BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 2 2009, pages 281–283
doi:10.1093/bioinformatics/btn587

Genetics and population analysis

The SNPMaP package for R: a framework for genome-wide association using DNA pooling on microarrays

Oliver S. P. Davis*, Robert Plomin and Leonard C. Schalkwyk

Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK

Received on July 14, 2008; revised on October 07, 2008; accepted on November 10, 2008

Advance Access publication November 12, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Large-scale genome-wide association (GWA) studies using thousands of high-density SNP microarrays are becoming an essential tool in the search for loci related to heritable variation in many phenotypes. However, the cost of GWA remains beyond the reach of many researchers. Fortunately, the majority of statistical power can still be obtained by estimating allele frequencies from DNA pools, reducing the cost to that of tens, rather than thousands of arrays. We present a set of software tools for processing SNPMaP (SNP microarrays and pooling) data from CEL files to Relative Allele Scores in the rich R statistical computing environment.

Availability: The SNPMaP package is available from http://cran.r-project.org/ under the GNU General Public License version 3 or later.

Contact: snpmap@iop.kcl.ac.uk

Supplementary information: Additional resources and test datasets are available at http://sgdp.iop.kcl.ac.uk/snpmap/

1 INTRODUCTION

Genetic variation has been an important factor in nearly every aspect of human health and disease. This takes forms ranging from single-locus (Mendelian) traits to small effects from numerous loci. The early triumphs of human genetics and positional cloning involved rare, drastic, single mutations. These were amenable to recombination mapping techniques using large multi-generation pedigrees, and allowed the genome to be scanned for linkage using just a few hundred DNA markers. The extension of these techniques to quantitative traits offered the same systematic genome coverage but poor statistical power for detecting loci of small effect (Balding, 2006).

Unrelated individuals from a population are much easier to recruit in large numbers than are families, and individual candidate mutations can be tested for association with a phenotypic trait with good statistical power. This association generally extends to nearby sequences as well, because of the relative rarity of new mutations and historical recombination in the population, giving rise to linkage disequilibrium (Slatkin, 2008). This indirect association allows scanning of candidate loci or regions. Given enough polymorphic markers, association scanning can be extended genome wide (McCarthy *et al*., 2008). The number of markers required for a truly comprehensive scan of the human genome is thought to be in the order of a million (Barrett and Cardon, 2006). SNP genotyping microarrays have made it possible to genotype individuals at a million loci quickly, and genome-wide association (GWA) studies are now being used to scan the human genome for loci related to heritable variation in many phenotypes (Wellcome Trust Case Control Consortium, 2007).

Many population samples with associated phenotypic information could yield valuable insights through GWA analysis, but the funding and industrial-scale infrastructure required for comprehensive genotyping of thousands of individuals will not be available; nor will funding be available to genotype samples again as new microarrays emerge. Additionally, given the extreme degree of multiple testing, the current generation of GWA experiments is far from definitive, and many more studies will be needed to fully confirm findings.

Fortunately, the majority of statistical power can be obtained by estimation of allele frequencies from DNA pools, in which small amounts of DNA from individuals in a group (e.g. cases or controls) are combined on the same microarray. This in effect averages allele frequencies biologically (rather than genotyping each individual and averaging their allele frequencies statistically). Importantly, the cost in time and effort for GWA studies using pooled DNA is well within the scope of a PhD project or post-doctoral contract even when several (usually 10 or more) independent pools are created to represent each group (Butcher *et al*., 2004; Kirov *et al*., 2006; Pearson *et al*., 2007). We present a set of software tools for processing SNPMaP (SNP microarrays and pooling) data in the increasingly popular R environment for statistical computing (R Development Core Team, 2008; http://www.r-project.org).

2 SOFTWARE OVERVIEW

The SNPMaP package has been designed to handle the processing of SNPMaP data from the CEL files generated by the Affymetrix GeneChip Command Console (AGCC) or GeneChip Operating Software (GCOS), through to the RAS (Relative Allele Scores—the pooling equivalent of a relative allele frequency; Butcher *et al*., 2008) used in most analyses. This can be as simple as typing

```
mras <- snpmap ()
```

at the R prompt. The package will identify and read in the CEL files from the current directory, extract the relevant probe intensities

*To whom correspondence should be addressed.

and calculate a mean RAS for each SNP on each chip, returning a SNPMaP S4 object containing the scores.

Given the large amount of data generated by current SNP arrays, even with the relatively modest numbers of arrays (tens) typical of SNPMaP experiments, we have provided a `lowMemory` option that uses memory-mapping (with the R.huge package) to allow analysis to be done on a 32-bit desktop PC with 1 GB of memory (naturally there is a speed penalty). If memory limits are exceeded in the course of the analysis, SNPMaP attempts to automatically switch from storing objects in memory to storing objects on disk. Most SNPMaP analyses use 20 to 30 arrays, so, as a severe test, we generated RAS summaries from 50 Affymetrix 6.0 arrays simultaneously on a 2.4 GHz Pentium 4 system with 1 GB of RAM, running Windows XP and reading and writing the data to a remote server. This took 7 h using the `lowMemory` option.

S4 methods for generic R functions such as `summary()`, `plot()` and `boxplot()` make it easy to query the SNPMaP object and visualize the data it contains. Accessors provide convenient access to the data. All functions are documented through the R help system. For example, typing `?snpmap` will bring up a page describing the `snpmap()` function and its usage. Similarly, `package?SNPMaP` and `class?SNPMaP` will bring up help pages for the SNPMaP package and the SNPMaP class, respectively.

Although the SNPMaP object is intended to be useable for further analyses, the data can also easily be extracted to a matrix using

```
mras.matrix <- as.matrix (mras)
```

A user who wants CEL files transformed into a spreadsheet of RAS in the simplest possible way need not use R interactively at all; example scripts that can be invoked from various shells are available from the web site, including a point-and-click front end for Windows. These steps comprise the simplest route from CEL files to the RAS used for association analysis.

On the other hand, a user who wants to examine all steps of the analysis and experiment with new methods has access to the data in a straightforward and convenient form. This flexibility is one of the major strengths of an implementation in R because of the impressive array of cutting-edge statistical techniques already available in the R environment.

A more involved approach might begin by extracting the raw probe intensities from the CEL files (running the workflow function `cel2raw` rather than the default `cel2rasS`):

```
raw <- snpmap (RUN='cel2raw')
```

This allows the user to plot the raw probe intensities and generate pseudoimages of the processed chips using the `image()` method, so the user can check for scanning artifacts, such as dust or fingerprints. The raw intensities can then be further processed to individual probe quartet RAS by a workflow function:

```
ras <- raw2ras (raw)
```

Other options available at the `snpmap()` or workflow stage include `normalize`, which quantile normalizes the raw probe intensities across chips; `log.intensities`, which causes SNPMaP to use the natural logarithm of the probe intensities; and `useMM`, which causes SNPMaP to subtract mismatch probe intensities (where available) before calculating RAS.

To calculate RAS, the package uses the method most commonly described in the literature; that is dividing the (possibly modified) intensity of allele A by the sum of the intensities of alleles A and B. If summary RAS scores for each array are required, this is calculated by a user-defined function (defaulting to the mean) of the RAS scores for each pair of probes. However, there are other ways of calculating summary RAS, such as taking the mean of allele A intensities across an array and dividing that by the summed means of allele A and B. Should users wish to use this alternative method of calculating RAS scores, a modified version of the SNPMaP package is available from the authors.

On recent Affymetrix arrays mismatch probes have been discarded in favor of greater perfect match probe density, so that mismatch intensities can no longer be subtracted before calculating RAS. Although this means that RAS scores based on the new arrays are no longer necessarily a good estimate of the absolute allele frequency in the pool, the critical comparison for an association analysis is the difference in allele frequencies between case and control pools. For this reason, the loss of the mismatch probes has relatively little effect on the resulting association analysis (Pearson *et al.*, 2007). For a discussion of the value of mismatch probes, see Millenaar *et al.* (2006).

Some authors have discussed the possibility of correcting pooled SNP assays for differential hybridization of the allele-specific probes, a process known as *k*-correction (Le Hellard *et al.*, 2002; Simpson *et al.*, 2005). The intention here is to produce more accurate estimates of absolute allele frequency in the pools. However, others have noted that again, since the critical comparison is between cases and controls rather than between absolute allele frequencies and a reference population, such correction has little effect on the results of the analysis (Macgregor *et al.*, 2006). Nevertheless, because of the way SNPMaP is implemented, for those users with access to a large reference sample of individuals typed on individual arrays, it is straightforward to apply the correction.

3 DISCUSSION

The SNPMaP package represents an evolution of R scripts that have been used by us (e.g. Meaburn *et al.*, 2006, 2008) and others (e.g. Wilkening *et al.*, 2007) for processing SNPMaP data. Although a few other software applications are available, such as GenePool (Pearson *et al.*, 2007) or MPDA (Yang *et al.*, 2008), the strength of the SNPMaP package is in making the data readily available in the rich R environment, allowing easy access to the early stages of the analysis, effective visualization using R's powerful graphics system and great flexibility in constructing association tests, comparing methods and making modifications (Barratt *et al.*, 2002; Sham *et al.*, 2002): a hands-on, rather than a hands-off approach. The package currently supports the most recent Affymetrix arrays: both Mapping250K arrays and the 5.0 and 6.0 GenomeWideSNP arrays. Support for other arrays will be added in future versions. We also have plans for a supporting package aimed at providing implementations of common association analyses described in the literature, along with some novel methods and further visualization tools. We have kept the data structures straightforward to facilitate further development of analysis methods by users; although the current version is specifically aimed at SNPMaP using Affymetrix SNP arrays, it is readily extended for other platforms. For example, our tests have included using the package to normalize the raw intensities across

265 Affymetrix Mouse Exon (gene expression) arrays. Above all, the SNPMaP package affords users the flexibility to make best use of the sophisticated tools already available in the R environment to analyze their SNP microarrays and pooling studies.

ACKNOWLEDGEMENTS

Thanks to Henrik Bengtsson, James Bullard and Kasper Daniel Hansen for the affxparser and R.huge packages.

Funding: The Wellcome Trust (GR75492); the US National Institute of Child Health and Human Development (HD49861); the UK Medical Research Council (G9424799, G0500079).

Conflict of Interest: none declared.

REFERENCES

Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781–791.

Barratt,B.J. *et al.* (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.,* **66**, 393–405.

Barrett,J.C. and Cardon,L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.

Butcher,L.M. *et al*. (2004) Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav. Genet.*, **34**, 549–55.

Butcher,L.M. *et al*. (2008) Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays. *Genes Brain Behav.*, **7**, 435–446.

Kirov,G. *et al*. (2006) Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics*, **7**, 27.

Le Hellard,S. *et al*. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res.*, **30**, e74.

Macgregor,S. *et al*. (2006) Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. *Nucleic Acids Res.*, **34**, e55.

Millenaar,F.F. *et al*. (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, **7**, 137.

McCarthy,M.I. *et al*. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Meaburn,E. *et al*. (2006) Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res.*, **34**, e28.

Meaburn,E.L. *et al*. (2008) Quantitative trait locus association scan of early reading disability and ability using pooled DNA and 100K SNP microarrays in a sample of 5760 children. *Mol. Psych.*, **13**, 729–740.

Pearson,J.V. *et al*. (2007) Identification of the genetic basis for complex disorders by use of pooling-based genome-wide single-nucleotide-polymorphism association studies. *Am. J. Hum. Genet.*, **80**, 126–139.

Sham,P. *et al*. (2002) DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.*, **3**, 862–871.

Simpson,C.L. *et al.* (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res.*, **33**, e25.

Slatkin,M. (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Wilkening,S. *et al*. (2007) Allelotyping of pooled DNA with 250K SNP microarrays. *BMC Genomics*, **8**, 77.

Yang,H.C. *et al*. (2008) MPDA: Microarray pooled DNA analyzer. *BMC Bioinformatics*, **9**, 196.