

Genome analysis

SNPchip: R classes and methods for SNP array dataRobert B. Scharpf¹, Jason C. Ting², Jonathan Pevsner² and Ingo Ruczinski^{1,*}¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205 and²Department of Neurology, Kennedy Krieger Institute, Baltimore, MD 21205, USA

Received on February 10, 2006; revised on November 15, 2006; accepted on December 14, 2006

Advance Access publication January 4, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Summary: High-density single nucleotide polymorphism microarrays (SNP chips) provide information on a subject's genome, such as copy number and genotype (heterozygosity/homozygosity) at a SNP. While fluorescence *in situ* hybridization and karyotyping reveal many abnormalities, SNP chips provide a higher resolution map of the human genome that can be used to detect, e.g., aneuploidies, microdeletions, microduplications and loss of heterozygosity (LOH). As a variety of diseases are linked to such chromosomal abnormalities, SNP chips promise new insights for these diseases by aiding in the discovery of such regions, and may suggest targets for intervention. The R package *SNPchip* contains classes and methods useful for storing, visualizing and analyzing high-density SNP data. Originally developed from the SNPscan web-tool, *SNPchip* utilizes S4 classes and extends other open source R tools available at Bioconductor. This has numerous advantages, including the ability to build statistical models for SNP-level data that operate on instances of the class, and to communicate with other R packages that add additional functionality.

Availability: The package is available from the Bioconductor web page at www.bioconductor.org

Contact: ingo@jhu.edu

Supplementary information: The supplementary material as described in this article (case studies, installation guidelines and R code) is available from <http://biostat.jhsph.edu/~iruczins/publications/sm/>

INTRODUCTION

Single nucleotide polymorphisms (SNPs), also called single base pair differences, are natural variations in the human genome, estimated to occur about every 1000 bp (The SNP Consortium, <http://snp.cshl.org/>). SNP arrays are a high-throughput tool to assess copy number changes, such as aneuploidy, deletions and cancer-associated amplifications (Bignell *et al.*, 2004; Matsuzaki *et al.*, 2004). For example, Affymetrix SNP chips provide data from 11 555 SNPs for the Mapping 10K array, to more than 500 000 loci for the Mapping 500K arrays (www.affymetrix.com). Several recent papers have developed algorithms for genotype calls (e.g. Carvalho *et al.*, 2006) and copy number estimation (e.g. Huang *et al.*, 2006, Laframboise *et al.*, 2006). Using pre-processed data as a starting point, a common goal is to identify chromosomal features spanning one or more SNPs.

Many chromosomal features can be detected by estimates of genotype calls and copy number. Copy number estimates are useful

for classifying homozygous deletions (zero copies), hemizygous deletions (one copy), gene duplications (greater than two copies) and mosaicism (non-integer copies). Gene function may also be modulated by mechanisms that do not affect copy number (copy neutral). For instance, uniparental isodisomy (UPD) occurs when a subject inherits two copies of a chromosome or chromosomal segment from 1 parent. Regions of LOH, of which UPD is a special case, are detectable in high-throughput SNP assays by a reduction in the proportion of SNPs called heterozygotes. Joint estimates of copy number and genotype calls can discriminate hemizygous deletion LOH from copy neutral LOH. Tools for organizing and visualizing SNP-level summaries of copy number and genotype calls are needed for the development of statistical models that identify chromosomal features.

R PACKAGE OVERVIEW

R is a 'free software environment for statistical computing and graphics' (<http://www.r-project.org>), available for all common platforms. The R package *SNPchip* is written using S4-style classes and methods (Chambers, 1998), and is freely available from the Bioconductor web page at www.bioconductor.org (Gentleman *et al.*, 2004). Additional details regarding implementation are available in the Bioconductor *SNPchip* vignette, and the Supplementary materials.

The Bioconductor package *Biobase* defines class structures that are useful for organizing high-dimensional genomic data, of which high-throughput SNP data is a special case. *SNPchip* extends the eSet *Biobase* classes for high-throughput data, defining three classes for SNP data corresponding to whether copy number estimates, genotype calls or both are available for each SNP. Annotation for each SNP (e.g. physical position) and chromosome (e.g. chromosome size and centromere location) are important aspects of the visualization. Methods for plotting and summarizing SNP-level data in the context of their physical position are illustrated in the Supplementary material. To provide immediate functionality, we have posted R objects of static annotation tables for download at our website (see the Supplementary material for loading instructions). The appropriate annotation for 10k, 100k and 500k Affymetrix SNP chips will be downloaded automatically; otherwise, explicit instructions are provided. Annotation for other high-throughput platforms are under development at Bioconductor. Integration of *SNPchip* with other open source software, such as the R package *RSNPper*, which maps SNPs to genes, can be useful for obtaining a richer annotation.

*To whom correspondence should be addressed.

APPLICATIONS

Visualizing the entire genome for aberrations in copy number estimates and genotype calls has been described by others (Beroukhim *et al.*, 2006, Laframboise *et al.*, 2006). In particular, *SNPchip* was spawned from a web-based tool (<http://pevsnerlab.kennedykrieger.org/snpscan.htm> and Ting *et al.*, 2006). Web-based tools are convenient for researchers to visualize and inspect the data, but are less useful as a means for building additional architecture. We have incorporated the plotting capabilities of the SNPscan web tool into our package. Users of *SNPchip* can further define methods of the class to generate diagnostics that may be useful for a particular dataset, or as a way of organizing data for higher-level analyses. Several examples are shown in the Supplementary material web site. To illustrate how we have extended the functionality of *SNPchip* in our own work, Figure 1 overlays the predicted states from a hidden Markov model (HMM) fit to high-throughput copy number estimates in chromosome 1 of a normal male. The deletion in the p-arm and amplification of the q-arm are features inserted for illustration.

DISCUSSION

Classes for efficient storage of high-dimensional genomic data have been developed in the R package *Biobase*. We extended the classes defined in *Biobase* to accommodate SNP chip data and have added methods useful for producing visual and descriptive summaries. In particular, the plotting methods are useful for identifying regions of probable chromosomal anomalies, with the capability of producing both broad (genome-wide) and focused (chromosome-specific) views of copy number and genotype data. A nice feature of having the three class structure for SNP-level data is that the visual tools and the statistical models built on such classes will work regardless of the software used to produce SNP-level estimates of genotype or copy number. One may proceed from raw data to HMMs for copy number within R using open source software. As the scope and breadth of annotation possible for each SNP evolves, tools to access the large amount of annotation without overburdening memory limits of personal computers will be necessary. Towards this end, annotation packages built through an SQL-interface are under development at Bioconductor. We have implemented a β -version, but this is currently not well supported. While we have provided annotation for the Affymetrix platform, *SNPchip* will be useful for other high throughput SNP platforms (e.g. Illumina) once annotation becomes available. There are many R packages in Bioconductor useful for visualizing, annotating and modeling high-throughput data. *SNPchip* utilizes this infrastructure in providing useful tools for high-throughput SNP data.

ACKNOWLEDGEMENTS

The authors thank Seth Falcon, Martin Morgan, Rafael Irizarry, Benilton Carvalho and Giovanni Parmigiani for suggestions, which greatly improved this manuscript. R.B.S. was supported by the training grant 5T32ES012871 from the U. S. National

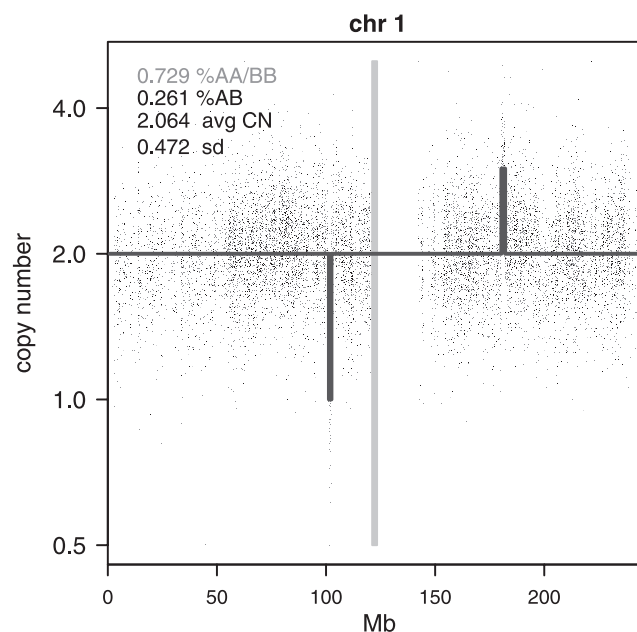


Fig. 1. Copy number estimates of SNPs on chromosome 1, plotted at their physical position on the chromosome. Typically the estimates are color coded, differentiating homozygous and heterozygous genotype calls (available in the Supplementary material). To illustrate the possibility to extend the functionality of *SNPchip*, the predicted copy number using a hidden Markov model is superimposed as a solid line on the data, indicating regions of gene copy number loss and amplification.

Institute of Environmental Health Sciences (P. I. Thomas Louis) and grant DMS034211 from the National Science Foundation (P. I. Giovanni Parmigiani). J.P. was supported by NIH grants R01 HD046598 and MRDDRC HD24061. I.R. was supported by NIH grant CA074841.

Conflict of Interest: none declared.

REFERENCES

- Bignell,G.R. *et al.* (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.*, **14**, 287–295.
- Beroukhim,R. *et al.* (2006) Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput. Biol.*, **2**, e41.
- Carvalho,B. *et al.* (2006) Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. *Biostatistics*, in press.
- Chambers,J.M. (1998) *Programming With Data*. Springer, New York.
- Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
- Huang,J. *et al.* (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*, **7**, 83.
- Laframboise,T. *et al.* (2006) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics*.
- Matsuzaki,H. *et al.* (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nat. Meth.*, **1**, 109–111.
- Ting,J.C. *et al.* (2006) Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics*, **7**, 25.