Graphical applications for visualization and analysis of genotype data sets

Iain Milne[1], Gordon Stephen[1], Micha Bayer[1], Paul D. Shaw[1], Sebastian Raubach[1], Sarah Hearne[2], Sukhwinder Singh[2], Peter Wenzl[2] and David Marshall[1]

1. Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, Scotland
2. International Maize and Wheat Improvement Center (CIMMYT), Mexico

bioinf@hutton.ac.uk

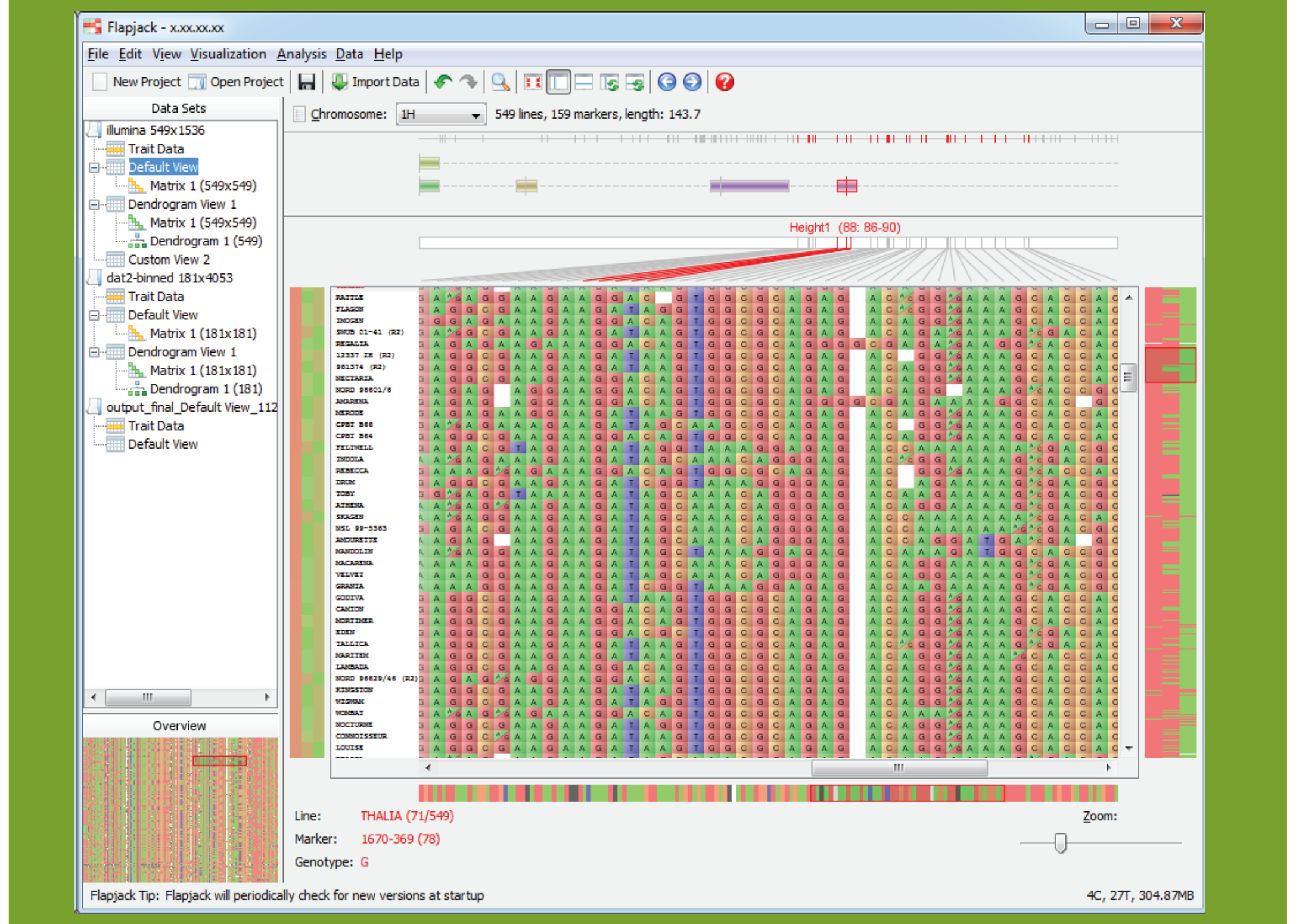
Our applications are available for multiple desktop platforms (Windows/OS X/Linux). You can download 32 and 64-bit versions from http://bioinf.hutton.ac.uk.

Visualization

Plant geneticists require a new generation of software tools to interrogate the increasingly high volume data sets that are being generated by high-throughput SNP platforms, genotyping-by-sequencing and other comparable genotyping technologies.

At the James Hutton Institute and within the MasAgro Seeds of Discovery project (**http://www.seedsofdiscovery.org**) we find that we are increasingly using data visualization to support and complement our analysis through the applications that we have developed.

These include Flapjack to visualize graphical genotypes from extensive SNP and GBS data sets, and CurlyWhirly to visualize spatial xyz plots such as those obtained from principal coordinates analysis.



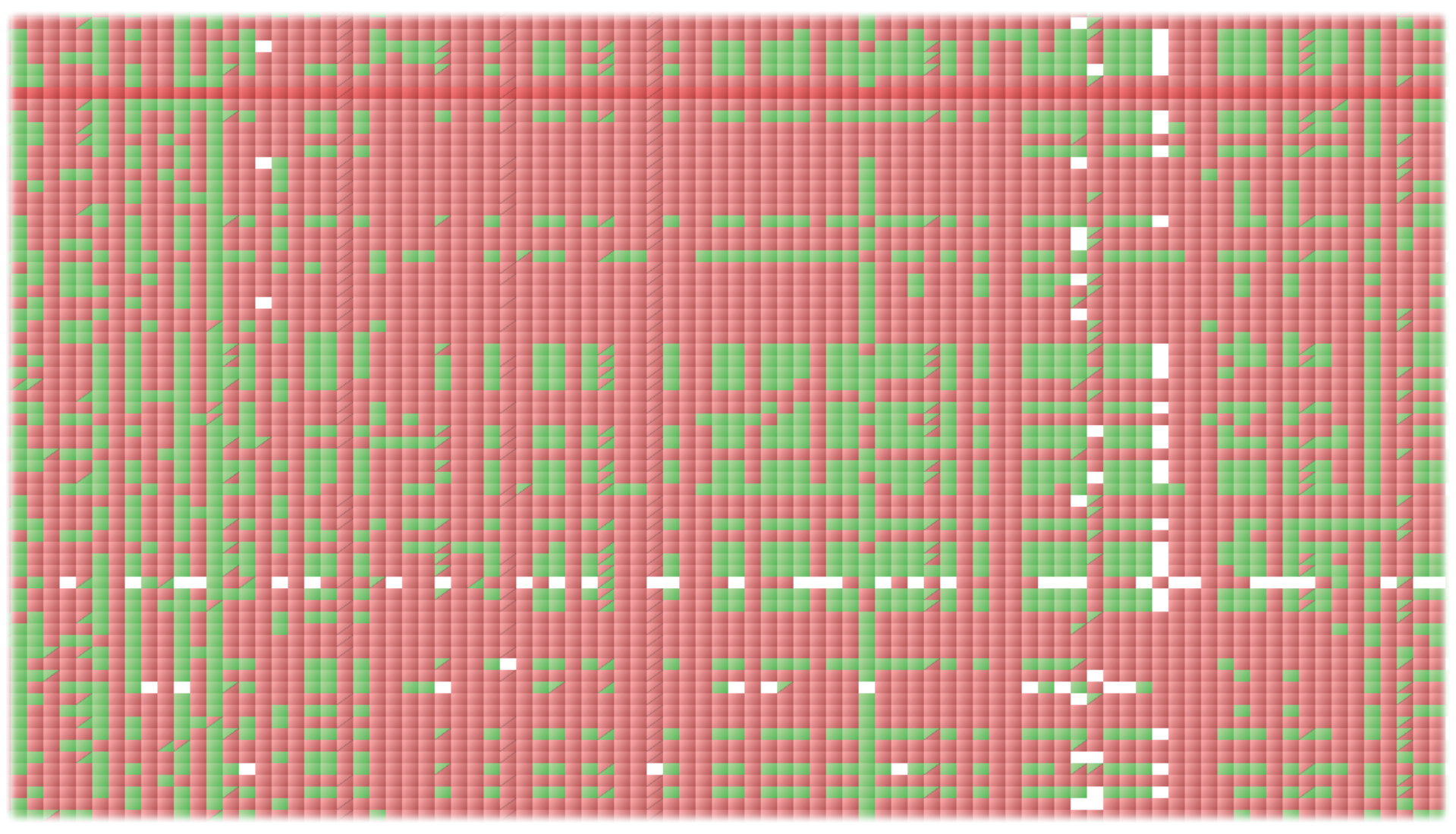
Although standalone, these applications also seamlessly integrate with a wider suite of tools forming part of this project (for example Germinate that handles and presents the storage of genetic resources and experimental data) so that information can be easily passed from database to analysis to visualization and back again.

Acknowledgements

This work is funded through the MasAgro initiative of SAGARPA and Strategic Initiative 8 of the maize CGIAR research program, and is also supported by the Scottish Government (RESAS, Theme 5).

We would also like to thank colleagues within the Cell & Molecular Sciences Group and Biomathematics & Statistics Scotland at The James Hutton Institute for their input to these applications.

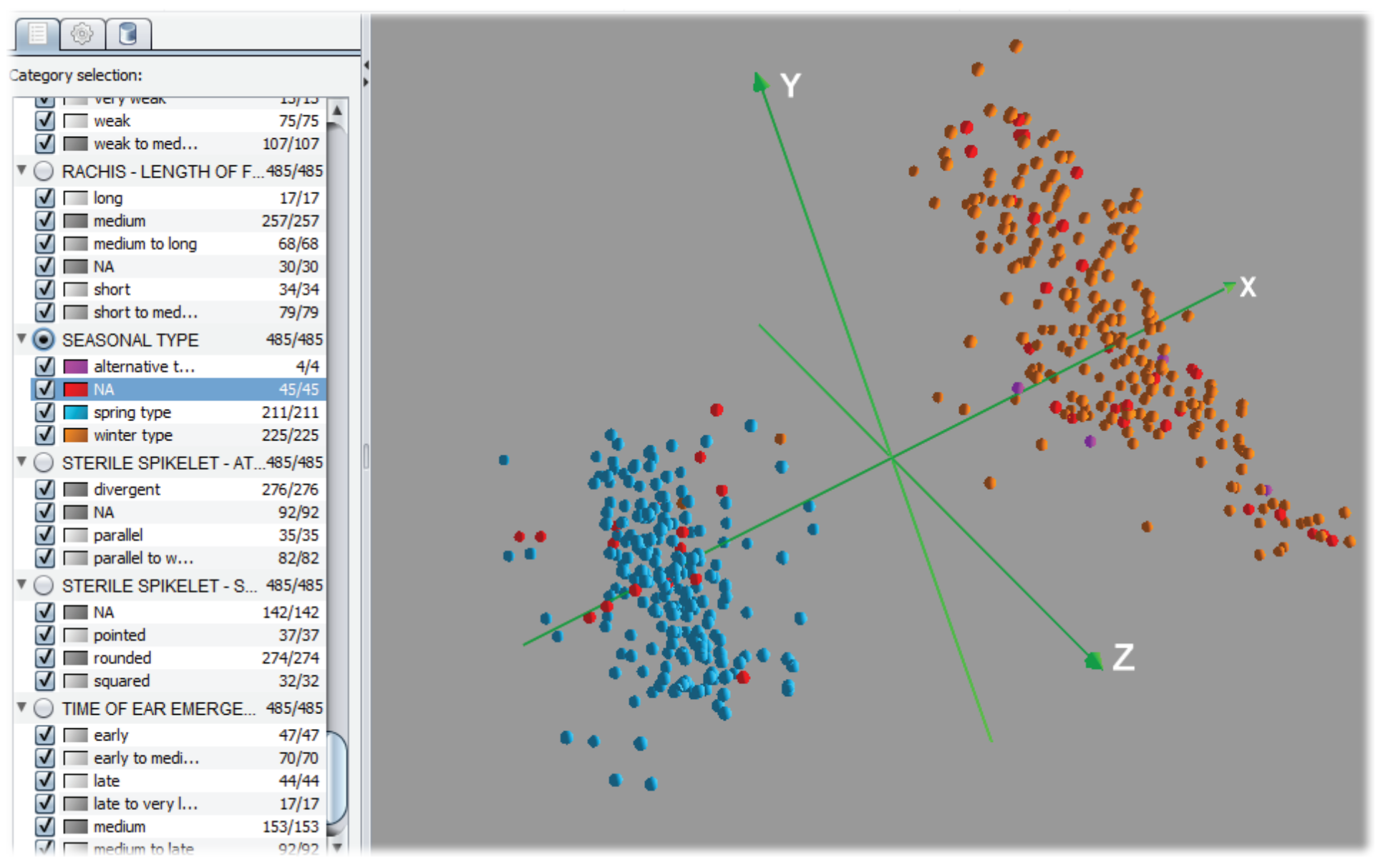
Flapjack



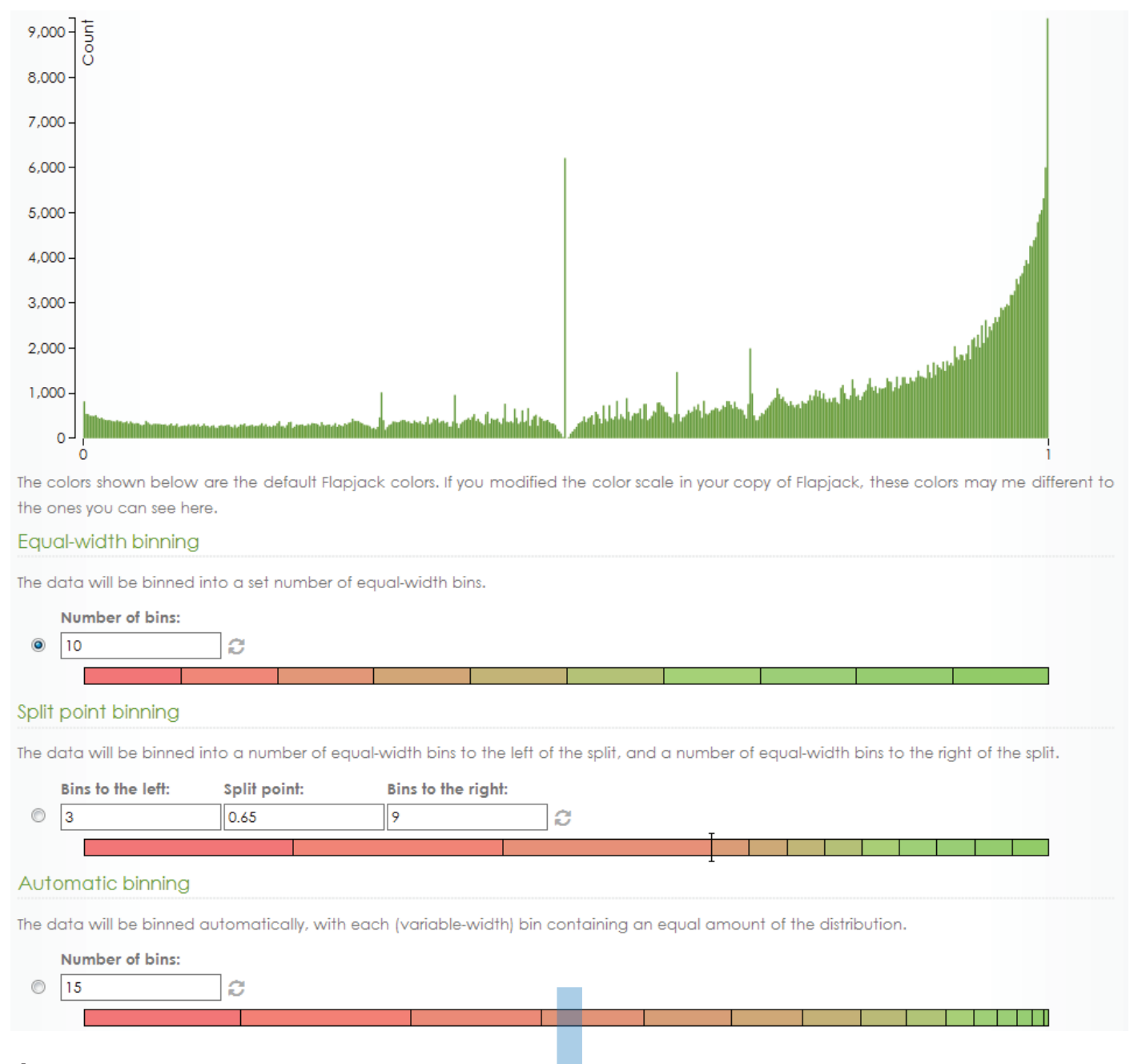
Flapjack's core visualization displays a graphical genotype of alleles for multiple germplasm accessions/cultivars, aligned against a map of markers displayed in linkage group order. It provides a number of alternative views with individual alleles coloured by state, frequency or similarity to a given reference line (as shown). This displays the selected line in a darker red, then colours each allele on every other line based on whether the allelic states are the same (red) or differ (green) from the comparison line.

CurlyWhirly

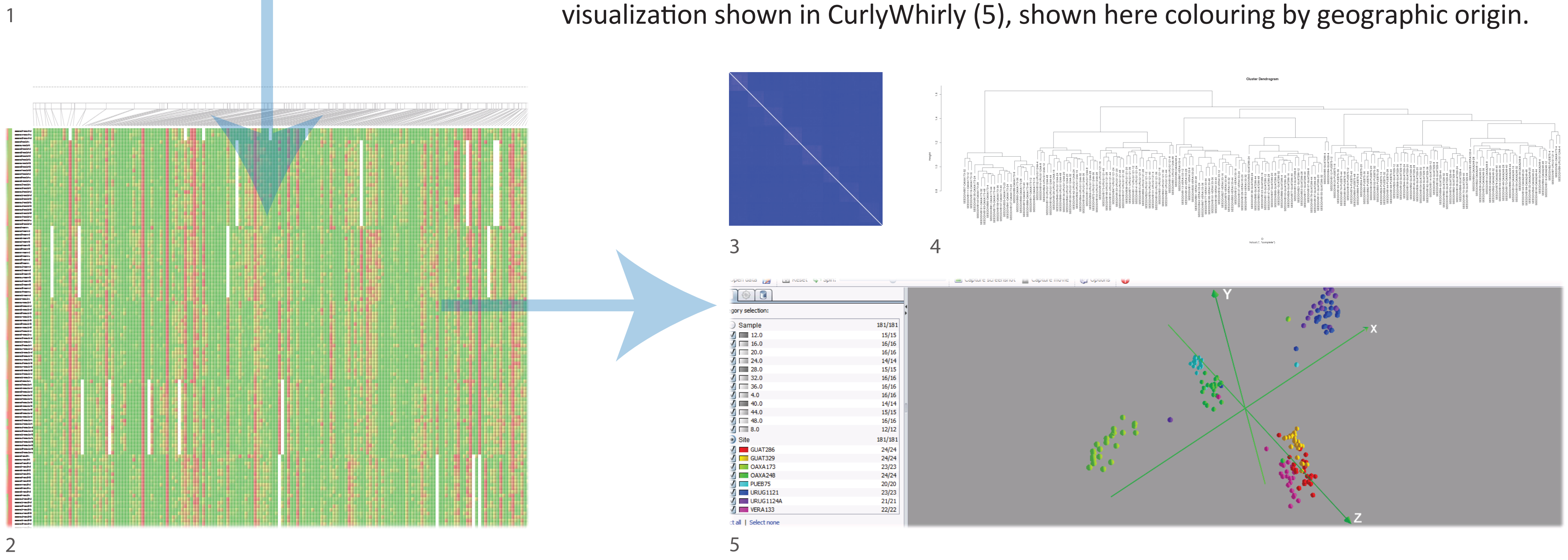
CurlyWhirly provides data visualization in a 3D context, including but not limited to the output from Principal Coordinate Analysis and Principal Components Analysis. Intuitive controls allow the data to be filtered or highlighted using categorical data such as from phenotypes, whilst its efficient memory usage and high-performance allows for real-time interactivity with very large data sets. This functionality enables exploration of multidimensional data in such a way that facilitates finding patterns and outliers within the data.



Exemplar allele-frequency analysis pipeline in Maize



In this example, we demonstrate the interaction available between our tools, which, although designed to be independent, are also loosely coupled for the purposes of basic data exchange. Starting with Germinate (1), here we see summary statistics on stored maize allele-frequency data presented as a histogram. The user can choose one of several binning schemes to apply to the underlying data so that it can be passed to Flapjack (2) for further visualization and exploration. URL mapping embedded in the file allows Flapjack to automatically link the user back to Germinate's web interface for a more detailed look at any given line or marker. A distance matrix can then be built (3) that is used (via R web services) to perform a cluster analysis, leading to the display of a dendrogram (4) whose order can be applied back to Flapjack's main display. The matrix also forms the basis of a PCoA analysis, ultimately leading to the 3D visualization shown in CurlyWhirly (5), shown here colouring by geographic origin.



Conclusions

- Data visualization is important for quality control and for exploring patterns in high-throughput genotype data.
- Our tools provide user-friendly management of large data sets through search and filtering functions and a multitude of navigation modes.
- They are designed for easy installation with no complex software dependencies, and provide intuitive, geneticist-friendly graphical user interfaces.