

An Integrated Analysis Tool for Analyzing Hybridization Intensities and Genotypes Using New-Generation Population-Optimized Human Arrays

> Mei-Chu Huang @JITMM 2017

# A number of GWAS for tropical medicine and global health using SNP arrays were carried out

### For **AIDS**



NIH-PA Author Manuscript

#### Author Manuscript NIH Public Access

Published in final edited form as: J Infect Dis. 2010 February 15; 201(4): 618-626. doi:10.1086/649842.

Multistage Genomewide Association Study Identifies a Locus at 1q41 Associated with Rate of HIV-1 Disease Progression to **Clinical AIDS** 

Joshua T. Herbeck<sup>1</sup>, Geoffrey S. Gottlieb<sup>2</sup>, Cheryl A. Winkler<sup>3</sup>, George W. Nelson<sup>3</sup>, Ping An<sup>3</sup>, Brandon S. Maust<sup>1</sup>, Kim G. Wong<sup>1</sup>, Jennifer L. Troyer<sup>3</sup>, James J. Goedert<sup>5</sup>, Bailey D. Kessing<sup>3</sup>, Roger Detels<sup>8</sup>, Steven M. Wolinsky<sup>10</sup>, Jeremy Martinson<sup>11</sup>, Susan Buchbinder<sup>9</sup>, Gregory D. Kirk<sup>6</sup>, Lisa P. Jacobson<sup>6</sup>, Joseph B. Margolick<sup>7</sup>, Richard A. Kaslow<sup>12</sup>, Stephen J. O'Brien<sup>4</sup>, and James I. Mullins<sup>1,2</sup>

## **For Malaria**

## LETTER

Genome-wide association study indicates two novel resistance loci for severe malaria

doi:10.1038/nature11334

Christian Timmann<sup>1,2</sup>, Thorsten Thye<sup>1,2</sup>, Maren Vens<sup>2</sup>, Jennifer Evans<sup>1,3</sup>, Jürgen May<sup>4</sup>, Christa Ehmen<sup>1</sup>, Jürgen Sievertsen<sup>1</sup>, Birgit nature

Kings Andre genetics

### Genome-wide and fine-resolution association analysis of malaria in West Africa

Muminatou Jallow<sup>1,34</sup>, Yik Ying Teo<sup>2,3,34</sup>, Kerrin S Small<sup>2,3,34</sup>, Kirk A Rockett<sup>2,3</sup>, Panos Deloukas<sup>3</sup>, Taane G Clark<sup>2,3</sup>, Katja Kivinen<sup>3</sup>, Kalifa A Bojang<sup>1</sup>, David J Conway<sup>1</sup>, Margaret Pinder<sup>1</sup>, Giorgio Sirugo<sup>1</sup>, Fatou Sisay-Joof<sup>1</sup>, Stanley Usen<sup>1</sup>, Sarah Auburn<sup>2,3</sup>, Suzannah J Bumpstead<sup>3</sup>, Susana Campino<sup>2,3</sup>, Alison Coffey<sup>3</sup>, Andrew Dunham<sup>3</sup>, Andrew E Fry<sup>2</sup>, Angela Green<sup>2</sup>, Rhian Gwilliam<sup>3</sup>, Sarah E Hunt<sup>3</sup>, Michael Inouye<sup>3</sup>, Anna E Jeffreys<sup>2</sup>, Alieu Mendy<sup>2</sup>, Aarno Palotie<sup>3</sup>, Simon Potter<sup>3</sup>, Jiannis Ragoussis<sup>2</sup>, Jane Rogers<sup>3</sup>, Kate Rowlands<sup>2</sup>, Elilan Somaskantharajah<sup>3</sup>, Pamela Whittaker<sup>3</sup>, Claire Widden<sup>3</sup>, Peter Donnelly<sup>2,4</sup>, Bryan Howie<sup>4</sup>, Jonathan Marchini<sup>2,4</sup>, Andrew Morris<sup>2</sup>, Miguel SanJoaquin<sup>2,5</sup>, Eric Akum Achidi<sup>6</sup>, Tsiri Agbenyega<sup>7</sup>, Angela Allen<sup>8,9</sup>, Olukemi Amodu<sup>10</sup>, Patrick Corran<sup>11</sup>, Abdoulaye Djimde<sup>12</sup>, Amagana Dolo<sup>12</sup>, Ogobara K Doumbo<sup>12</sup>, Chris Drakeley<sup>13,14</sup>, Sarah Dunstan<sup>15</sup>, Jennifer Evans<sup>7,16</sup>, Jeremy Farrar<sup>15</sup>, Deepika Fernando<sup>17</sup>, Tran Tinh Hien<sup>15</sup>, Rolf D Horstmann<sup>16</sup>, Muntaser Ibrahim<sup>18</sup>, Nadira Karunaweera<sup>17</sup>, Gilbert Kokwaro<sup>19</sup>, Kwadwo A Koram<sup>20</sup>, Martha Lemnge<sup>21</sup>, Julie Makani<sup>22</sup>, Kevin Marsh<sup>19</sup>, Pascal Michon<sup>8</sup>, David Modiano<sup>23</sup>, Malcolm E Molyneux<sup>5</sup>, Ivo Mueller<sup>8</sup>, Michael Parker<sup>24</sup>, Norbert Peshu<sup>19</sup>, Christopher V Plowe<sup>25,26</sup>, Odile Puijalon<sup>27</sup>, John Reeder<sup>8</sup>, Hugh Reyburn<sup>13,14</sup>, Eleanor M Riley<sup>13,14</sup>, Anavaj Sakuntabhai<sup>27</sup>, Pratap Singhasivanon<sup>28</sup>, Sodiomon Sirima<sup>29</sup>, Adama Tall<sup>30</sup>, Terrie E Taylor<sup>25,31</sup>, Mahamadou Thera<sup>12</sup>, Marita Troye-Blomberg<sup>32</sup>, Thomas N Williams<sup>19</sup>, Michael Wilson<sup>20</sup> & Dominic P Kwiatkowski<sup>2,3</sup>, Wellcome Trust Case Control Consortium<sup>33</sup> & Malaria Genomic Epidemiology Network33

# A number of studies showed CNVs are highly relevant to tropical diseases and global health

#### **Enrique Gonzalez et al. Science 2005** Eastman et al., Antimicrob. Agents Chemother., 2011 Risk of acquiring HIV relative to **Piperaquine resistance** is population median/"switch point" HIV+ -HIV-Α 95% CI associated with a CNV on chr 5 50 $\chi^2 = 39.97$ 0.60-2.74 0.5287 1.28 Switch point 5 8 40 $P = 7.7 \times 10^{\circ}$ 1.67 1.18-2.36 0.0037 Children 1.00 2 30 0.64 0.42-0.99 0 0477 Lrequen 10 0.60 0.35-1.03 0.0615 0.41 0.17-0.97 0.0433 HIV+ n=407 0.27 0.07-0.98 0.1389 **Piperaquine + dihydroartemisinin** HIV- n=395 0.02-1.70 0.1390 7 + 0.20(B has recently become the official 50 $\chi^2 = 44.7$ 0 8.12 0.93-70.8 0.0574 % 40 $P = 1.6 \times 10^{-7}$ 2.40-7.93 1 x 10<sup>-6</sup> first-line therapy in several 4.36 1.55 1.03-2.33 0.0332 2 30 1.04-2.33 0.0307 1.56 Southeast Asian countries. Prequent 10 1.00 >4 0.90 0.61-1.32 0.5961 >4 HTV+ n=409 HIV- n=497 C Cheeseman et al., Mol. Biol. Evol., 2016 $\chi^2 = 58.8$ 50 0 2.75 1.53-4.92 0.0007 % 40 1 2.41 1.84-3.16 1 x 10<sup>-</sup> Adults →2 1.00 2 30 1.08 0.81-1.44 0.5891 Lequen 10 **Deletions Amplifications** nerica A Malaria HIV<sup>+</sup> n=620 HIV<sup>-</sup> n=675 AFR AFR SEA SEA D SAM SAM 50 $\chi^2 = 25.3$ 0 10.17 0.46-226\* 0.1212 % 40 = 0.0007 1.78-13.8 0.0021 4.96 **CNVs** were 1.80 0.78-4.17 0.1707 Q 30 >3 1.00 significantly 20 Iner 04 0.17-1.39 0.1802 0 49 mericans HIV<sup>+</sup> n=69 more common in 02 02 HTV- n=101 8 10 South America 4 6 CCL3L1 gene copy number 00. 00-00 0.6 10 12 0.0 01 02 03 0.4 05 04 08 06 % of the Genome % of the Genome The **amplification** of the CCL3L1

gene was found to reduce the

risk of **HIV** progression

CNVs show diverse genetic distributions in geographically dispersed populations

## **Motivation 1**

Genotyping solution





25



\* \*Array 6.0 vs. Axiom TWB (350,075/95 = 3685 TWD/sample)



**New-Generation Population-Optimized Human Arrays:** 

**Affymetrix Axiom genotyping solution** 

# The study highlights the potential uses of Axiom genotyping solution prevalent in tropical infectious diseases



### Axiom\_aegypti1

INVESTIGATION

### A Multipurpose, High-Throughput Single-Nucleotide Polymorphism Chip for the Dengue and Yellow Fever Mosquito, *Aedes aegypti*

#### Benjamin R. Evans,\*<sup>1</sup> Andrea Gloria-Soria,\* Lin Hou,<sup>†</sup> Carolyn McBride,<sup>‡</sup> Mariangela Bonizzoni,<sup>§</sup> Hongyu Zhao,<sup>†</sup> and Jeffrey R. Powell\*

\*Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06511, <sup>†</sup>Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520, <sup>‡</sup>Princeton Neuroscience Institute and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08540, and <sup>§</sup>Department of Molecular Biology and Biochemistry, University of California Irvine, Irvine, California 92697



Major vector of dengue and yellow fever viruses

## Motivation 2

## Axiom array was originally developed for genotyping

## No public software is available

for the **integrated** genomic analysis of **hybridization intensities** and **genotypes** for this new-generation population-optimized genotyping platform



an integrated genomic analysis software

CNV

# ALICE (<u>AF/LOH/LCSH/AI/CNV/CNA Enterprise</u>) software

Adin Functions       Genome Browser       Aberration Integration         1. Type of analysis:       Unpaired analysis       Paired analysis         2. Input/output path:	
1. Type of analysis:       © Unpaired analysis       © Paired analysis         2. Input/output path:       Directory of data input:       Example         Directory of fault output:	
2. Input/output path:	
Directory of data input:       Example         Directory of result output: <b>3. Data format:</b> Genome-wide SNP array:       Affymetrix: Axiom         Input data format:       CEL-based         Genotype/Intensity-based       Path of directory of "bin" of APT (Affymetrix power tools):       C/Program Files/Affymetrix Power Tools/APT-1.19.0bin         C       Genotype/Intensity-based       "SNP markes"       NA string:       , Skip Row #:       , SNP col:       , Chr col:       , Posi col:       , Call (A) col         C       RData-based       "SNP markes"       NA string:       , Skip Row #:       , SNP col:       , Chr col:       , Posi col:       , Call (A) col         C       RData-based       Path of the list:	
Directory of result output:	
3. Data format:       Genome-wide SNP array:       Affymetrix: Axiom         Input data format:       C       CEL-based       Path of directory of "bin" of APT (Affymetrix power tools):       C/Program Files/Affymetrix Power Tools/APT-1.19.0/bin         C       Genotype/Intensity-based       "SNP marker"       NA string:       , Skip Row #:       , SNP col:       , Chr col:       , Posi col:       , Call (A) col         C       RData-based       "SNP marker"       NA string:       , Skip Row #:       , SNP col:       , Chr col:       , Posi col:       , Call (A) col         C       RData-based       "SNP marker"       NA string:       , Intensity (A)Log2Ratio col:       , Intensity (B) Strength col:       , BAF col:       .         Attachased       "SNP sto exclude from the analysis of ALICE:       Yes       Path of the list:          Attachased       C       No       © Yes       Significance level:       0.01       Genotype-specific reference:       O No         Log2-scale transformation:       O       No       © Yes       Significance level:       0.01       Genotype-specific reference:       O No         Quantile normalization:       O       No       © Yes       Number of permutation:       10000       (Window size, N of consecutive sig. markers):       Upper b	)
Genome-wide SNP array:       Affymetrix: Axiom         Input data format:       CEL-based       Path of directory of "bin" of APT (Affymetrix power tools):       C:/Program Files/Affymetrix Power Tools/APT-1.19.0bin         C       Genotype/Intensity-based       "SNP marker"       NA string:       , Strip Row #:       , SNP col:       , Chr col:       , Pooi col:       , Call (A) col         C       RData-based       "SNP marker"       NA string:       , Strip Row #:       , SNP col:       , Chr col:       , Pooi col:       , Call (A) col         C       RData-based       "SNP marker"       NA string:       , Strip Row #:       , SNP col:       , Chr col:       , Pooi col:       , Call (A) col         C       RData-based       "SNP marker"       NA string:       , SNP col:       , Intensity (B)Strength col:       , BAF col:       .         4.       Statistical analysis:       Path of the list:	
Input data format: C CEL-based Path of directory of "bin" of APT (Affymetrix power tools): C/Program Files/Affymetrix Power Tools/APT-1.19.0/bin C Genotype/Intensity-based "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) co Call (B) col: , Intensity (A)Log2Ratio col: , Intensity (B)Strength col: , BAF col: . C RData-based Provide a list of SNPs to exclude from the analysis of ALICE: C Yes Path of the list: A. Statistical analysis: Intensity data preprocessing: Log2-scale transformation: C No   Yes Significance level: 0.01 Confidence level: 0.95 Chip effect removal: C Mean C Median Minimum num. of markers: 2 Confidence level: 0.95 Import large-size data into: C RAM C Hard drive Cut-off for HI values of sig. segments: 0 Segmentation algorithm: C Raw CB8  Quick CB8	
C       Genotype/Intensity-based       "SNP marker"       NA string:       , Skip Row #:       , SNP col:       , Chr col:       , Posi col:       , Call (A) co         C       Call (B) col:       , Intensity (A)/Log2Ratio col:       , Intensity (B)/Strength col:       , BAF col:       .         C       RData-based         Provide a list of SNPs to exclude from the analysis of ALICE:       Yes	
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .   Provide a list of SNPs to exclude from the analysis of ALICE: C Yes Path of the list: Intensity data preprocessing: CNV/CNA segmentation: CNV/CNA	1:
Intensity data preprocessing:   CNV/CNA segmentation:   Log2-scale transformation:   O   No   CNV   CNV   Mean   O   Median   Minimum num. of markers:   Quantile normalization:   O   No   CNV   Provide a list of SNPs to exclude from the analysis of ALICE:   No   CNV/CNA segmentation:   Log2-scale transformation:   O   No   CNV   Segmentation:   O   No   CNV   No   CNV   CNV   CNV   CNV   Segmentation:   O   No   CNV   Segmentation:   O   No   CNV   Segmentation:   O   No   CNV   Segmentation:   O   No   CNV   Segmentation:   CNV   CNU   CNU   Confidence level:   O   O   CNU	
Provide a list of SNPs to exclude from the analysis of ALICE: O Yes Path of the list:	
4. Statistical analysis:       Intensity data preprocessing:       CNV/CNA segmentation:       AILOHLCSH/CNV/CNA detection:         Log2-scale transformation:       C       No       Yes       Significance level:       0.01       Genotype-specific reference:       C       No         Chip effect removal:       Mean       Median       Minimum num. of markers:       2       Confidence level:       0.95         Quantile normalization:       No       Yes       Number of permutations:       10000       (Window size, N of consecutive sig. markers):         Import large-size data into:       RAM       Hard drive       Cut-off for HI values of sig. segments:       0         Segmentation algorithm:       Raw CBS       Quick CBS       Quick CBS	
Intensity data processing:       CNV/CNA segmentation:       AI/LOH/LCSH/CNV/CNA detection:         Log2-scale transformation:       O       No       Import large-size data into:       O       No       Import large-size data into:       O       No       Import large-size data into:       O       Naddat       AI/LOH/LCSH/CNV/CNA detection:       Genotype-specific reference:       O       No         Import large-size data into:       O       Mean       O       Hard drive       Cut-off for HI values of sig. segments:       0       Outor       Upper bound of reference:       0.95         Import large-size data into:       Import large-size data into:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for this cance lay the segmentation algorithm:       Import large for the segmentalgorithm:       Import large for the segmentation algore	
Log2-scale transformation:       C       No       • Yes       Significance level:       0.01       Genotype-specific reference:       C       No         Chip effect removal:       • Mean       • Median       Minimum num. of markers:       2       Confidence level:       0.95         Quantile normalization:       • No       • Yes       Number of permutations:       10000       (Window size, N of consecutive sig. markers):       Upper bound of reference:       0.95         Import large-size data into:       • RAM       • Hard drive       Cut-off for HI values of sig. segments:       0	
Chip effect removal:          • Mean         • Median        Minimum num. of markers:         2           2           Confidence level:         0.95         (Window size, N of consecutive sig. markers):          Quantile normalization:          • No         • Yes           Number of permutations:         10000         (Window size, N of consecutive sig. markers):         Proportion of data to be trimmed:         0.025         (Window size, N of consecutive sig. markers):         Upper bound of reference:         0.95         (Begin and algorithm:         • Raw CBS         • Quick CBS         )         )         )	• Ye
Quantile normalization:       No       Yes       Number of permutations:       10000       (Window size, N of consecutive sig. markers):         Proportion of data to be trimmed:       0.025       Upper bound of reference:       0.95         Import large-size data into:       RAM       Hard drive       Cut-off for HI values of sig. segments:       0       0         Segmentation algorithm:       Raw CBS       Quick CBS       Quick CBS       0       0	
Proportion of data to be trimmed: 0.025 Upper bound of reference: 0.95 Import large-size data into:      RAM      Hard drive Cut-off for HI values of sig. segments: 0     Segmentation algorithm:      Raw CBS      Quick CBS	(51, 3)
Import large-size data into:	•
Segmentation algorithm: 🔿 Raw CBS 💿 Quick CBS	
5. Output:	
Numerical output: Graphical output:	
Save raw R data (* RData): 🔿 No 💽 Yes Indiv-sample figure: 🔽 AF figure Cross-sample figure: 🔽 AI figure	
Save APT output: O No O Yes 🔽 Six-panel figure 🔽 LOH/LCSH figure	
Data description: O No O Yes 🔽 CNV/CNA figure	
Individual numerical output: C No 🖸 Yes	
6. Parallel processing:	

25

## Three main components of ALICE

ALICE (AF/LOH/LCSH	ALC?	NV/CN	4 E)	nterprise,	) supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illum	ina platforms
Main Functions Genome Bro	vser A	berratio	n Inte	gration		
1. Type of analysis:		Unpaired :	inalysi	a 0	Paired analysis	
2. Input/output path:						
Directory of data input: E	ample					
Directory of result output:						
3. Data format:						-
Genome-wide SNP array: A	fymetri	ix Axion			-	
Input data format: C Cl	L-based			Pat	th of directory of "bin" of APT (Affymetrix power tools): C:Progr	ram Files/Affymetrix Power Tools/APT-1.19.0/bin
C 6	netype/	Intensity-I	hased	"22	SP marker" NA string:, Skip Row #:, SNP col: [	, Chr col: , Posi col: , Call (A) col:
				Ca	fl (B) col:, Intensity (A):Log2Ratio col:, Intensity	(B)Strength col: , BAF col: .
C 8	lata-base	ed.				
Provide a list of SNPs to em	lude from	n the anal	yaia of	ALICE: C	Yes Path of the list:	
4. Statistical analysis:						_
Intensity data preprocessi	ig:				CNV/CNA segmentation:	AI/LOH/LCSH/CNV/CNA detection:
Log2-scale transformation	с С	No	e	Yes	Significance level: 0.01	Genotype-specific reference: C No 🖲 Y
Chip effect removal:		Mean	0	Median	Minimum num. of markers: 2	Confidence level: 0.95
Quantile normalization:	- C	No	¢	Yes	Number of permutations: 10000	(Window size, N of consecutive sig. markers): (51, 3)
					Proportion of data to be trimmed: 0.025	Upper bound of reference: 0.95
		RAM	0	Hard drive	Cut-off for HI values of sig. segments: 0	
Import large-size data int	e @					
Import large-size data int	× ®				Segmentation algorithm: C Raw CBS @ Quick CBS	
Import large-size data int	× ©				Segmentation algorithm: C Raw CBS @ Quick CBS	
Import large-size data int 5. Output: Numerical output:	× ®			G	Symunitation algorithm: C Raw CBS @ Quick CBS	
Import large-size data int 5. Output: Numerical output: Save raw 8. data (* 8.Data	•	No C	Yes	G	Segmentation algorithm: C Raw CBS @ Quick CBS rephical output: Index-sample figure [7] AF figure Cross-sa	mple figure: 🔽 Al figure
Import large-size data int 5. Output: Numerical output: Save raw R data (* RData Save APT output:		No @ No C	Yes Yes	G	Segmentation algorithm: C Raw CBS & Quick CBS replated output: Indiv-sample figure: IP AF figure Cross-sa IP Inc-panel figure	engle figure: 🔽 Al figure Run 🔽 LOR LCRH figure Run
Import large-size data int 5. Output: Numerical output: Save are R data (* RData Save APT output: Data description:	 	No 💿 No 🖸 No 😨	Yes Yes Yes	G	Segmentation signifier: C Raw CBS & Quick CBS implifical output: Indiv-sample figure: P. AF figure Cross-sa P. Iin-pand figure	emple figure: 🗗 Al figure 🔽 LOHLCOH figure 🗗 CONCOA figure

Component 2: "Genome Browser."

### **Component 3: "Aberration Integration."**

WALKE interface	78 AUC method	
ALICE (AF/LOH/LCSH/AU/CNV/CNA Enterprise) — supports Adymetrix: 1106, 500K, Array 6.0, Assort and Ilumina platforms	ALICE (AF/LOH/LCSH/AUCNV/CNA Enterprise) — supports Adjunter 1016, 5016, Aray 6.0. Axism and Humina plathums	
Main Functions Genome Brower Abenation Integration	Main Functional Generate Browser Abstraction Integration	
	#USC (FF.MURCH)       - appex. diputers 100, 600. key 68 Junn and times pathes         Handle for the start brue form       - appex. diputers 100, 600. key 68 Junn and times pathes         Detersor of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Detersor of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Detersor of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Detersor of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Option of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Option of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Option of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Option of an asys       - appex. diputers 100, 600. key 68 Junn and times pathes         Option on any test of appex. diputers 100, 600. key 68 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 Junn and test of appex. diputers 100, 600. key 64 J	
	Image: Control of the second organ     Control of the second organ       Image: Control of the second organ	

## The structure of ALICE software



## Methods for the integrated genomic analysis

### Extraction of HI

 $-\begin{cases} h_A = 2^{S+0.5L} \\ h_B = 2^{S-0.5L} \\ \end{cases}$  where S: Strength, L: Log<sub>2</sub> ratio

### • Preprocessing of HI

 $-t_m = s_m - \frac{\sum_{i=1}^M s_i \cdot I[\Delta_i]}{\sum_{i=1}^M I[\Delta_i]}$ 

### AF estimation with a CPA + LIM adjustment

- (1) CPA adjustment

- $\hat{h}_{i,m} = \frac{h_{i,m}}{h_{i,m} + \kappa_{m} \cdot (1-h_{i,m})}$
- where  $\kappa_m = \frac{1}{n_m(AB)} \sum_{i=1}^{n_m(AB)} \frac{h_{i,m}}{1-h_{i,m}} + \frac{n_m(AB)}{n_m(AB)-1} \left[ \frac{\sum_{i=1}^{n_m(AB)} h_{i,m}}{1-\sum_{i=1}^{m_m(AB)} h_{i,m}} \frac{1}{n_m(AB)} \sum_{i=1}^{n_m(AB)} \frac{h_{i,m}}{1-h_{i,m}} \right]$
- (2) LIM adjustment

$$\begin{array}{l} \begin{array}{l} 1, & \text{if } h_{+,m}(AA) < h_{i,m} \\ \frac{1}{2} + \frac{1}{2} \cdot \frac{\hat{h}_{i,m} - \bar{h}_{+,m}(AB)}{\bar{h}_{+,m}(AA) - \bar{h}_{+,m}(AB)}, & \text{if } \bar{h}_{+,m}(AB) < \hat{h}_{i,m} \leq \bar{h}_{+,m}(AA) \\ \frac{1}{2} \cdot \frac{\hat{h}_{i,m} - \bar{h}_{+,m}(BB)}{\bar{h}_{+,m}(AB) - \bar{h}_{+,m}(BB)}, & \text{if } \bar{h}_{+,m}(BB) < \hat{h}_{i,m} \leq \bar{h}_{+,m}(AB) \\ 0, & \text{if } \hat{h}_{i,m} \leq \bar{h}_{+,m}(BB) \end{array}$$

### • Single-point index of AI detection

- $\bar{f}_{+,m}(g) = \frac{1}{n-(q)} \sum_{i=1}^{n_m(g)} \hat{f}_{i,m}$
- $-S_{+,m}(g) = \left[\frac{1}{n_{m}(g)-1}\sum_{i=1}^{n_{m}(g)} (\hat{f}_{i,m} \bar{f}_{+,m}(g))^{2}\right]^{1/2}$

- $-\begin{cases} CI_{+,m}^{\mathcal{A},SP}(AA) = \left[\bar{f}_{+,m}(AA) Z_{1-\frac{\alpha}{3M}} \cdot S_{+,m}(AA), 1\right] \\ CI_{+,m}^{\mathcal{A},SP}(AB) = \left[\bar{f}_{+,m}(AB) Z_{1-\frac{\alpha}{6M}} \cdot S_{+,m}(AB), \bar{f}_{+,m}(AB) + Z_{1-\frac{\alpha}{6M}} \cdot S_{+,m}(AB)\right] \\ CI_{+,m}^{\mathcal{A},SP}(BB) = \left[0, \bar{f}_{+,m}(BB) + Z_{1-\frac{\alpha}{3M}} \cdot S_{+,m}(BB)\right] \end{cases}$
- $I_{i,m}^{\mathcal{A},SP} = \begin{cases} 1, & \text{if } \hat{f}_{i,m} \notin CI_{+,m}^{\mathcal{A},SP}(AA), CI_{+,m}^{\mathcal{A},SP}(AB), \text{ or } CI_{+,m}^{\mathcal{A},SP}(BB) \\ 0, & \text{ otherwise} \end{cases}$

### Single-point index of LOH/LCSH detection

### $- CI_{+,m}^{L,SP}(AB) = \left[ \bar{f}_{+,m}(AB) - Z_{1-\frac{\alpha}{2M}} \cdot S_{+,m}(AB), \bar{f}_{+,m}(AB) + Z_{1-\frac{\alpha}{2M}} \cdot S_{+,m}(AB) \right]$

 $- I_{i,m}^{L,SP}(AB) = \begin{cases} 1, & \text{if } \hat{f}_{i,m} \notin CI_{+,m}^{L,SP}(AB) \\ 0 & \text{otherwise} \end{cases}$ 

- Single-point index of CNV/CNA detection
  - $\bar{t}_{+,m}(g) = \frac{1}{n-(q)} \sum_{i=1}^{n_m(g)} t_{i,m}$
  - $\hat{\sigma}_{+,m}(g) = \left[\frac{1}{n_m(g)-1} \sum_{i=1}^{n_m(g)} (t_{i,m} \bar{t}_{+,m}(g))^2\right]^{1/2}$
  - $CI_{+,m}^{\mathcal{C},SP}(g) = \left[\bar{t}_{+,m}(g) Z_{1-\frac{\alpha}{24}} \cdot \hat{\sigma}_{+,m}(g), \bar{t}_{+,m}(g) + Z_{1-\frac{\alpha}{24}} \cdot \hat{\sigma}_{+,m}(g)\right]$
  - $I_{i,m}^{\mathcal{CSP}}(g) = \begin{cases} 1, & \text{if } t_{i,m} > \bar{t}_{+,m}(g) + Z_1 \frac{\alpha}{2M} \cdot \hat{\sigma}_{+,m}(g) \\ -1, & \text{if } t_{i,m} < \bar{t}_{+,m}(g) Z_1 \frac{\alpha}{2M} \cdot \hat{\sigma}_{+,m}(g) \\ 0 & \text{otherwise} \end{cases}$
  - $p_{im}^{\mathcal{C},SP} = \min \left\{ 2 \left( 1 \Phi(Z_{im}^{\mathcal{C},SP}) \right) \cdot M, 1 \right\} \text{ where } Z_{im}^{\mathcal{C},SP} = (t_{im} \bar{t}_{+m}) / \hat{\sigma}_{+m} \text{ is the test statistic}$

### Multipoint indices of AI, LOH/LCSH, and **CNV/CNA detection**

- $W_{i,m}^{\mathcal{E},MP}(v,n_c) = \frac{1}{2v+1} \sum_{x \in \{m-v,m-v+1,\cdots,m,\dots,m+v-1,m+v\}} J_{i,x}^{\mathcal{E},SP} \text{ where }$  $J_{l,x}^{\mathcal{E},SP} = I[\sum_{l=1,\dots,n_c} \prod_{z=x-n_c+l,x-n_c+l+1,\dots,x+l-1} I_{l,z}^{\mathcal{E},SP} > 0]$
- (1) Confidence interval method
  - Smoothed the WAPs using the local regression LOESS function for every sample:  $\widetilde{W}_{im}^{\mathcal{E},MP}(v,n_c)$
- Calculate the Q%-quantile of the smoothed WAPs from reference samples:  $\tilde{Q}_{im}^{\mathcal{E},MP}(v,n_c)$
- First multipoint detector:  $I_{im}^{\mathcal{E},MP}(v, n_c, 1) = I[\widetilde{W}_{im}^{\mathcal{E},MP}(v, n_c) > \widetilde{Q}_{im}^{\mathcal{E},MP}(v, n_c)]$
- (2) LIM adjustment
- · Calculate the mean and standard deviation of the WAP statistics for E for all the normal reference samples:  $\hat{\mu}_{i,m}^{\mathcal{E},MP}(v,n_c)$  and  $S_{i,m}^{\mathcal{E},MP}(v,n_c)$
- Calculate the test statistic:  $Z_{l,m}^{\mathcal{E},MP}(v,n_c) = \frac{W_{l,m}^{\mathcal{E},MP}(v,n_c) \tilde{\mu}_{l,m}^{\mathcal{E},MP}(v,n_c) + \frac{1}{M}}{S_{l,m}^{\mathcal{E},MP}(v,n_c)}$
- Calculate the adjusted p value after Bonferroni correction:  $p_{i,m}^{\mathcal{E},MP}(v,n_c) = min\{[1 \phi(Z_{i,m}^{\mathcal{E},MP}(v,n_c))] \cdot M, 1\}$
- Second multipoint detector:  $I_{i,m}^{\mathcal{E},MP}(v, n_c, 2) = I[p_{i,m}^{\mathcal{E},MP}(v, n_c) < 0.05]$

In a real data analysis, we consider their combination  $I_{im}^{\mathcal{E},MP}(v,n_c) = I_{im}^{\mathcal{E},MP}(v,n_c,1) \times I_{im}^{\mathcal{E},MP}(v,n_c,2)$ 

- Ouick-CBS algorithm
  - $\quad d_{i,m}^{\mathcal{E}} = W_{i,m}^{\mathcal{E},MP}(v,n_c) Q_{i,m}^{\mathcal{E},MP}(v,n_c) \twoheadrightarrow \tilde{d}_{i,m}^{\mathcal{E}} = 2 \cdot \frac{d_{i,m}^{\mathcal{E}} c_{i,min}^{\mathcal{E}}}{c_{i,max}^{\mathcal{E}} c_{i,min}^{\mathcal{E}}} 1$
  - The weight for the *m*<sup>th</sup> SNP of the *i*<sup>th</sup> individual is calculated as follows:  $\widetilde{w}_{i,m} = \frac{w_{i,m}}{\sum_{m=1}^{M} w_{i,m}}$
  - where  $w_{i,m} = 10^{-10} + max\{\tilde{d}_{i,m}^{\mathcal{A}}, \tilde{d}_{i,m}^{\mathcal{L}}\} \cdot I[max\{\tilde{d}_{i,m}^{\mathcal{A}}, \tilde{d}_{i,m}^{\mathcal{L}}\} > 0]$
  - A weighted t test statistic based on weight Wiim is used to analyze the difference in the averages of two segments in a region of AI or LOH/LCSH.
  - A permutation test, which randomly shuffles the data in two segments, is used to calculate an empirical p value.

## Visualization of the integrated analysis results 1

### A normal sample

![](_page_10_Figure_2.jpeg)

## Visualization of the integrated analysis results <sub>2</sub>

A tumor sample

![](_page_11_Figure_2.jpeg)

# Evaluation of the performance

- Simulation study
- Real data analysis
- Validation using qPCR experiment
- Quick-CBS algorithm

## Evaluation of the performance: based on **Simulation study**

![](_page_13_Figure_2.jpeg)

# Performance of the CNV/CNA analysis using ALICE

Level of noise	Suggested	FPR and	TPR	N(SNI	<b>Ps) in the</b>	target r	egion
interference	setting $(w, n_c)$	(Simulation	scenario)	11	51	101	501
Without noise	(11, 2)	Average FPR	(neutral)	1.12	1.12	1.12	1.12
interference $(\alpha^{0}) = 00()$		Average TPR	(loss)	95.55	94.08	95.39	95.43
(q% = 0%)		Average TPR	(gain)	99.39	96.69	95.59	95.83
With noise	(51, 3)	Average FPR	(neutral)	2.19	2.19	2.19	2.19
interference ( $q$		Average TPR	(loss)	42.09	85.45	94.03	93.54
% = 23%)		Average TPR	(gain)	50.67	89.37	92.81	93.55

### The proposed CNV/CNA detection,

which integrates AI and LOH/LCSH detection,

had a **promising** *TPR* and **well-controlled FPR** in simulation studies.

## Evaluation of the performance: based on

Otherwise

## **Real data analysis**

## We analyzed **3,236 samples** genotyped using different platforms

- 3, 235 unrelated noncancerous samples
- **11 admixed samples** based on 1 lung cancer patient

![](_page_15_Figure_5.jpeg)

## Evaluation of the performance: based on

## **qPCR** experiment

![](_page_16_Figure_2.jpeg)

The results **successfully validated** the CNVs/CNAs

identified in the Axiom data analyses using the proposed methods

![](_page_17_Picture_0.jpeg)

## **ALICE** website

## http://hcyang.stat.sinica.edu.tw/software/ALICE.html

# Office (AF/LOH/LCSH/AJ/CNV/CNA Enterprise)

### Announcement (2016-10-01):

To fulfill the requirements of different Axiom SNP arrays and laboratories, we can provide customized reference databases of ALICE for you. If you have the need, please email Dr. Hsin-Chou Yang (hsinchou@stat.sinica.edu.tw).

### Introduction:

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) is user-friendly software for an integrated analysis of allele frequency (AF), allelic imbalance (AI), loss of heterozygosity (LOH), long contiguous stretch of homozygosity (LCSH), and copy number variation or alteration (CNV/CNA) on the basis of SNP probe hybridization intensities and genotypes. The software, user manual, library files for APT, annotation files, test examples, and reference databases can be downloaded below.

![](_page_17_Figure_8.jpeg)

#### Download software & user manual:

Version	Content	Screen resolution	Architecture	Size	Update date
		101100h < 000	22-bs	164 MB	2015/10/04
	Wigh < 800		64-bit	164 MB	2015/10/0
	Software	100000 - 0000	32-bit	164 MB	2015/10/04
ALICE V. 1.0		Witch 2 800 64-bit		164 MB	2015/10/0
	User manual - a brief g	1.07 MB	2015/10/0		
	User manual - a full gu		5.30 MB	2015/10/04	
	Frequently asked quest		349.5 KB	2015/10/0	

\* We provide a default example in the ALICE software. Therefore, the file is somewhat large!

Download librar	y files for Affy	metrix Power	Tools (APT):	
Affymetrix Platform	100K	500K	Array.6.0	Axiom (CHB.1)
File Size	19.6 MB	65.2 MB	79.8 MB	73.9M8
Original resource	Hind, Xba	Nsp. Sty	Attack	Axiom CHB1

#### Download annotation files:

Version	na28 (hg18/db5NP 128)	na29 (hg18/db5NP 128)	na30 (hg18/db5NP 128)	na31 (hg19/db5NP 131)	na32 (hg19/db5NP 132)	na33 (hg19/dbSNP 137)
100K	<u>ma28</u>	na29	na30	na31	na22	
500K	0228	0829	0829	na21	0822	
Array 6.0	na28	0829	6839	na31	na22	na33
Axiom					0232	na33

For Illumina: HumanHap550v3\_B.cay

ownloa	d examp	les:						
Example	1:1	1-2	1.1	2.1	2.2	2:2	2-1	2.2
Batch					Batch	Batch		
File size	390 MB	189 MB	189 MB	172 MB	147 MB	150 MB	219 MB	219 MB

Popultaion	Support platform	Innut data format	Sample	Parameter	Database index							
		anger and rotate	size	(w, nc)	1	2	3	4	5	6	7	8
	202003000000	and an and a second	210	(11.1)								1
	Affymetrix	GTC-exported		(51,3)								1
	1000000			(1001,5)								1
	Affymetrix SOOK	Alfymetrix GTC-exported 500K text file	210	(11,1)								1
Combined CEU + CHR + JPT +				(51,3)								
YRI)				(1001,5)								1
			210	(11,1)		-		-		-		
	Affymetrix Array 6.0	Affymetrix Probe results Array 6.0 file (*.CEL)		(51,3)	-			-	-	-		1
	Acres 6.0			(1001.5)	100	100	-	100	-		-	1

### **Reference (\* correspondence author):**

Mei-Chu Huang, Tze-Po Tsuang, Chien-Hsiun Chen, Jer-Yuarn Wu, Yuan-Tsong Chen, Ling-Hui Li\*, and Hsin-Chou Yang\* (2015) An Integrated Analysis Tool for Analyzing Hybridization Intensities and Genotypes Using New-Generation Population-Optimized Human Arrays. BMC Genomics 17:266. (<u>http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2478-8</u>)

## ALICE paper: Huang et al., BMC Genomics, 2016

Huang et al. BMC Genomics (2016) 17:266 DOI 10.1186/s12864-016-2478-8

BMC Genomics

Dr. Li

**Open Access** 

CrossMark

**Dr. Yang** 

### **RESEARCH ARTICLE**

An integrated analysis tool for analyzing hybridization intensities and genotypes using new-generation population-optimized human arrays

Mei-Chu Huang<sup>1,2,3</sup>, Tzu-Po Chuang<sup>4,5</sup>, Chien-Hsiun Chen<sup>6</sup>, Jer-Yuarn Wu<sup>6</sup>, Yuan-Tsong Chen<sup>6</sup>, Ling-Hui Li<sup>6\*</sup> and Hsin-Chou Yang<sup>1,2,7,8,9,10\*</sup>

Welcome for collaboration

## Acknowledgments

- We gratefully acknowledge the
  - Translational Resource Center for Genomic Medicineand National Center for Genome Medicineat Academia Sinica for providing DNA samples andgenotyping support.
- This work was supported by the Career Development Award of Academia Sinica [grant number AS-100-CDA-M03 to H.C.Y.] and a research grant from the Ministry of Science and Technology of Taiwan [grant number MOST 103-2314-B-001-008-MY3 to H.C.Y.].

![](_page_19_Picture_4.jpeg)

![](_page_19_Picture_5.jpeg)

![](_page_20_Picture_0.jpeg)

# Academia Sinica

![](_page_20_Picture_2.jpeg)

## sinica.edu.tw

- Founded in 1928 as the National Academy
- 3 main divisions in Academia Sinica, <u>division of Mathematics and Physical</u> <u>Science</u>, <u>life science</u> and <u>humanities and social science</u>. Totally, 24 research institutes and 7 research centers
- 7900+ researchers and students from Taiwan and abroad

![](_page_20_Picture_7.jpeg)

# TIGP-Bioinformatics Program

![](_page_21_Picture_1.jpeg)

TIGPBP

http://tigpbp.iis.sinica.edu.tw/tigpbio

## Algorithm and Computational Method Development

![](_page_21_Figure_4.jpeg)

# Thank you very much for the attention.

![](_page_22_Picture_1.jpeg)