

Visualizing Genetic Transmission Patterns in Plant Pedigrees

Paul D. Shaw^{**†‡}
The James Hutton
Institute

Jessie Kennedy[‡]
Edinburgh Napier
University

Martin Graham[‡]
Edinburgh Napier
University

Iain Milne[†]
The James Hutton
Institute

David F. Marshall[†]
The James Hutton
Institute

1 INTRODUCTION

The use of pedigrees in plant breeding allows breeders to target specific crosses in order to maximise the potential of achieving desirable agriculturally important characteristics such as yield and disease resistance. The ability to predict inheritance of genes that facilitate resistance to pathogens or other commercially important characteristics is of benefit to experimental plant genetics and commercial plant breeding programmes. Derivation of this information by traditional molecular techniques is expensive and time consuming, especially in industrial settings where, due to time constraints relating to growing seasons, many thousands of plant lines may need to be screened quickly and efficiently every year.

Statistical analysis can be used to predict transmission of genes to progeny. There is however a limitation trying to conceptualise large pedigree structures. However, a combination of visual and statistical analysis would allow researchers and breeders to gain a better understanding of the transmission of genetic elements within a pedigree based framework. By developing a pedigree based framework for visualization of genetic information we hope to uncover patterns and associations when combined with specialist domain knowledge.

Using a unique and extensive pedigree, genotypic and phenotypic barley dataset covering UK elite germplasm derived from genome wide association studies (GWAS), we discuss the challenges of visualizing the transmission of alleles encoding traits and characteristics of agricultural importance in a pedigree based framework. This will allow predictions relating to patterns of transmission of genes or alleles through breeding programmes and present a prototype pedigree visualization tool and initial feedback from users.

The availability of the unique and real datasets gathered over the last 20 years should allow us to create a prototype system and make predictions based on this for lines for which we do not have data. Through focussing on characters such as *Mlo* and *Mla* resistance, rachilla hair length or aleurone colour which can be determined either through visual observation or molecular testing, we can accurately test our predictions and hypothesis. In addition we can compare results that we gain to those obtained through non-visual statistical methods both in terms of analysis time and analysis quality, both of which can be accurately quantified.

2 PLANT PEDIGREES

A pedigree (see Figure 1) is a representation of the genetic relationship through each sexual generation showing how genetically discrete individuals are related in time to one another.

Pedigrees are often used in human contexts to show the transmission of alleles responsible for genetic conditions of medical importance. In plants they are often used as a framework on which statistical analysis can be used to determine factors such as mode of inheritance or to check for potential genotyping errors.

While there are standard nomenclatures for pedigrees in humans [1], [2] there is no single formal system for plant pedigrees due to the hermaphrodite nature of most plant species, the complexity of mating designs possible in plant genetics and the absence of any coordinating organisation.

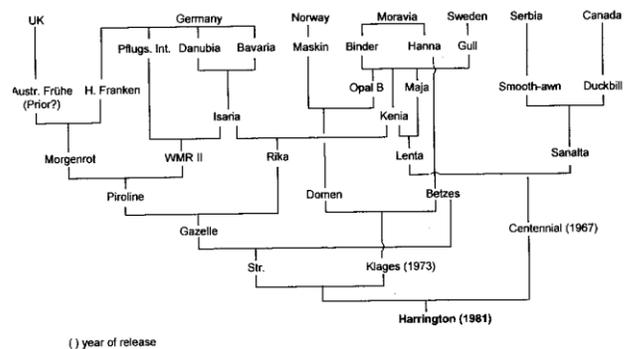


Figure 1: Traditional barley pedigree showing Elite cultivars [3].

3 EXPERIMENTAL DATASETS

Our experimental data centers on barley pedigree data for 600 UK Elite cultivars as well as Single Nucleotide Polymorphism (SNP) genotypic data for these lines across approximately 3,000 genetic markers. In addition we have phenotypic data for these lines for around 33 characters across multiple years and sites (1980 – present which is around 500,000 data points). We will however focus only on a small subset of these relating to disease resistance for this study. This is a large and unique dataset and of significant agricultural importance.

The ability to display multivariate data is a key component of the pedigree visualizations that we implement. This takes the form of phenotypic, genotypic and similarity data. The ability to combine this information will be important in any software prototypes. Independent variables from empirical datasets will be supplemented with derived data from similarity calculations between lines.

The similarity of lines can be described in two ways. Firstly, at the phenotypic level whereby two individuals are phenotypically identical based on some morphological or physiological trait.

* e-mail: paul.shaw@hutton.ac.uk

† Genetics Programme, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK

‡ School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, Scotland, UK

Secondly, they may be genetically identical at the DNA level; this may lead to the same expressed characteristics.

The breadth of data in our datasets is limited to the number of new varieties which are produced in a given year. The depth is constrained by time.

4 PREVIOUS WORK

Pedigree visualization has primarily been carried out by work in the human genetics domain. As plant breeding programmes involve phenomena such as routine inbreeding there are additional visualization challenges that need to be addressed. In addition, there are often a larger number of plant lines in the average pedigree. This is down to factors such as generation time which is relatively low in most crop species. Current solutions do not address all of these differences [4].

5 PROTOTYPE SYSTEM

We developed a prototype visualization system using GraphViz to explore issues of visualizing our exemplar data sets (Figures 2,3). Early tests using a directed acyclic graph (DAG) based approach on exemplar datasets have already proved valuable to our target users. Taking the UK Elite Barley pedigree we were able to test a prototype visualization system utilising both colour and size of node to represent information (see figures 2,3). When presented with our results breeders told us that it gave them an overview of their data that was not currently available to them, indeed these representations uncovered interesting information relating to the relative frequency of use of particular “key” lines in the UK Elite Barley germplasm collections, something which while known in terms of written records was never shown in a way where users could see it explicitly.



Figure 2: A. Early experimentation with node size and colour to display a simple plant pedigree. Colour was used to distinguish between the winter/spring phenotypic grouping, and node size and weighted edges to represent the number of derived progeny. While techniques such as these have been seen in other applications they tend to offer either colour or node size features. Expanded views such as this give an overall picture and not implicit detail.

Techniques such as Dorling cartograms are an example of a technique which preserves both position and shows additional data types. While our data does not have a geographical location it does have a position in time in a pedigree making a hybrid of these charts and DAGs a potential solution for our problem. Additionally the use of flow mapping techniques to show flow of alleles from parent to offspring may prove to be valuable.

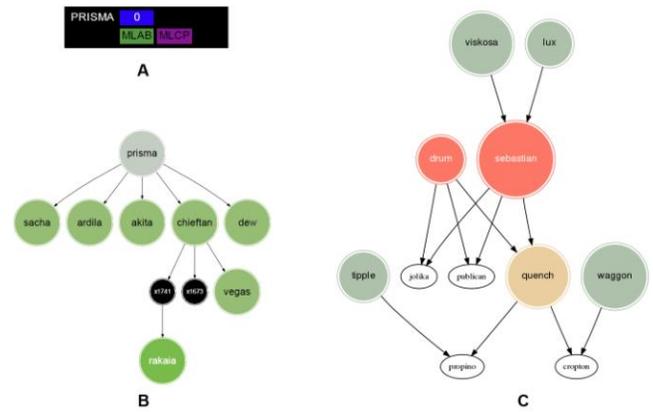


Figure 3: Early prototype system. A. Alternate node representation showing increased levels of detail for each line. B. Use of colour gradient to show genetic similarity between lines. C. While figure 2 shows an overall picture users can select more detail by selecting sub-graphs.

6 CONCLUSION

We have identified the challenges of visualizing plant pedigree datasets and initial visualization experiments show the potential techniques which will be valuable in developing an exploratory analytical tool for plant breeders. Our current prototype system lacks interactive features which would allow us to further explore our data. We propose, in cooperation with end users, to refine how our data is displayed and implement a more interactive visual exploration system. This will allow us to add and remove data from our pedigree scaffold and try to identify novel patterns in the data or errors which will allow better decisions to be made in breeding programmes.

7 ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from the Scottish Government and Edinburgh Napier University.

We would also like to thank colleagues at The James Hutton Institute, in particular Bill Thomas for providing pedigree data and advice, Linda Milne, Gordon Stephen, Jordi Comadran, Luke Ramsay and Robbie Waugh for their feedback and colleagues from NIAB and the AGOUEB consortium for the use of phenotypic and genotypic data relating to our test datasets.

REFERENCES

- [1] R. L. Bennett et al., “Recommendations for standardized human pedigree nomenclature. Pedigree Standardization Task Force of the National Society of Genetic Counselors,” *American Journal of Human Genetics*, vol. 56, no. 3, pp. 745-752, Mar. 1995.
- [2] R. L. Bennett, K. S. French, R. G. Resta, and D. L. Doyle, “Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors,” *Journal of Genetic Counseling*, vol. 17, no. 5, pp. 424-433, Oct. 2008.
- [3] G. Fischbeck, “Chapter 3 Diversification through breeding,” in *Diversity in Barley - Hordeum vulgare*, vol. 7, Elsevier, 2003, pp. 29-52.
- [4] P. D. Shaw, J. Kennedy, M. Graham, I. Milne, and D. F. Marshall, “Visualizing Genetic Transmission Patterns in Plant Pedigrees.” [Online]. Available: http://bioinf.hutton.ac.uk/paul/technical_documents/20110429_pedigree_visualization.pdf.