Flapjack—graphical genotype visualization

Iain Milne1, Paul Shaw1, Gordon Stephen1, Micha Bayer1, Linda Cardle1, William T. B. Thomas1, Andrew J. Flavell2 and David Marshall1,∗

1Genetics Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA and 2Division of Plant Sciences, University of Dundee at S CRI, Invergowrie, Dundee DD2 5DA, UK

ABSTRACT

Summary: New software tools for graphical genotyping are required that can routinely handle the large data volumes generated by the high-throughput single-nucleotide polymorphism (SNP) platforms, genotyping-by-sequencing and other comparable genotyping technologies. Flapjack has been developed to facilitate analysis of these data, providing real time rendering with rapid navigation and comparisons between lines, markers and chromosomes, with visualization, sorting and querying based on associated data, such as phenotypes, quantitative trait loci or other mappable features.

Availability: Flapjack is freely available for Microsoft Windows, Mac OS X, Linux and Solaris, and can be downloaded from http://bioinf.scri.ac.uk/flapjack

Contact: flapjack@scri.ac.uk

Received on September 14, 2010; revised on October 8, 2010; accepted on October 8, 2010

1 INTRODUCTION

The concept of a graphical genotype to visualize haplotype diversity between chromosomes has been widely adopted since Young and Tanksley (1989) used it in the context of restriction fragment length polymorphism (RFLP) mapping populations. Existing software tools to display graphical genotypes include GGT (van Berlo, 2008) and Geneflow (geneflowinc.com). The advent of new high-throughput genotyping technologies have given a renewed stimulus to the concept of graphical genotyping, through a combination of dramatic reduction in cost per data point and vastly increased marker density and throughput. The resultant high-density data underpin new genetic approaches such as genome-wide association analysis (Rostoks et al., 2006). It also leads to the possibility of visually comparing many lines (e.g. samples or individuals) or sorting and selecting based on phenotype, identified groupings and genome features, such as quantitative trait loci (QTL) or gene models mapped to the genetic or physical genome. However, the ability to generate datasets with many thousands of markers (McMullen et al., 2009) on many thousands of lines imposes a significant demand on both software tools and the underlying computer hardware. Flapjack provides a high performance visual interface into graphical genotyping applications in genetics and plant breeding.

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
of markers or to export sections of data into their own custom view. Selections can be made either manually or by markers under a QTL. Flapjack allows the user to create any number of these custom views, each containing its own set of lines, markers, ordering, colour schemes, bookmarked locations and so on.

The data for a given view—either in graphical or in its underlying raw format—can be exported back to disk. Images can be produced and saved in PNG format for the current view, or the on-screen subsections of a view, and the user can select whether to include all components (allele data, chromosome maps, line names, traits, etc.) or pick and choose only the ones of interest. When exporting the underlying data, similar options are available to export the entire dataset or to only include data from specific chromosomes or the currently selected lines and markers. The data are saved in tab-delimited plain text files identical in format to the files original imported into Flapjack.

Although completely standalone, with data imported via simple plain-text formats, integration with external data sources such as Germinate (bioinf.scri.ac.uk/germinate) is also possible. This provides easy selection and export of data directly into Flapjack, along with web-links back to the line and marker data in the original database. This feature has been designed to work with any external data source, by means of supplying Flapjack with a custom URL that can be queried with key/value pairs.

Flapjack projects are persistent, with all data, views, user selections and so on being saved to either an XML-based file or an experimental binary format more suited to very large datasets. The XML and text formats are documented on our web site, and are also currently supported by iMAS (icrisat.org/bi-biomatrics-imas.htm), QU-GENE (Podlich and Cooper, 1998), Gramene (Liang et al., 2008), Genstat (vsni.co.uk/software/genstat) and The Hordeum Toolbox (hordetntontoolbox.org). Projects can also be created using a command-line utility, which provides a convenient integration with custom analysis pipelines and databases.

3 IMPLEMENTATION

Flapjack is written in Java and is compatible with any system running Java 1.6 or higher. For convenience, we provide installable versions with everything required to run the application, including a suitable Java run-time. These are available for Windows, Mac OS X, Linux and Solaris. Flapjack regularly monitors our server for new versions and will prompt, download and update quickly and easily when a new release is available. The code is internationalized and is distributed with translations in English (UK/US) and German.

The code can take advantage of multicore processors, a feature especially significant for the rendering code, which—among its other optimizations—is capable of simultaneous rendering across all cores, greatly improving the end-user experience when navigating around large or complex datasets. We have designed Flapjack to be very memory efficient, and are confident that it can comfortably handle datasets with hundreds of millions of alleles even on a machine with just 1 GB of main memory.

4 FUTURE WORK

Future development with Flapjack will entail enhancing its visualizations to provide better support for very small datasets, primarily by enabling the display of all markers across the genome in a single view. We want to extend support for rendering features beyond QTL to include more generic features, such as gene models for SNP data anchored to physical maps, and to provide a graph track to display summary information such as PIC values or test statistics. We are also working with academic and breeding company partners to explore supporting additional data formats such as HapMap and PLINK, and on closer integration with Germinate, by allowing its databases to be automatically populated by the data imported into Flapjack.

ACKNOWLEDGEMENTS

We would like to thank colleagues within the Genetics and BioSS Programmes at SCRI for their input to this project.

Funding: Scottish Government (RERAD, Programme 1); Scottish Funding Council; Scottish Enterprise through the Scottish Bioinformatics Research Network (SBRN) project.

Conflict of Interest: none declared.

REFERENCES

Liang,C. et al. (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res, 36, D947-D953.
McMullen,M.D. et al. (2009) Genetic properties of the maize nested association mapping population. Science, 325, 737-740.
Podlich,D.W. and Cooper,M. (1998) QU-GENE: a platform for quantitative analysis of genetic models. Bioinformatics, 14, 632-653.
Rostoks,N. et al. (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl Acad. Sci. USA, 103, 18656–18661.
van Berloo,R. (2008) GGT 2.0: versatile software for visualization and analysis of genetic data. J Hered., 99, 232-236.
Young,N.D. and Tanksley,S.D. (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor. Appl. Genet., 77, 101.