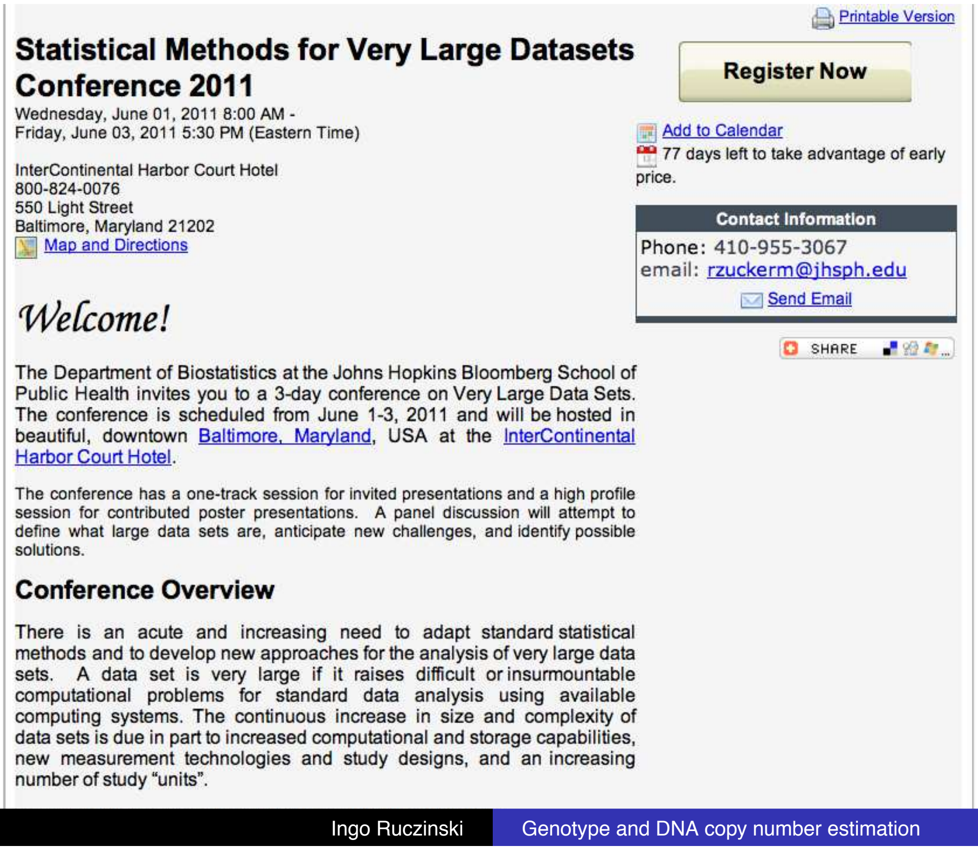
Genotype and DNA Copy Number Estimation

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

November 17, 2010

Very large data sets



Printable Version

Statistical Methods for Very Large Datasets Conference 2011

Wednesday, June 01, 2011 8:00 AM -
Friday, June 03, 2011 5:30 PM (Eastern Time)

InterContinental Harbor Court Hotel
800-824-0076
550 Light Street
Baltimore, Maryland 21202
Map and Directions

Register Now

Add to Calendar
77 days left to take advantage of early price.

Contact Information

Phone: 410-955-3067
email: rzuckerm@jhsph.edu
Send Email

SHARE

Welcome!

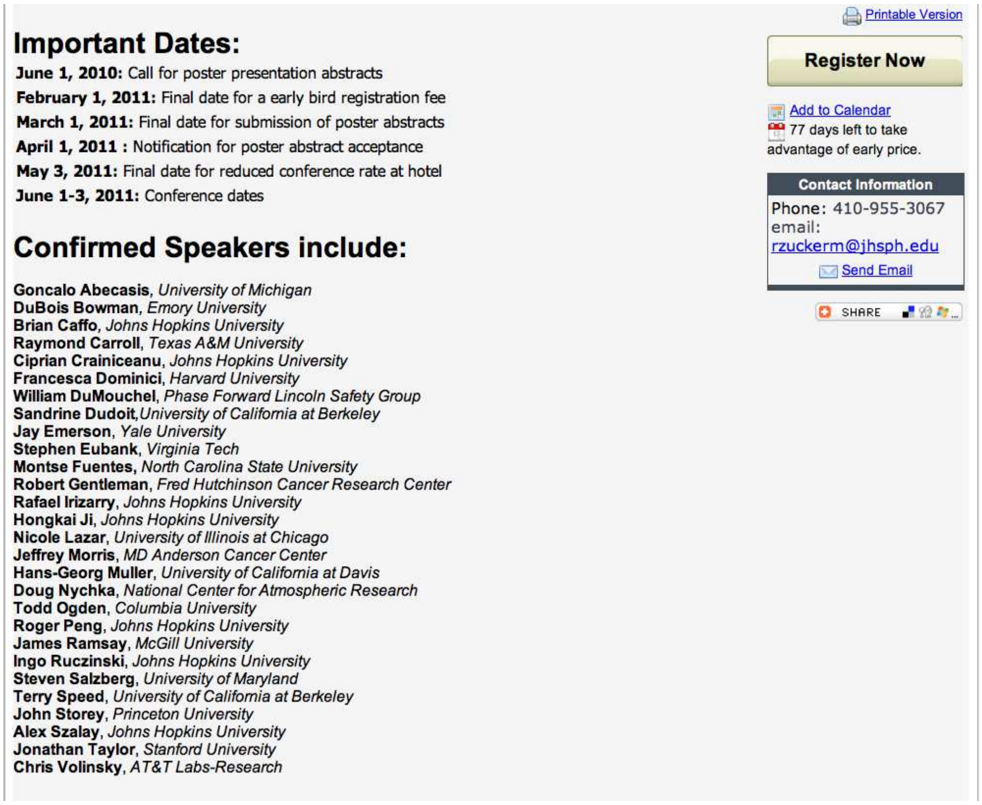
The Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health invites you to a 3-day conference on Very Large Data Sets. The conference is scheduled from June 1-3, 2011 and will be hosted in beautiful, downtown Baltimore, Maryland, USA at the InterContinental Harbor Court Hotel.

The conference has a one-track session for invited presentations and a high profile session for contributed poster presentations. A panel discussion will attempt to define what large data sets are, anticipate new challenges, and identify possible solutions.

Conference Overview

There is an acute and increasing need to adapt standard statistical methods and to develop new approaches for the analysis of very large data sets. A data set is very large if it raises difficult or insurmountable computational problems for standard data analysis using available computing systems. The continuous increase in size and complexity of data sets is due in part to increased computational and storage capabilities, new measurement technologies and study designs, and an increasing number of study "units".

Very large data sets



Printable Version

Important Dates:

June 1, 2010: Call for poster presentation abstracts
February 1, 2011: Final date for a early bird registration fee
March 1, 2011: Final date for submission of poster abstracts
April 1, 2011 : Notification for poster abstract acceptance
May 3, 2011: Final date for reduced conference rate at hotel
June 1-3, 2011: Conference dates

Register Now

Add to Calendar
77 days left to take advantage of early price.

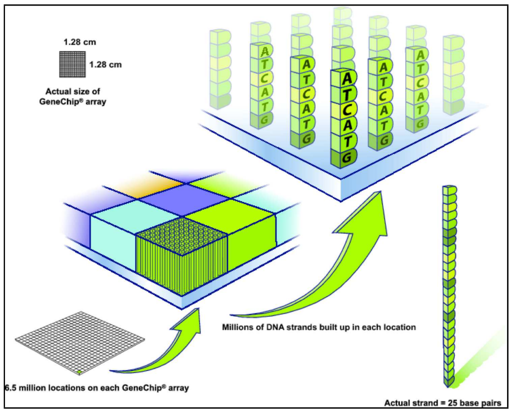
Contact Information
Phone: 410-955-3067
email:
rzuckerm@jhsph.edu
Send Email

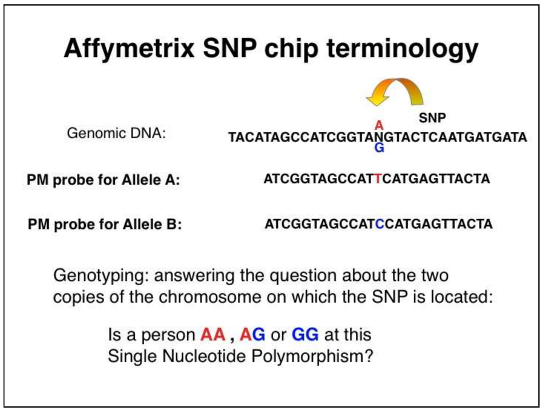
SHARE

Confirmed Speakers include:

Goncalo Abecasis, *University of Michigan*
DuBois Bowman, *Emory University*
Brian Caffo, *Johns Hopkins University*
Raymond Carroll, *Texas A&M University*
Ciprian Crainiceanu, *Johns Hopkins University*
Francesca Dominici, *Harvard University*
William DuMouchel, *Phase Forward Lincoln Safety Group*
Sandrine Dudoit,*University of California at Berkeley*
Jay Emerson, *Yale University*
Stephen Eubank, *Virginia Tech*
Montse Fuentes, *North Carolina State University*
Robert Gentleman, *Fred Hutchinson Cancer Research Center*
Rafael Irizarry, *Johns Hopkins University*
Hongkai Ji, *Johns Hopkins University*
Nicole Lazar, *University of Illinois at Chicago*
Jeffrey Morris, *MD Anderson Cancer Center*
Hans-Georg Muller, *University of California at Davis*
Doug Nychka, *National Center for Atmospheric Research*
Todd Ogden, *Columbia University*
Roger Peng, *Johns Hopkins University*
James Ramsay, *McGill University*
Ingo Ruczinski, *Johns Hopkins University*
Steven Salzberg, *University of Maryland*
Terry Speed, *University of California at Berkeley*
John Storey, *Princeton University*
Alex Szalay, *Johns Hopkins University*
Jonathan Taylor, *Stanford University*
Chris Volinsky, *AT&T Labs-Research*

Genomic arrays


1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
6.5 million locations on each GeneChip® array
Millions of DNA strands built up in each location
Actual strand = 25 base pairs



Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA (A / G)

PM probe for Allele A: ATCGGTAGCCATTCATGAGTTACTA

PM probe for Allele B: ATCGGTAGCCATCCATGAGTTACTA

Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

Is a person **AA** , **AG** or **GG** at this Single Nucleotide Polymorphism?

http://www.affymetrix.com

Copy number estimates are noisy

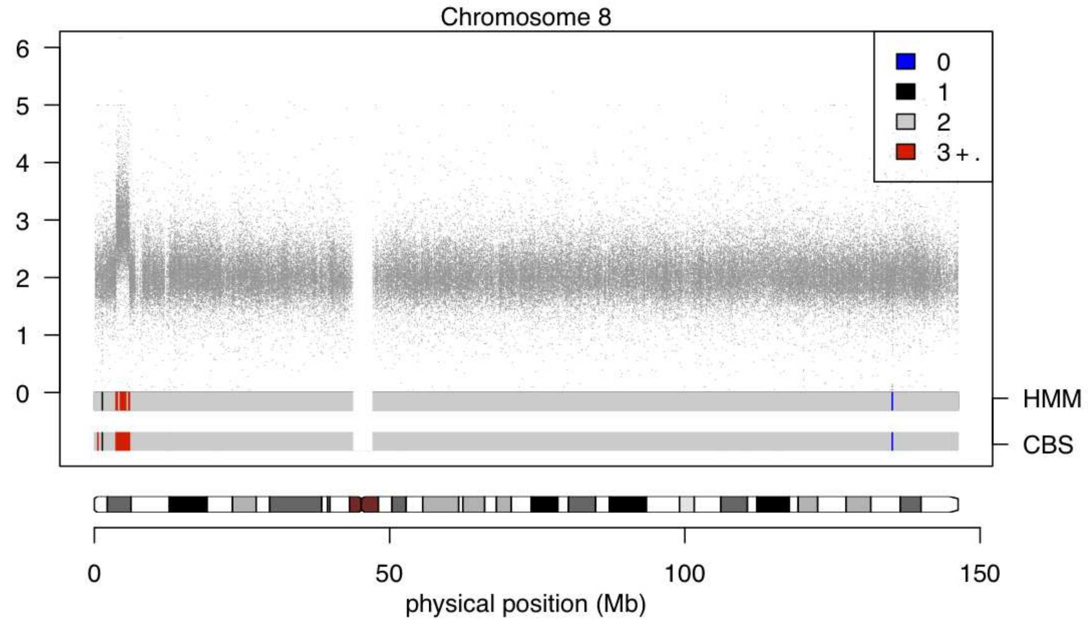
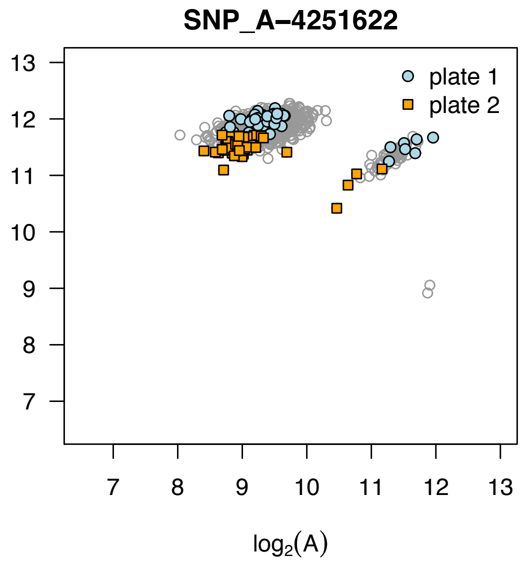
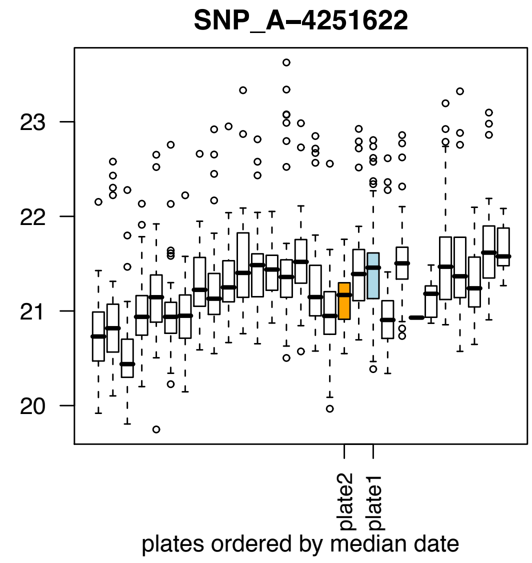
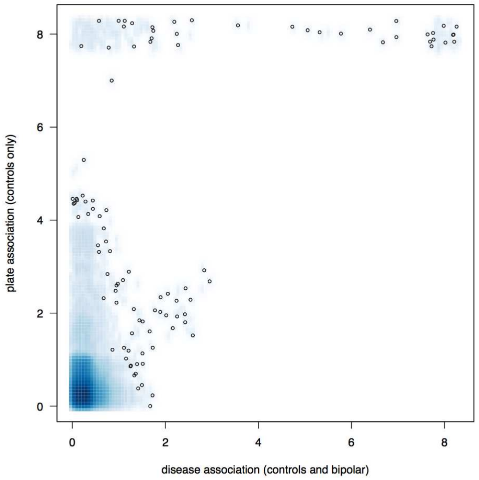


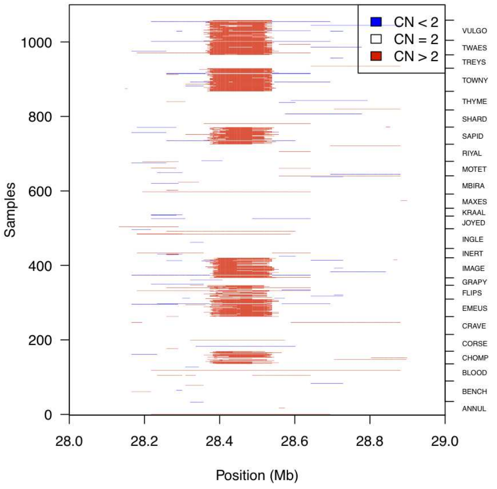
Plate effects





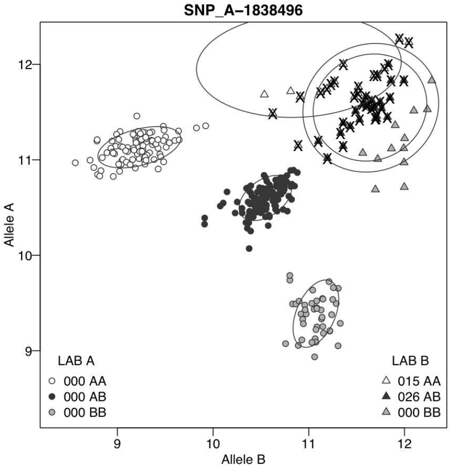
Confounding of plate and disease



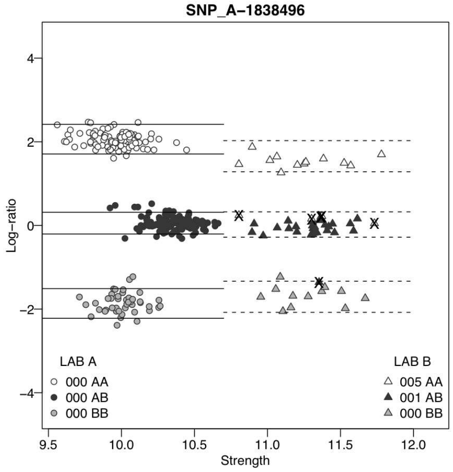


Genotype estimates are more robust

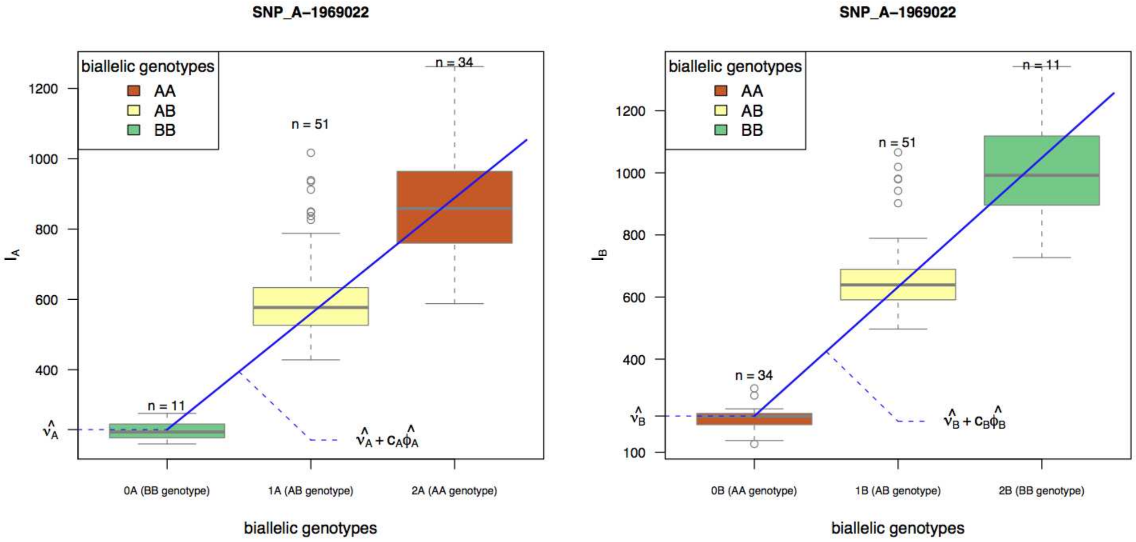
Birdseed



CRLMM



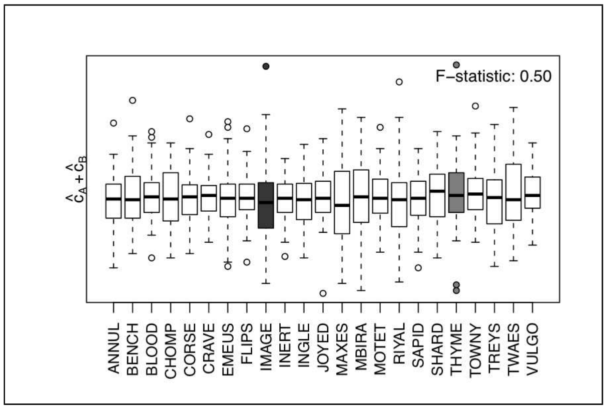
Allele specific copy numbers

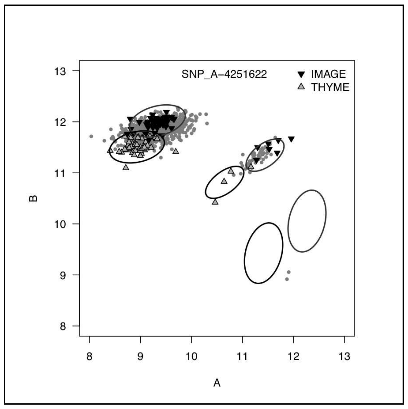


Allele specific copy numbers

At locus *i*, for subject *j* in plate *p*, we have for allele $k \in \{A, B\}$

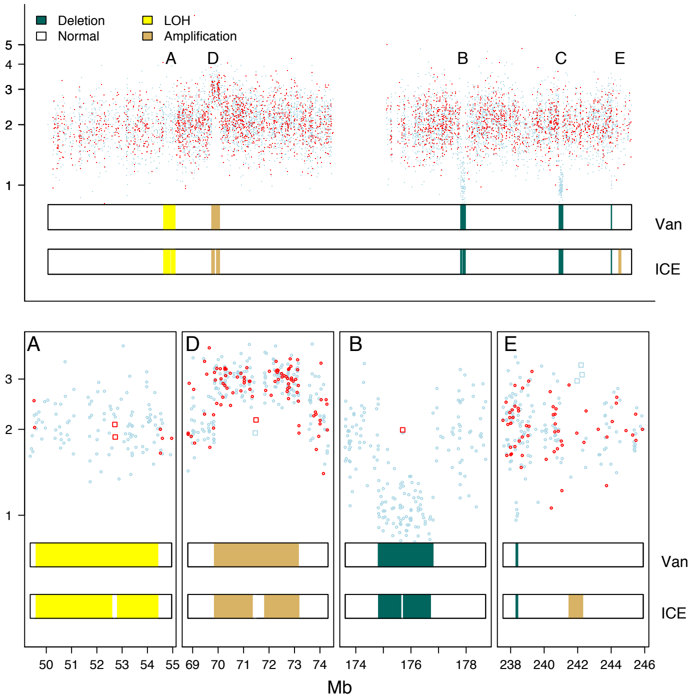
$$I_{kijp} = \nu_{kip}\delta_{kijp} + \phi_{kip}c_{kijp}\epsilon_{kijp} \implies \hat{c}_{kijp} = \max\left\{ \frac{1}{\hat{\phi}_{kip}} \left(I_{kijp} - \hat{\nu}_{kip}\right) ,\ 0 \right\}$$





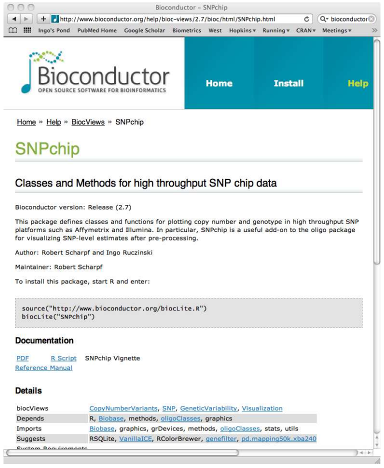
[Sch · Iri · Rit · Car · Ruc | Tech·Rep 2010] ● [Sch · Ruc · Car · Doa · Cha · Iri | Biostat 2010]

Vanilla and ICE HMMs for genotype and copy number



[SCH · PAR · PEV · RUC | AOAS 2008]

Open source software



Home » Help » BiocViews » SNPchip

SNPchip

Classes and Methods for high throughput SNP chip data

Bioconductor version: Release (2.7)

This package defines classes and functions for plotting copy number and genotype in high throughput SNP platforms such as Affymetrix and Illumina. In particular, SNPchip is a useful add-on to the oligo package for visualizing SNP-level estimates after pre-processing.

Author: Robert Scharpf and Ingo Ruczinski

Maintainer: Robert Scharpf

To install this package, start R and enter:

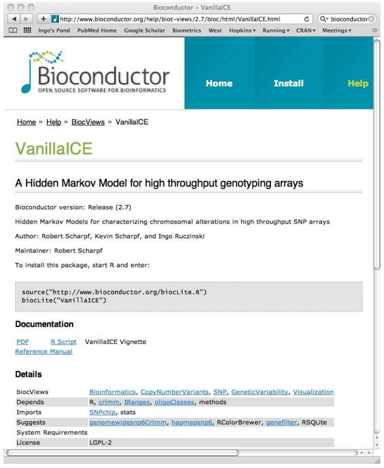
```
source("http://www.bioconductor.org/biocLite.R")
biocLite("SNPchip")
```

Documentation

PDF R Script SNPchip Vignette
Reference Manual

Details

biocViews	CopyNumberVariants, SNP, GeneticVariability, Visualization
Depends	R, Biobase, methods, oligoClasses, graphics
Imports	Biobase, graphics, grDevices, methods, oligoClasses, stats, utils
Suggests	RSQLite, VanillaICE, RColorBrewer, genefilter, pd.mapping50k.xba240



Bioconductor - VanillaICE
http://www.bioconductor.org/help/bioc-views/2.7/bioc/html/VanillaICE.html
bioconductor
Ingo's Pond
PubMed Home
Google Scholar
Biometrics
West
Hopkins
Running
CRAN
Meetings
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Home
Install
Help

Home » Help » BiocViews » VanillaICE

VanillaICE

A Hidden Markov Model for high throughput genotyping arrays

Bioconductor version: Release (2.7)

Hidden Markov Models for characterizing chromosomal alterations in high throughput SNP arrays

Author: Robert Scharpf, Kevin Scharpf, and Ingo Ruczinski

Maintainer: Robert Scharpf

To install this package, start R and enter:

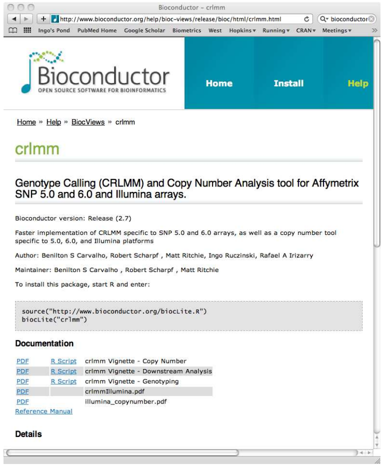
```
source("http://www.bioconductor.org/biocLite.R")
biocLite("VanillaICE")
```

Documentation

PDF R Script VanillaICE Vignette
Reference Manual

Details

biocViews	Bioinformatics, CopyNumberVariants, SNP, GeneticVariability, Visualization
Depends	R, crlmm, IRanges, oligoClasses, methods
Imports	SNPchip, stats
Suggests	genomewidesnp6Crlmm, hapmapsnp6, RColorBrewer, genefilter, RSQLite
System Requirements	
License	LGPL-2



Home » Help » BiocViews » crlmm

crlmm

Genotype Calling (CRLMM) and Copy Number Analysis tool for Affymetrix SNP 5.0 and 6.0 and Illumina arrays.

Bioconductor version: Release (2.7)

Faster implementation of CRLMM specific to SNP 5.0 and 6.0 arrays, as well as a copy number tool specific to 5.0, 6.0, and Illumina platforms

Author: Benilton S Carvalho, Robert Scharpf , Matt Ritchie, Ingo Ruczinski, Rafael A Irizarry

Maintainer: Benilton S Carvalho , Robert Scharpf , Matt Ritchie

To install this package, start R and enter:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("crlmm")
```

Documentation

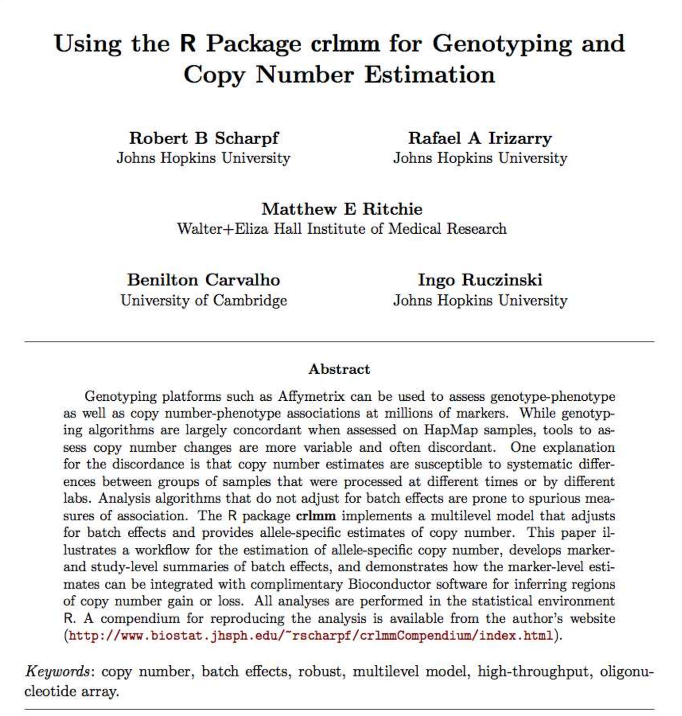
PDF	R Script	crlmm Vignette - Copy Number
PDF	R Script	crlmm Vignette - Downstream Analysis
PDF	R Script	crlmm Vignette - Genotyping
PDF		crlmmIllumina.pdf
PDF		illumina_copynumber.pdf

Reference Manual

Details

[SCH · · · RUC | BIOINF 2007] ● [SCH · RUC | M·MOL·BIO 2010] ● [SCH · RUC · · · IRI | BIOSTAT 2010]

A software vignette



Using the R Package crlmm for Genotyping and Copy Number Estimation

Robert B Scharpf
Johns Hopkins University

Rafael A Irizarry
Johns Hopkins University

Matthew E Ritchie
Walter+Eliza Hall Institute of Medical Research

Benilton Carvalho
University of Cambridge

Ingo Ruczinski
Johns Hopkins University

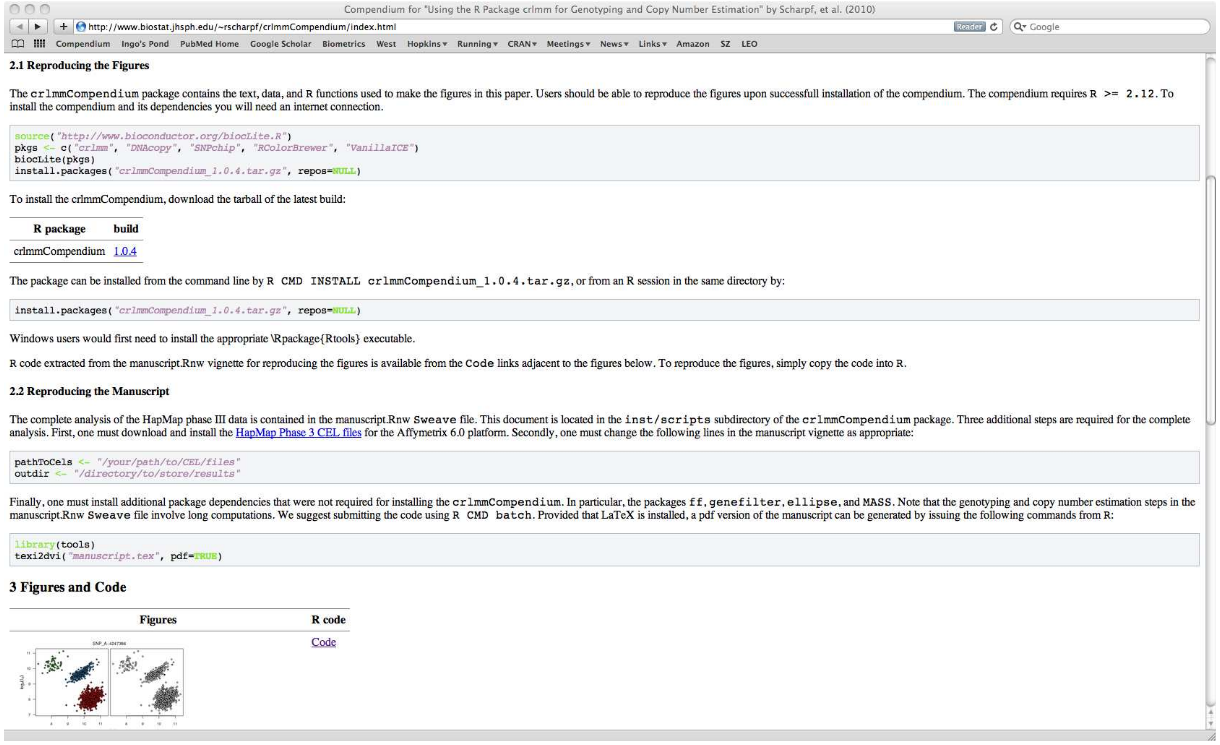
Abstract

Genotyping platforms such as Affymetrix can be used to assess genotype-phenotype as well as copy number-phenotype associations at millions of markers. While genotyping algorithms are largely concordant when assessed on HapMap samples, tools to assess copy number changes are more variable and often discordant. One explanation for the discordance is that copy number estimates are susceptible to systematic differences between groups of samples that were processed at different times or by different labs. Analysis algorithms that do not adjust for batch effects are prone to spurious measures of association. The R package **crlmm** implements a multilevel model that adjusts for batch effects and provides allele-specific estimates of copy number. This paper illustrates a workflow for the estimation of allele-specific copy number, develops marker- and study-level summaries of batch effects, and demonstrates how the marker-level estimates can be integrated with complimentary Bioconductor software for inferring regions of copy number gain or loss. All analyses are performed in the statistical environment R. A compendium for reproducing the analysis is available from the author's website (`http://www.biostat.jhsph.edu/~rscharpf/crlmmCompendium/index.html`).

Keywords: copy number, batch effects, robust, multilevel model, high-throughput, oligonucleotide array.

[SCH · IRI · RIT · CAR · RUC | TECH·REP 2010]

Compendium



Compendium for "Using the R Package crlmm for Genotyping and Copy Number Estimation" by Scharpf, et al. (2010)

http://www.biostat.jhsph.edu/~rscharpf/crlmmCompendium/index.html

2.1 Reproducing the Figures

The `crlmmCompendium` package contains the text, data, and R functions used to make the figures in this paper. Users should be able to reproduce the figures upon successfull installation of the compendium. The compendium requires `R >= 2.12`. To install the compendium and its dependencies you will need an internet connection.

```
source("http://www.bioconductor.org/biocLite.R")
pkgs <- c("crlmm", "DNAcopy", "SNPchip", "RColorBrewer", "VanillaICE")
biocLite(pkgs)
install.packages("crlmmCompendium_1.0.4.tar.gz", repos=NULL)
```

To install the crlmmCompendium, download the tarball of the latest build:

R package	build
crlmmCompendium	1.0.4

The package can be installed from the command line by `R CMD INSTALL crlmmCompendium_1.0.4.tar.gz`, or from an R session in the same directory by:

```
install.packages("crlmmCompendium_1.0.4.tar.gz", repos=NULL)
```

Windows users would first need to install the appropriate \Rpackage{Rtools} executable.

R code extracted from the manuscript.Rnw vignette for reproducing the figures is available from the `Code` links adjacent to the figures below. To reproduce the figures, simply copy the code into `R`.

2.2 Reproducing the Manuscript

The complete analysis of the HapMap phase III data is contained in the manuscript.Rnw `Sweave` file. This document is located in the `inst/scripts` subdirectory of the `crlmmCompendium` package. Three additional steps are required for the complete analysis. First, one must download and install the HapMap Phase 3 CEL files for the Affymetrix 6.0 platform. Secondly, one must change the following lines in the manuscript vignette as appropriate:

```
pathToCels <- "/your/path/to/CEL/files"
outdir <- "/directory/to/store/results"
```

Finally, one must install additional package dependencies that were not required for installing the `crlmmCompendium`. In particular, the packages `ff`, `genefilter`, `ellipse`, and `MASS`. Note that the genotyping and copy number estimation steps in the manuscript.Rnw `Sweave` file involve long computations. We suggest submitting the code using `R CMD batch`. Provided that LaTeX is installed, a pdf version of the manuscript can be generated by issuing the following commands from `R`:

```
library(tools)
texi2dvi("manuscript.tex", pdf=TRUE)
```

3 Figures and Code

Figures	R code
	Code

References

Scharpf RB, Ting JC, Pevsner J, Ruczinski I (2007).
SNPchip: R classes and methods for SNP array data.
Bioinformatics, 23(5): 627-8.

Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I (2008).
Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays.
The Annals of Applied Statistics, 2(2): 687-713.

Scharpf RB, Ruczinski I (2010).
R classes and methods for SNP array data.
Methods in Molecular Biology 593: 67-79.

Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA (2010).
A multi-level model to address batch effects in copy numbers using SNP arrays.
Biostatistics (to appear).

Scharpf RB, Irizarry RA, Ritchie M, Carvalho B, Ruczinski I (2010).
Using the R package crlmm for genotyping and copy number estimation.
Technical Report.

http: //biostat.jhsph.edu/∼iruczins/

ingo@jhu.edu