



Universiteit  
Leiden  
The Netherlands

## **Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping**

Rippe, R.C.A.

### **Citation**

Rippe, R. C. A. (2012, November 13). *Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping*. Retrieved from <https://hdl.handle.net/1887/20118>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/20118>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/20118> holds various files of this Leiden University dissertation.

**Author:** Rippe, Ralph Christian Alexander

**Title:** Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping

**Issue Date:** 2012-11-13

---

# ADVANCED STATISTICAL TOOLS FOR SNP ARRAYS

*Signal Calibration, Copy Number Estimation*

*and Single Array Genotyping*

RALPH C.A. RIPPE

---

LEIDEN UNIVERSITY  
LEIDEN 2012

This book was formatted using the L<sup>A</sup>T<sub>E</sub>X class memoir,  
with custom chapter style definitions.

Printed by Mostert en Van Onderen, Leiden.

Advanced Statistical Tools for SNP Arrays  
Signal Calibration, Copy Number Estimation and Single Array Genotyping  
PhD Thesis, Leiden University

ISBN/EAN: 978-94-90858-14-8

Copyright ©2012, Ralph C.A. Rippe

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronically, mechanically, by photocopy, by recording, or otherwise, without prior written permission from the author.

# Proefschrift

ter verkrijging van  
de graad van Doctor aan de Universiteit Leiden  
op gezag van Rector Magnificus prof. dr. mr. P.F. van der Heijden,  
volgens besluit van het College voor Promoties  
te verdedigen op dinsdag 13 november 2012  
klokke 16:15 uur  
door

Ralph C.A. Rippe

Geboren te Delft  
in 1982

## **Promotiecommissie**

### *Promotores*

Prof. Dr. J.J. Meulman

Prof. Dr. ing. P.H.C. Eilers (Erasmus Medisch Centrum)

### *Overige leden*

Prof. Dr. D.I. Boomsma (Vrije Universiteit Amsterdam)

Prof. Dr. F.A. van Eeuwijk (Wageningen Universiteit)

Prof. Dr. W.J. Heiser (Universiteit Leiden)

Prof. Dr. P. Slagboom (Leids Universitair Medisch Centrum)

# CONTENTS

---

<b>List of Figures</b>	<b><i>x</i></b>
<b>List of Tables</b>	<b><i>xi</i></b>
<b>1 Introduction</b>	<b>1</b>
1.1 Human DNA and disease	1
1.2 SNP chips and recent developments in DNA measurement	5
1.3 Applications of SNP fluorescence signals	7
1.4 Data format and signal properties	9
1.5 Contributions in this thesis	10
1.6 Thesis outline	13
 <b><i>Research Articles</i></b>	 <b><i>14</i></b>

---

<b>2 Correction of Fluorescence Bias on Affymetrix     Genotyping Microarrays</b>	<b>17</b>
2.1 Introduction	17
2.2 Methods	19
2.3 Models and estimation	20
2.4 Results	24
2.5 Application: CNV and imbalance maps	28
2.6 Conclusions and Discussion	29
 <b>3 Single chip genotyping with semi-parametric log-concave mixtures</b>	 <b>31</b>
3.1 Introduction	31
3.2 Semi-parametric single-array genotyping	36
3.3 Comparisons	41
3.4 Conclusion and discussion	46

<b>4</b>	<b>Optimal use of low quality SNP samples by the single array approach</b>	<b>49</b>
4.1	Introduction	49
4.2	Methods	56
4.3	Results	58
4.4	Discussion	63
<b>5</b>	<b>Genomic waves: where they come from, and how to eliminate them</b>	<b>65</b>
5.1	Introduction	65
5.2	Methods	68
5.3	Empirical results	72
5.4	Discussion	75
<b>6</b>	<b>Visualizing genomic changes by segmented smoothing using an <math>L_0</math> penalty</b>	<b>77</b>
6.1	Introduction	77
6.2	Statistical methods	80
6.3	Simulations	91
6.4	Applications	93
6.5	Discussion	94
<b>7</b>	<b>The SCALA software suite</b>	<b>101</b>
7.1	Introduction	101
7.2	Functions and implementation	103
7.3	Illustrative examples	108
7.4	Technical model details	113
7.5	Discussion	117
<hr/>		
<b>8</b>	<b>Discussion</b>	<b>119</b>
8.1	Advantages of single array analysis	119
8.2	A short review	120
8.3	Ideas for future research	122
<hr/>		



<b>Bibliography</b>	<b>127</b>
<b>Samenvatting (summary in Dutch)</b>	<b>135</b>
<b>Summary</b>	<b>140</b>
<b><i>Appendices</i></b>	<b>143</b>
<hr/>	
<b>A Fluorescence bias: calibration result tables</b>	<b>147</b>
<b>B Genotyping: coding scheme</b>	<b>153</b>
<b>C Waves correction: result tables</b>	<b>155</b>
C.1 Sample GBM 139	156
C.2 Sample GBM 180	160
<b>D Manual: SCALA Suite</b>	<b>165</b>
D.1 Introduction	165
D.2 The SCALA object class	167
D.3 SCALA.convert: CEL file conversion	168
D.4 SCALA.global: calibration	170
D.5 SCALA.call: single array genotyping	172
D.6 SCALA.map: CNV / LOH mapping	174
D.7 Appendix: The SCALA model	178
<hr/>	
<b>Subject Index</b>	<b>181</b>
<b>Notes</b>	<b>184</b>
<b>Curriculum Vitae</b>	<b>187</b>



# LIST OF FIGURES

---

1.1	Nucleotides and nucleosides: molecules and bases	2
1.2	DNA: a double helix structure	3
1.3	An Affymetrix chip	6
1.4	An Illumina well-plate	6
1.5	Three-mode representation of SNP data	10
1.6	Structural signal levels in SNP data	11
2.1	Model fit example for chromosome 1 on an Affymetrix SNP6.0 chip	25
2.2	Signal bandwidths before and after calibration	26
2.3	Reduced noise in CNV profiles before and after calibration	28
2.4	Reduced noise in allelic imbalance before and after calibration	29
3.1	Multi-array genotype calling	32
3.2	Data transformation to Ratio-Sum orientation	34
3.3	Single array genotype calling	35
3.4	Comparing per-chromosome clusters to overall clusters	37
3.5	Calling methods on (a)symmetric arrays	40
3.6	Semi-parametric probabilities for a symmetric and asymmetric array	42
3.7	Call disagreements with HapMap	45
3.8	Distribution of $p_{max}$ against MAF	46
4.1	Ratio-Sum plot of high and low quality arrays	51
4.2	CNV and imbalance plots for high and low quality arrays	53
4.3	Illustration of genotype-free calibration on a high quality array	55
4.4	Single chip callrates for 10 threshold levels, for uncalibrated signals	59
4.5	Call rate improvements after genotype-free calibration	60
4.6	Aberration profiles after threshold signal selection	62
4.7	Gradient selection of thresholded signals	63
5.1	Waves are systematic bias	67

5.2	How smoothing produces “waves”	68
5.3	Wave patterns before and after calibration with SCALA	73
5.4	Calibration effects for healthy chromosome 1 and tumor chromosome 9	74
6.1	Illustrations of copy numbers and allelic ratio, expressed as logarithms, for healthy and tumor tissue	78
6.2	Illustration of smoothing with different norms (2,1,0) in the roughness penalty	82
6.3	Odd-even cross-validation for finding an optimal $\lambda$	85
6.4	Illustration of convergence behavior in zero-norm smoothing with little noise	87
6.5	Illustration of convergence behavior in zero-norm smoothing with moderate noise	88
6.6	Comparing normal and segmented scatterplot smoothing	89
6.7	ZEN smoothing of CNV in tumor data (sample GBM139.CEL)	94
6.8	ZEN smoothing of log allelic ratio (sample GBM 139.CEL)	95
6.9	Examples of smoothed CNV and allelic imbalance in clinical samples, using ZEN	96
6.10	Examples of smoothed CNV in clinical samples, using CNAG software	97
6.11	Histograms and estimated normal mixtures for the log allelic ratio	98
7.1	Graphical User Interface for SCALA.map	106
7.2	Example CNV and imbalance map after user-tuned analysis	112
7.3	Raw data with estimated smooth genotype densities	114
7.4	Illustration of signal calibration	116
8.1	Unsuccessful signal calibration on asymmetric arrays	122
8.2	Three models combined	124
8.3	The Michelin model for normal and diseased tissue	125
8.4	The Michelin model: three single views	126
8.5	Segment-wise genotype components	127

# LIST OF TABLES

---

3.1	Frequencies of total number of different genotypes	33
3.2	Genotype agreement in one Affymetrix SNP6.0 array: SCALA vs HapMap	43
3.3	Genotype agreement in one Affymetrix 100k array: SCALA vs HapMap	43
3.4	Average genotype call agreement: SCALA vs HapMap	43
3.5	Average genotype call agreement: CRLMM vs HapMap	44
3.6	Average genotype call agreement: SCALA vs GenoSNP	44
4.1	Call rate improvements after genotype-free calibration	60
6.1	Zero-norm performance (ZEN)	92
A.1	SCALA fit for Affymetrix 100k Hind	147
A.2	SCALA fit for Affymetrix 100k Xba	148
A.3	SCALA fit for Affymetrix 500k NSP	149
A.4	SCALA fit for Affymetrix 500k STY	150
A.5	SCALA fit for Affymetrix SNP6.0	151
C.1	Sample GBM 139: Wave removal for two methods, $\lambda = 1$	156
C.2	Sample GBM 139: Wave removal for two methods, $\lambda = 10$	157
C.3	Sample GBM 139: Wave removal for two methods, $\lambda = 100$	158
C.4	Sample GBM 139: Wave removal for two methods, $\lambda = 1000$	159
C.5	Sample GBM 180: Wave removal for two methods, $\lambda = 1$	160
C.6	Sample GBM 180: Wave removal for two methods, $\lambda = 10$	161
C.7	Sample GBM 180: Wave removal for two methods, $\lambda = 100$	162
C.8	Sample GBM 180: Wave removal for two methods, $\lambda = 1000$	163

