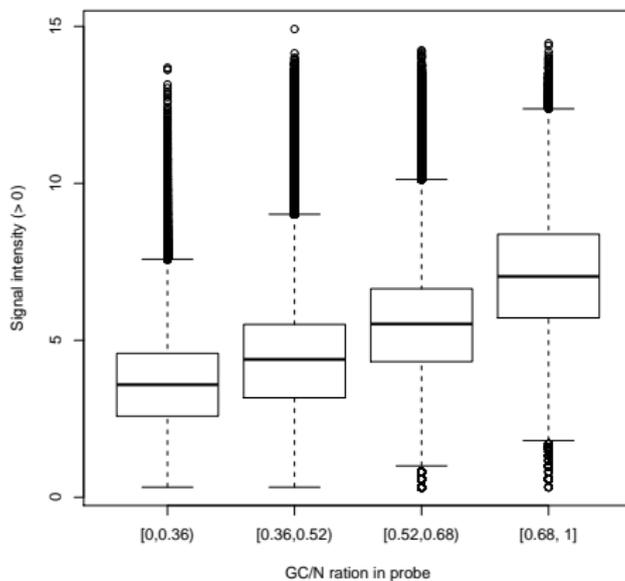
Statistical segmentation method
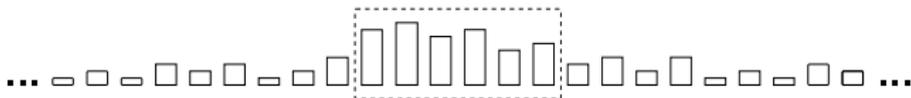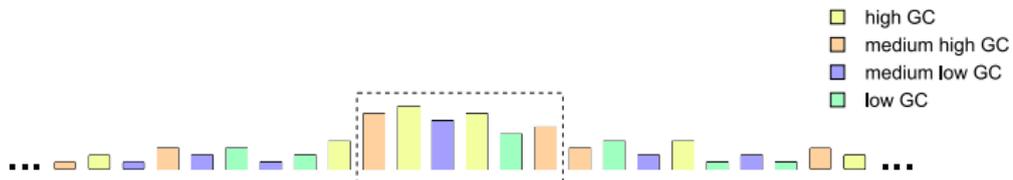
# Statistical segmentation method

Statistical segmentation method

□ high GC
□ medium high GC
□ medium low GC
□ low GC

GC-based shuffling
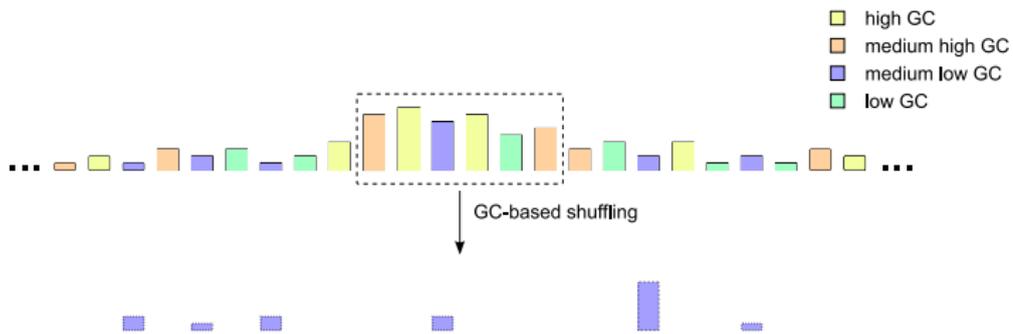
Statistical segmentation method
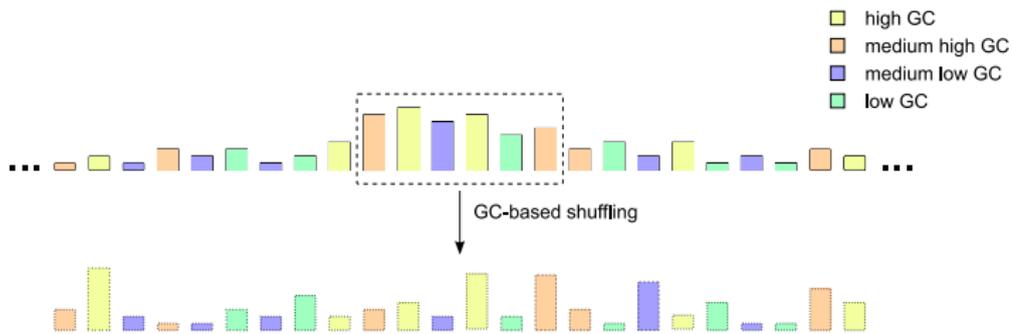
Statistical segmentation method
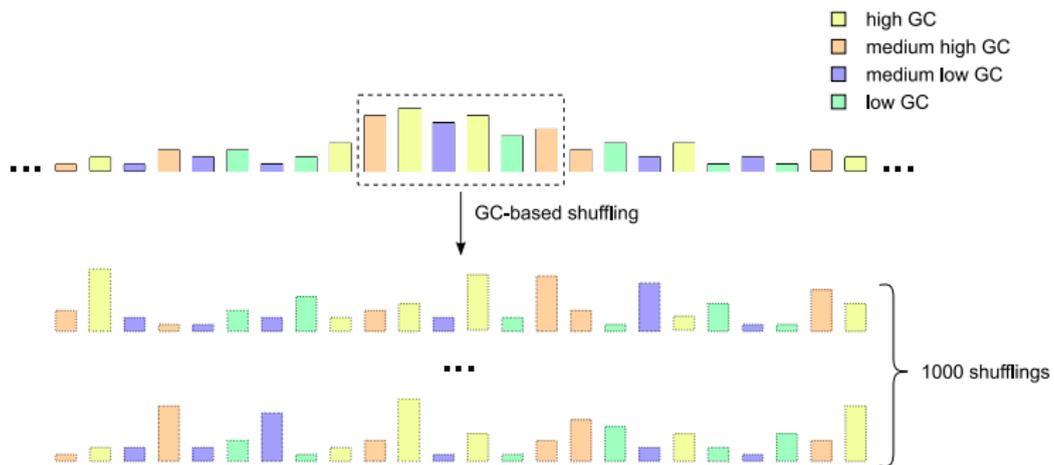
# Statistical segmentation method
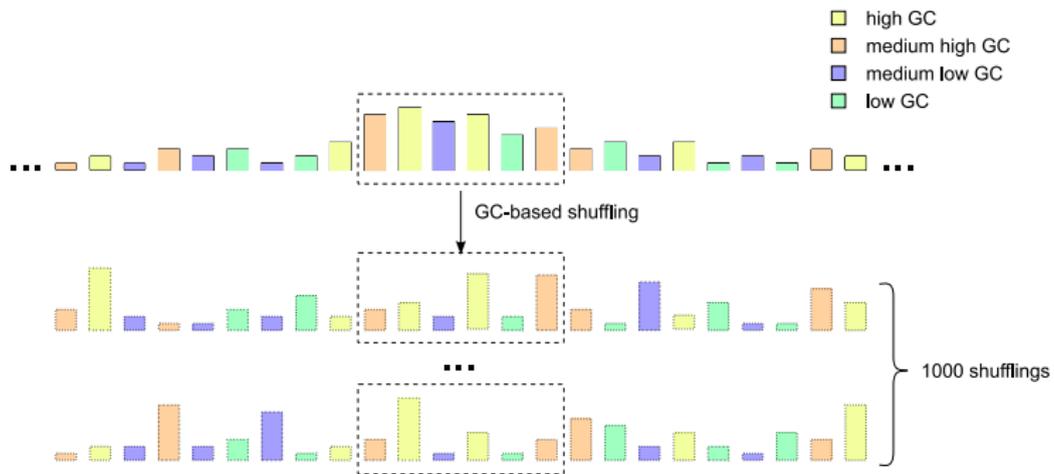
# Influence of GC classification
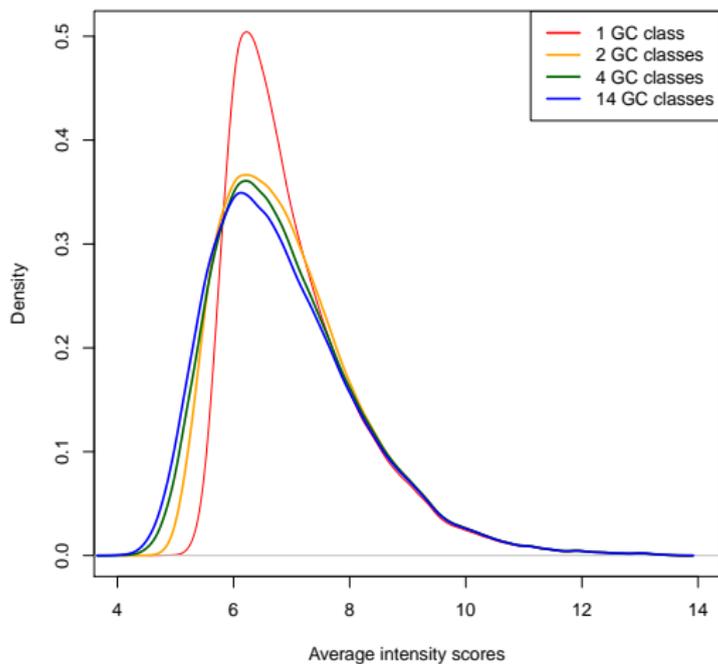
Comparison of segment lengths
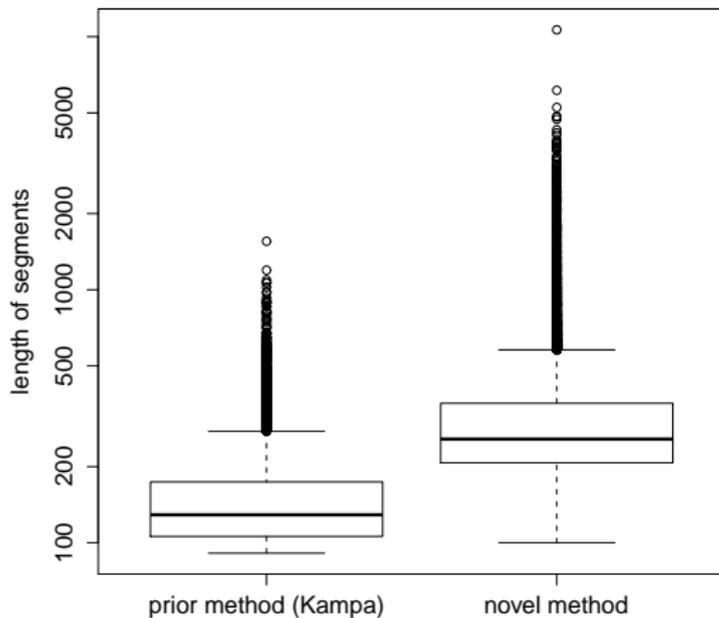
Example of exon structure
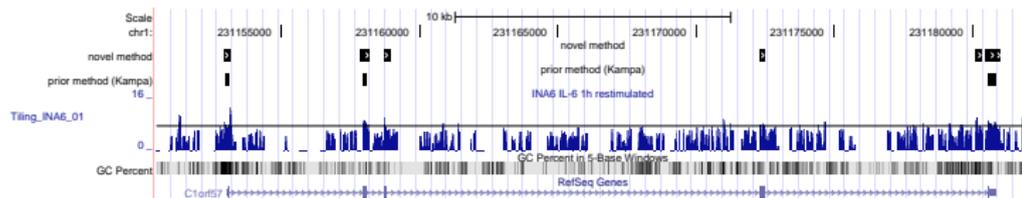
Example of exon structure
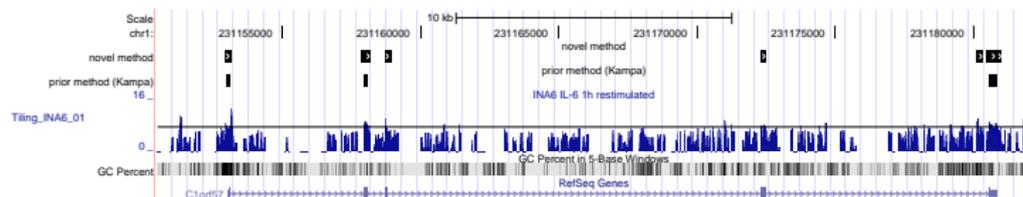
Preliminary conclusions:

- fast statistical method to detect expressed segments
- consideration of GC bias
- detection of longer segments

Future work

- evaluation of reported segments
- use of median or pseudomedian over windows
- consider other sequence-specific biases
- support of other tiling array platforms

Bibliography

📄 J M Johnson, S Edwards, D Shoemaker, and E E Schadt.
Dark matter in the genome: evidence of widespread transcription detected by
microarray tiling experiments.
Trends Genet, 21(2):93–102, Feb 2005.

📄 T E Royce, J S Rozowsky, P Bertone, M Samanta, V Stolc, S Weissman,
M Snyder, and M Gerstein.
Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping.
Trends Genet, 21(8):466–75, Aug 2005.

📄 D Kampa, J Cheng, P Kapranov, M Yamanaka, S Brubaker, S Cawley,
J Drenkow, A Piccolboni, S Bekiranov, G Helt, H Tammana, and T R Gingeras.
Novel RNAs identified from an in-depth analysis of the transcriptome of human
chromosomes 21 and 22.
Genome Res, 14(3):331–42, Mar 2004.

📄 M Guttman, I Amit, M Garber, C French, M F Lin, D Feldser, M Huarte, O Zuk,
B W Carey, J P Cassady, M N Cabili, R Jaenisch, T S Mikkelsen, T Jacks,
N Hacohen, B E Bernstein, M Kellis, A Regev, J L Rinn, and E S Lander.
Chromatin signature reveals over a thousand highly conserved large non-coding
RNAs in mammals.
Nature, 458(7235):223–7, Mar 2009.

The end

Thank you for listening!

Feel free to ask some questions.