Visualizing Genetic Transmission Patterns in Plant Pedigrees

Paul David Shaw

A thesis submitted in partial fulfilment of the requirements of Edinburgh Napier University, for the award of Doctor of Philosophy

February 2016

Abstract

Ensuring food security in a world with an increasing population and demand on natural resources is becoming ever more pertinent. Plant breeders are using an increasingly diverse range of data types such as phenotypic and genotypic data to identify plant lines with desirable characteristics suitable to be taken forward in plant breeding programmes. These characteristics include a number of key morphological and physiological traits, such as disease resistance and yield that need to be maintained and improved upon if a commercial plant variety is to be successful.

The ability to predict and understand the inheritance of alleles that facilitate resistance to pathogens or any other commercially important characteristic is crucially important to experimental plant genetics and commercial plant breeding programmes. However, derivation of the inheritance of such traits by traditional molecular techniques is expensive and time consuming, even with recent developments in high-throughput technologies. This is especially true in industrial settings where, due to time constraints relating to growing seasons, many thousands of plant lines may need to be screened quickly, efficiently and economically every year. Thus, computational tools that provide the ability to integrate and visualize diverse data types with an associated plant pedigree structure will enable breeders to make more informed and subsequently better decisions on the plant lines that are used in crossings. This will help meet both the demands for increased yield and production and adaptation to climate change.

Traditional family tree style layouts are commonly used and simple to understand but are unsuitable for the data densities that are now commonplace in large breeding programmes. The size and complexity of plant pedigrees means that there is a cognitive limitation in conceptualising large plant pedigree structures, therefore novel techniques and tools are required by geneticists and plant breeders to improve pedigree comprehension.

Taking a user-centred, iterative approach to design, a pedigree visualization system was developed for exploring a large and unique set of experimental barley (*H. vulgare*) data. This work progressed from the development of a static pedigree visualization to interactive prototypes and finally the Helium pedigree visualization software. At each stage of the development process, user feedback in the form of informal and more structured user evaluation from domain experts guided the development lifecycle with users' concerns addressed and additional functionality added.

Plant pedigrees are very different to those from humans and farmed animals and consequently the development of the pedigree visualizations described in this work focussed on implementing currently accepted techniques used in pedigree visualization and adapting them to meet the specific demands of plant pedigrees. Helium includes techniques to aid problems with user understanding identified through user testing; examples of these include difficulties where crosses between varieties are situated in different regions of the pedigree layout. There are good biological reasons why this happens but it has been shown, through testing, that it leads to problems with users' comprehension of the relatedness of individuals in the pedigree. The inclusion of visual cues and the use of localised layouts have allowed complications like these to be reduced. Other examples include the use of sizing of nodes to show the frequency of usage of specific plant lines which have been shown to act as positional reference points to users, and subsequently bringing a secondary level of structure to the pedigree layout. The use of these novel techniques has allowed the classification of three main types of plant line, which have been coined: *principal*, *flanking* and *terminal* plant lines. This technique has also shown visually the most frequently used plant lines, which while previously known in text records, were never quantified.

Helium's main contributions are two-fold. Firstly it has applied visualization techniques used in traditional pedigrees and applied them to the domain of plant pedigrees; this has addressed problems with handling large experimental plant pedigrees. The scale, complexity and diversity of data and the number of plant lines that Helium can handle exceed other currently available plant pedigree visualization tools. These techniques (including layout, phenotypic and genotypic encoding) have been improved to deal with the differences that exist between human/mammalian pedigrees which take account of problems such as the complexity of crosses and routine inbreeding. Secondly, the verification of the effectiveness of the visualizations has been demonstrated by performing user testing on a group of 28 domain experts. The improvements have advanced both user understanding of pedigrees and allowed a much greater density and scale of data to be visualized. User testing has shown that the implementation and extensions to visualization techniques has improved user comprehension of plant pedigrees when asked to perform real-life tasks with barley datasets. Results have shown an increase in correct responses between the prototype interface and Helium. A SUS analysis has sown a high acceptance rate for Helium.

Table of contents

List of tables

List of figures

Acknowledgements

I would like to start by thanking my supervisors Jessie Kennedy, Martin Graham and David Marshall for their support throughout this work and to Edinburgh Napier University for funding a large proportion of my fees. I would also like to thank my employer, The James Hutton Institute (Formerly the Scottish Crop Research Institute) and the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS) for their continued support allowing me to take on this research part-time while continuing with my day job.

This research was carried out while working within the Genetics, and more recently Information and Computational Sciences group, at The James Hutton Institute in Dundee. Because of this I thank my work colleagues for having to put up with me over the duration of this work and perhaps having to wait a little longer than usual to get things done!

Pedigree data used in this work was obtained from a number of sources including Science and Advice for Scottish Agriculture (SASA) in Edinburgh, the Home Grown Cereals Authority (HGCA) and a number of commercial companies. I would also like to express my appreciation to Bill Thomas for his tireless work in spotting errors in data and his enthusiasm for this work.

In no particular order I thank my colleagues: Luke Ramsay, Robbie Waugh, Bill Thomas, Joanne Russell, Iain Milne, Linda Milne, Jordi Comadran (Limagrain), Runxuan Zhang, Sebastian Raubach and Helena Oakey (University of Dundee) for their encouragement over the last few years, being willing test subjects for both ideas and software testing, and for the knowledge from their diverse fields that each has imparted.

Lastly I would like to thank all those who took part in the user testing for this work especially those at Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) in Mexico and from Cell and Molecular Sciences at The James Hutton Institute who,

without complaint and infections enthusiasm, gave up so much of their time to take part in the testing of Helium.

Institute for Informatics & Digital Innovation
Part of Edinburgh Napier University

The James Hutton Institute

The Scottish Government
Riaghaltas na h-Alba

Dedication

To Fiona who has suffered and our recently born first child Ailsa who has avoided the stress and strains of this research.

Contributing Publications

Refereed Publications

Shaw, P., Graham, M., Kennedy, J., Milne, I., & Marshall, D. 2014. Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics*, *15*(1), 259. doi:10.1186/1471-2105-15-259*

**This paper won the best paper award at BioVis 4th Symposium on Biological Data Visualization (Special Interest Group of ISMB) 11-12 July 2014, Boston, USA.*

Posters and Conference Abstracts (Peer Reviewed)

Shaw, P.D, Kennedy, J., Graham, M., Milne, I. and Marshall, D.F. 2014. Helium: Visualization of Large Plant Pedigrees. *Monogram Network Meeting*, Reading, UK, 27-28 March 2014.

Milne, I., Stephen, G., Bayer, M., Shaw, P., Raubach, S., Hearne, S., Singh, S., Wenzl, P., Marshall, D. 2014. Graphical Applications for Visualizing and Analysing Genotype Data Sets. *Plant and Animal Genome XXII*, San Diego, CA, 11-15 January 2014

Shaw, P.D., Kennedy, J., Graham, M., Marshall, D. 2013. Evaluation of Helium: Visualization of Large Scale Plant Pedigrees. *BioVis 2013 3rd IEEE Symposium on Biological Data Visualization*. Atlanta, GA, 13-14 October 2013.

Shaw, P., Thomas, B., Ramsay, L., Waugh, R., Comadran, J., Stephen, G., Milne, I., Graham, M., Kennedy, J., Marshall, D. 2013. Visualizing Genetic Transmission Patterns in Plant Pedigrees. *Plant & Animal Genome XXI*, San Diego, CA, 12-16 January 2013

Shaw, P.D., Kennedy, J., Graham, M., Milne, I., Marshall, D.F. 2012. Using a Mathematical Graph Framework for Visualization of Inheritance Patterns in Commercial Plant Pedigrees. *BioVis 2012 2nd IEEE Symposium on Biological Data Visualization*. Seattle, WA, 14 – 15 October 2012.

Shaw, P., Milne, I., Cardle, L., Waugh, R., Thomas, W.T.B., Ramsay, L., Comadran, J., Stephen, G., Bayer, M., Graham, M., Kennedy, J., Oakey, H., Marshall, D. 2012. Visualizing Genetic Transmission Patterns in Plant Pedigrees. *Plant & Animal Genome XX*, San Deigo, CA, 14-18 January 2012

Shaw, P.D., Kennedy, J., Graham, M., Milne, I., Marshall, D.F. 2011.Visualizing Genetic Transmission Patterns in Plant Pedigrees. *BioVis 2011 1st IEEE Symposium on Biological Data Visualization*. Providence, RI, 23 – 24 October 2011.

Posters and Conference Abstracts (Non Peer Reviewed)

Shaw, P.D., Kennedy, J., Graham, M., Marshall, D. 2014. Evaluation of Helium: Visualization of Large Scale Plant Pedigrees. *BiVi 1st Annual Meeting*. Edinburgh, Scotland, 16-17 December 2014.

Stephen, G., Milne, I, Bayer, M., Shaw, P.D., Raubach, S., Hearne, S., Singh, S., Wenzl, P. and Marshall, D. 2014. Graphical applications for visualization and analysis of genotypic data sets. *VIZBI*, Heidelberg, Germany, 5th-7th March 2014.

Glossary

While these terms may have meaning outside of plant genetics and breeding they are presented here in a plant context only.

Abiotic – Non-living chemical or physical environmental components.

Allele – Alternate form of a gene at a specific locus.

Biotic – Living environmental component.

Crop – Cultivated plant that is harvested by humans. The term crop is used to refer, in this work, to a cultivated plant which is grown for food although the term can equally apply to other organisms such as algae or fungi which are grown for food and biofuel production.

Cross – The mating of two plants (or a single plant if the result of self-fertilisation).

Domestication – Process whereby a population of plants is modified at the genetic level to increase or develop characters which are desirable and beneficial to humans.

Genotype – The genetic composition of an organism but in this thesis refers to the state at a specific genetic locus.

Germplasm – A collection of genetic material for an organism such as a collection of wild collected plants or cultivated varieties.

Haplotype – A set of genes which are inherited together from a single parent.

Heritable character – Characters whose observed effects are attributed to genetic differences.

Heterozygous – Genetic locus which is composed of two varying allelic states.

Homozygous – Genetic locus which is composed of the same allelic states on each homologous chromosome.

Hybrid – The offspring of two plants of different species or varieties (crossing of two genetically discrete individuals).

Inheritance – passing of genetic material from parents to progeny.

Monogenic – Effect caused allelic states in a single gene.

Pathogen – An organism that can cause disease in another organism.

Pedigree – Genetic relationship between individuals in a population.

Phenotype – An organism’s observable or measureable physiological, biochemical, behavioural or morphological characters.

Plant breeding – Manipulation of a plants traits to create a plant with more desirable characteristics.

Polygenic – Effects caused by allelic states within multiple genes.

Selfing / Self-pollination – Flower has both compatible male and female organs (stamen and carpel) which make contact with each other to achieve pollination.

Single Nucleotide Polymorphism (SNP) – DNA sequence variation at a single locus which occurs commonly within a population.

Chapter 1 Introduction

Fundamental problems are facing humans in maintaining food security. Issues ranging from sourcing of phosphate based fertiliser, disease and pest resistance, the increase in water demand tied with decreasing availability, the requirement to increase yields and the impending issue associated with climate changes will challenge current agricultural systems to, and beyond breaking point. More efficient nutrient-use systems and the development of plant varieties bred to be more tolerant to stress conditions such as drought, waterlogging and disease and pest resistance while maintaining yields are desperately required. While advances in fertilisers, technology transfer, better breeding strategies and mechanised automation have improved crop production, only the development of new varieties to combat both abiotic and biotic stresses can be viewed as a sustainable practice going forward.

Visualization tools that bring together the diverse data types used in plant breeding will have major impact in this process.

Within the next few decades the human population faces very real environmental and food security problems which will have a disproportionate effect on developing countries that are less able to afford increases in commodity prices (Winkler 2005; Unfccc 2007; Mertz et al. 2009). Current genotyping and phenotyping technologies allow large volumes of data to be generated both quickly, and in many cases cost effectively. These categories of data form the foundations for both plant genetics and plant breeding programmes. Subsequently, computer visualization tools that allow breeders to track the inheritance of agriculturally important alleles (alternate gene forms at the same genetic locus) and bring together and visualize these diverse experimental data types will allow more informed decisions on which biological crosses are performed, leading to more productive and efficient plant breeding programmes.

In recent decades plant breeding (or more appropriately plant selection before the introduction of targeted molecular techniques) and genetics has been principally focussed on increasing plant yields, disease resistance and product quality traits but is now focusing increasingly on adaptation to climatic change, water and nutrient availability. Examples of this include the breeding of plant lines which are more efficient at nitrogen uptake from the environment, thereby mitigating the requirements for addition of fertilisers and subsequently reducing N_2O (Nitrous oxide) emissions, a known greenhouse gas (Burney, Davis, and Lobell 2010; Rosario et al. 2003; Raun and Johnson 1999), and the development of varieties more resistant to biotic pests and disease (Miedaner and Korzun 2012; Piffanelli et al. 2004; Buschges et al. 1997).

Drought is a major problem facing cereal production worldwide (Hoekstra and Mekonnen 2012; Pimentel et al. 2004). Recent documents point towards greater fluctuation and variation in rainfall and statistical climate models suggest longer warmer summers (Field 2014). Additionally, an increasing population will place a higher demand on available water for industrial and domestic utilisation leading to a scarcer and more expensive commodity for crop irrigation purposes, a process that is often necessary for food production. It is important to recognise that while drought is often held up as an example of the problems associated with climate change, the problem of increased water for some geographical regions is also important (Karl, Melillo, and Peterson 2009). Increased water goes hand-in-hand with drought as hard compacted soil, which can result from longer warmer summers, is a major factor in surface flooding.

Commercially important agronomic traits are under the control of one (monogenic) or more (polygenic) genes plus environmental interaction. Deoxyribonucleic acid (DNA) based technologies have allowed geneticists to study the underlying genetic factors affecting many of the current agriculturally important characters (either biologically significant or traits used in plant registration processes) and have shown that many are under the influence of complex genetic regulatory systems such as Na+ tolerance (Mian et al. 2011; Wu et al. 2013; Colmer, Munns, and Flowers 2005).

The ability to determine the underlying genetic basis for expression of phenotype is critical in plant breeding as it facilitates the targeted selection of plants showing desirable heritable characteristics for inclusion into new genetic backgrounds. This is significant in increasing yields and in identifying potential new sources for pest and pathogen resistance. There is however increasing acceptance that new breeding strategies must be developed to produce innovative varieties suitable for growth in fluctuating environments to protect and maintain food security and sustainability for society and consequently profitability for commercial plant breeding companies and farmers.

Current genotyping technologies give researchers access to more data than ever available before. The efficient integration of this data with phenotypic and genotypic data will lead to advances in the association of traits with their genetic foundations and subsequently, improvements in plant breeding.

Barley (*Hordeum vulgare* spp.) is currently the fourth most important small grained cereal crop worldwide and the largest arable crop grown in Scotland accounting for 300,000 hectares and an annual production of 1.7 million tonnes (140 million tonnes worldwide (FAOSTAT CEST 2014). Barley has particular value not only to the Scottish economy but also that of Europe which produces 61.9% of worldwide barley according to figures published by the Food and Agricultural Organization of the United Nations (FAO).

Barley has major significance to Scottish industry due to its use in the production of beer, Scotch whisky and animal feed. While worldwide use in the brewing and distilling industries account for around 15% of total output, in Scotland this increases to between 40 and 50% so is of significant importance. In Scotland barley production outnumbers that of wheat which is the second most farmed crop by 2:1 although its growth requires 3 times as much available land due to the greater yield per hectare of wheat.

Due to the high-throughput nature of many modern genetic and genomics techniques, the ability to generate data far in excess of current limitations of desktop computers is common. The ability to deal with such data requires the capability to analyse the data and report on calculated findings, or to be able to efficiently visualize large quantities of data. Recent software has incorporated statistical analysis using high performance computer clusters tied to end user visualizations. The ability to utilize such compute resources allows end users to perform analysis far quicker and more efficiently than before. In addition, the use of visualization components is essential for data exploration, analytics and problem data identification.

Using genotypic data derived from SNP (Single Nucleotide Polymorphism) assays, around 1000 barley plant lines of which many are in the current Home Grown Cereals Authority (HGCA 2014) recommended list for barley varieties in the United Kingdom (UK) and phenotypic data in the form of Distinctiveness Uniformity and Stability (DUS) data, this work examines the merging of phenotypic and genotypic data in a pedigree context and advances the current pedigree visualization techniques for plants.

1.1 Aims and contribution

The aim of this work is to adapt and critically evaluate current pedigree visualization and visual analytics techniques as a means to apply and develop methods to help address the problems and difficulties that modern geneticists and plant breeders have when working and interacting with large and complex pedigree and associated experimentally derived data. These problems include the complexity of plant crosses and data diversity and density and are specific to plant breeding.

Plant breeders need to be able to visualize plant pedigrees in order to gain a deeper understanding of the genetic composition of commercial varieties. While it is known that plant pedigrees are often complex the lack of suitable tools for the visualization of plant pedigree data has meant that this complexity has not been effectively quantified. Plant breeders and geneticists have been looking for visualization techniques and applications to allow them to accurately model these complex pedigrees as a means to select plant lines which could lead to new plant varieties containing important traits either of agricultural importance of useful for plant varietal testing systems. This work will examine how traditional pedigree visualizations can be improved to allow for the diversity of data types that are routinely used in modern plant breeding programmes and experimental plant genetics and identify the problems that exist that mean current pedigree visualization tools are not suitable for plant breeding.

Discussion with potential users showed that there was a need for tools to allow them to accurately determine lineage (parental, both paternal and maternal contribution) within complex pedigree structures. Current tracking was focussed around the use of lists of textual pedigree strings and user-curated spreadsheets. They highlighted that tools that would facilitate the overlaying of phenotypic and genotypic information in order to try and identify both errors, and patterns in the underlying pedigree structure were unavailable and desperately required. This work aimed to address these concerns.

There are clear problems that exist that mean current pedigree visualization tools are not suitable for plant breeding and genetics and a key question will be how these problems can be addressed by the modification of current Information Visualization (IV) principals to integrate experimental data with pedigrees and handle plant as opposed to mammalian pedigrees which are fundamentally different.

A computer based application to visualize the inheritance of genes of agricultural importance in a pedigree context would allow plant breeders and researchers to identify and track potential plant lines carrying beneficial alleles for agriculturally important traits.

This work advances the current state of the art knowledge within the biological visualization domain by applying established visualization principals to the specific and well accepted problem of plant pedigree visualization. It adopts techniques used in traditional animal and human pedigree representations and information visualization and applies these to handle the specific problems which are associates with the visualization of plant pedigrees. It has improved these techniques to account for the differences that exist between animal and plant pedigrees and provide a pedigree visualization tool (Helium) to bring together different data types (pedigree, phenotype and genotype) and utilise these data sources to visualize transmission patterns within complex plant pedigrees.

This research has allowed the definitions of terms to describe pedigree shapes (pedigree *delta* and pedigree *net* types) which describe the topology of the pedigree when visualized in Helium as well as the identification of plant lines, which have been coined *principal*, *flanking* and *terminal* to describe their relative position within their defined pedigree.

Finally the effectiveness of the Helium tool was verified through carrying out of two distinct rounds of user testing on a group of domain experts (16 and 28 expert users), asking them carry out the tasks and problems which were identified at the start of this research.

1.2 Thesis Organisation

Chapter 2 introduces plant breeding and the biological problems that can be addressed using plant pedigrees. It details the importance of plant breeding (in particular barley) to the Scottish economy then concludes by describing the datatypes and datasets that will form the basis of this work and the initial feedback and requirements gained from users.

Chapter 3 examines how visualization techniques can help in exploring large biological data sets. Current visualization techniques including pedigree visualization tools are evaluated in terms of their suitability in exploring plant breeding data and limitations of existing systems identified.

Chapter 4 describes a static paper-based prototype pedigree visualization system called Orb and the user feedback process used to establish user acceptance of this tool and the abstract visualization techniques it used. It begins to set the scene for a more advanced interactive pedigree visualization tool.

Chapter 5 describes the move from the paper-based prototype in Chapter 4 and how this was developed into a prototype Java based interactive pedigree visualization tool called Helium and the iterative process used in development and refinement of this application based on expert user feedback.

Chapter 6 details the user-centred subjective evaluation of the interactive prototype then presents these test results.

Chapter 7 highlights the changes and refinements made to Helium based on the evaluation data in Chapter 6. This chapter highlights the major issues that testing uncovered and subsequently describes solutions implemented to overcome these problems.

Chapter 8 describes the second round of testing on Helium after the modifications and tweaks had been made which were described in Chapter 7 and details the results from a second round of user testing culminating in a comparison between the two Helium interactive prototypes and what the user testing uncovered.

Chapter 9 discusses the outcomes of this work and where this work sits in the body of literature in this area. It then goes on to describe the main conclusions, highlights weaknesses, and suggests possible future work to advance this field.

Chapter 2 Plant Breeding and plant pedigrees

Both climate change and an increasing population are giving rise to new challenges for plant breeders. There is a requirement to develop new varieties, which both increase yield and disease resistance to meet these demands. Plant pedigrees are a representation of how genetically similar plants are related to one another and, when coupled with phenotypic and genotypic data, can be utilised to help plant breeders and geneticists make more informed decisions as to which plant lines to cross to achieve favourable outcomes.

There are however many problems in dealing with large and diverse phenotypic and genotypic datasets. Incorporating this data and pedigree data into breeding programmes and experimental plant genetics will facilitate the creation of new plant varieties that are both higher yielding and more adapted to changing environmental conditions.

2.1 Overview

This chapter reviews the history of plant breeding and its importance to food security. It then introduces the complex problem of tracking alleles of interest in breeding programmes using pedigrees and the potential problems that are facing the industry going forwards. It outlines the background behind the biological data and biological concepts that have formed the foundations for this work, a description of the data which forms the basis of this work and concludes with the results of an initial feedback session with 6 plant breeders and geneticists which would be used to direct any subsequent work.

2.2 History of plant breeding

A prerequisite for what is now defined as plant breeding was the domestication of plants. Domestication is an anthropological phenomenon where human populations have, over time, selected plants that have some measureable qualities which made them better than wild plants for procedures valuable to humans such as processing, storage and digesting. Over time this domestication leads to plants that are morphologically very different from their wild ancestors.

Domestication brings about the selective advantage of rare mutants or alleles which are necessary for survival in cultivation but not for survival in the wild. It has often been said that modern domesticated plants are as reliant on humans for their survival as we are of them (Kingsbury 2009). In addition, changes in allele frequency, gradations between species, the fixation of major genes and the improvement of quantitative traits are all genetic outcomes of domestication.

Plant breeding, and in particular that of the Triticeae (grasses) is thought to have originated around 12,000 years ago (Zohary, Hopf, and Weiss 2012; Kilian et al. 2009) when sedentism overtook nomadic lifestyles across the Middle East. This settling of populations depended on the availability of accessible agricultural land and was a prerequisite for the establishment of modern agricultural practices including both crop and animal cultivation.

Early plant selection processes are thought to have focussed on reducing adverse phenotypes relating to seed dispersal, such as shattering, so seeds only break from the rachis during threshing. This character would not be evolutionary advantageous for a wild species. Other examples include the removal of husks or glumes to increase processability and digestibility, reduction in branching, reduction in height, synchronous tillering (flowering and ripening), reduction in internode length, and determinacy (simultaneous flowering) (Peterson 2011). Other phenotypes, which would have been targeted by early plant cultivators, include the production of larger seeds and reduction in toxic compounds such as phenolics, which are known to cause bitterness. Techniques such as vegetative propagation essentially lead to instant domestication. In early domestication, farmers selected plants which had desirable characteristics, took cuttings or tubers and repeated year on year; a process known as selection.

The discovery of plants that are self-pollinated was also an important development in the cultivation of modern crops. Self-pollination allows uniformity and stability in progeny and while not particularly common in the plant kingdom generally, is commonly seen in domesticated cereals such as barley, legumes such as soybean and sunflowers. It does of course have disadvantages but in the context of commercial cultivation these are harnessed as advantages, such as in the case of maintaining stable traits.

2.2.1 *Modern cereals*

Current cereals such as barley, wheat, maize and rice are thought to have originated from 5,000 to 8,000 years ago in the Fertile Crescent (Nesbitt, M. Samuel 1995), a phrase coined by James Henry Breasted, which encompassed what is now defined as the Middle East. During the Neolithic epoch the region was seen as the birthplace of civilisation and was the first known area to use mass irrigation, a technique fundamental in the development of mass agriculture.

The different climates and topologies of the Fertile Crescent region gave rise to a large and diverse range of plants which were cultivated for human and animal consumption. The ecological succession of annual plants that produce large grain is well documented (Brown et al. 2009; Fuller 2007). These, which are termed Neolithic Founder Crops (Weiss and Zohary 2011), include Emmer wheat (*Triticum dicoccum*), Einkorn wheat (*Triticum monococcum*), lentils (*Lens culinaris*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), bitter vetch (*Vicia ervilla*), flax (*Linum usitatissimum*) and barley (*Hordeum vulgare*) which was descended from the wild *Hordeum spontaneum* which still exist in southwest Asia (Harlan and Zohary 1966). These primary domesticates now form the basis for much of the worldwide commercial agriculture along with maize (*Zea mays*) and rice (*Oryza sativa*) and are extremely important both for feeding humans and animals and maintaining food industries.

Since the first crops were selected in the Neolithic period farmers have been ‘genetically engineering’ them, through selection techniques, in order to provide characteristics which are beneficial to humans. These features which are selected will be referred to as traits with phenotypes referring to their specific expression variant. Examples include; larger grain sizes, greater disease resistance, better drought or frost tolerance and shorter or longer stems.

2.3 **Why change is needed**

Changing environmental conditions coupled with an increasing populace is placing increasing demand on farmers to produce enough crops to effectively feed the world’s population. Factors such as climate change (Figure 2-1) with a trend towards increased temperatures in crop growing regions (Figure 2-2) are posing new and important challenges and questions on how the world's population can be fed through changing times. The observed 'cool spot' (Figure 2-1) found in the North Atlantic south of Greenland is thought to be caused by weakening ocean currents (Rahmstorf et al. 2015). The Food and Agriculture Organisation (FAO) puts the figure of people who are suffering from chronic undernourishment in 2012 at close to 870 million. This figure is 15% of the population of developing countries (FAO 2012).

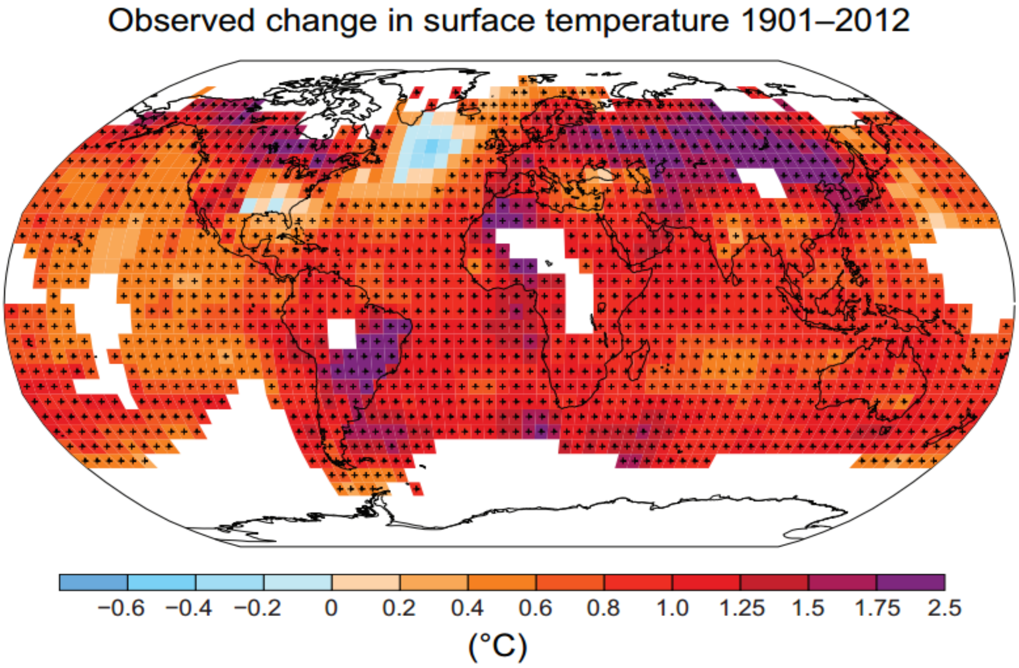


Figure 2-1 Observed surface temperature change 1901-2012 (IPCC 2013)

It is important to note that this is not a problem focussed purely in the developing world. Developed nations still need to produce enough food to feed their populations and changes in environment will put an unimaginable strain on mass agriculture. It is estimated that by 2030 there will be a 400% increase in the amount of water required for crop production (Foley et al. 2011).

In order to address these global issues, plant breeders will need to adopt more efficient selection strategies for choosing plant varieties which are better suited to current global climate trends and include desirable characteristics (or phenotypes) which aid in increased yield, disease and pathogen resistance and more efficient nutrient and water utilisation in commercial production systems.

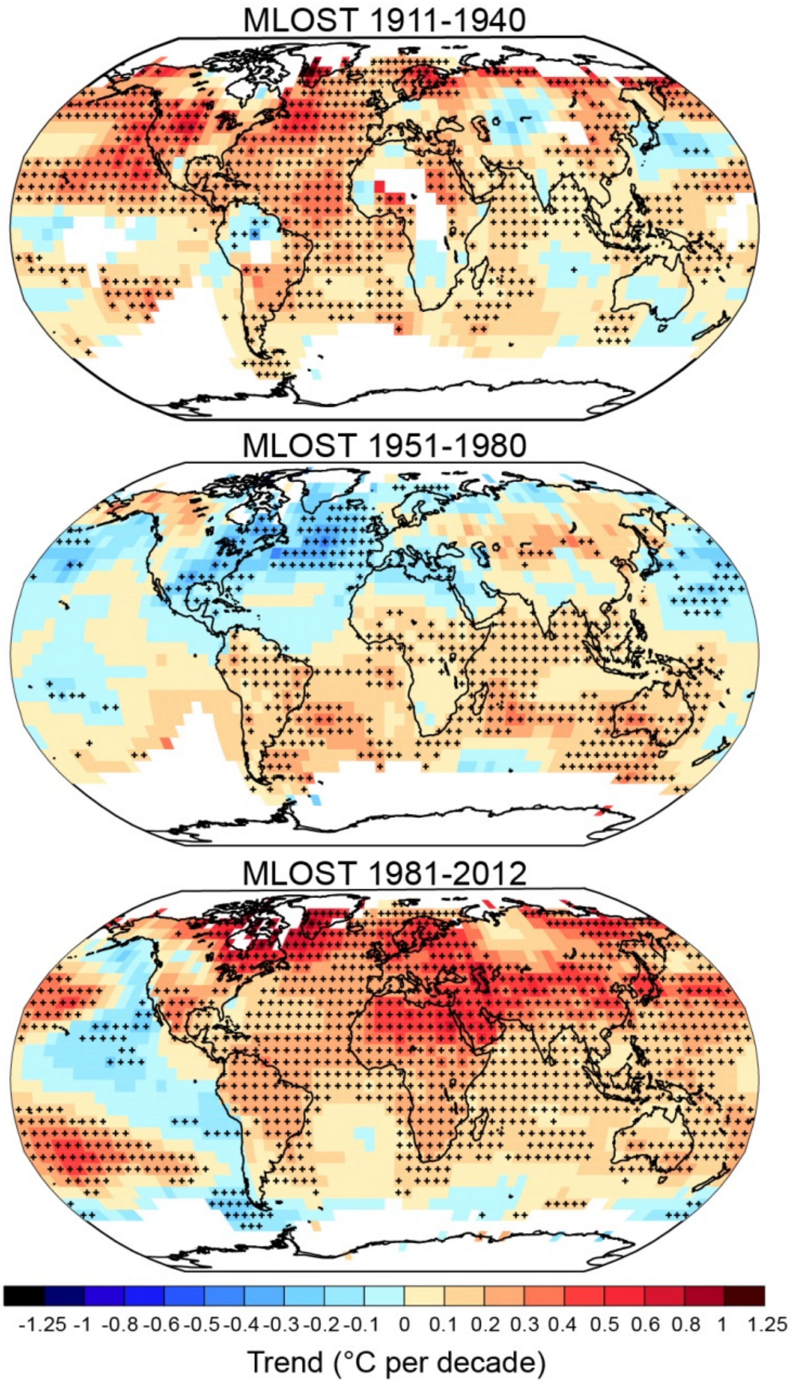


Figure 2-2 Surface temperature trend 1911-2012 (IPCC 2013)

While countries like the UK could adapt more easily than other European countries to changing temperatures by the utilisation of germplasm currently grown in either regions of higher or lower latitude, there is a more difficult problem with regards to increased rainfall in Northern Europe and a decline in regions which currently have water availability issues such as Southern Europe (Figure 2-3).

In addition to the problems associated with climate change and input availability (water and nutrients), countries such as Scotland also have additional problems. Due to Scotland's large export market for cereal-derived products such as Scotch whisky (barley), distilled spirits (predominantly but not exclusively barley and wheat) and brewing (barley) there is a requirement to increase grain production in order to meet the increased worldwide demand for such drinks based products.

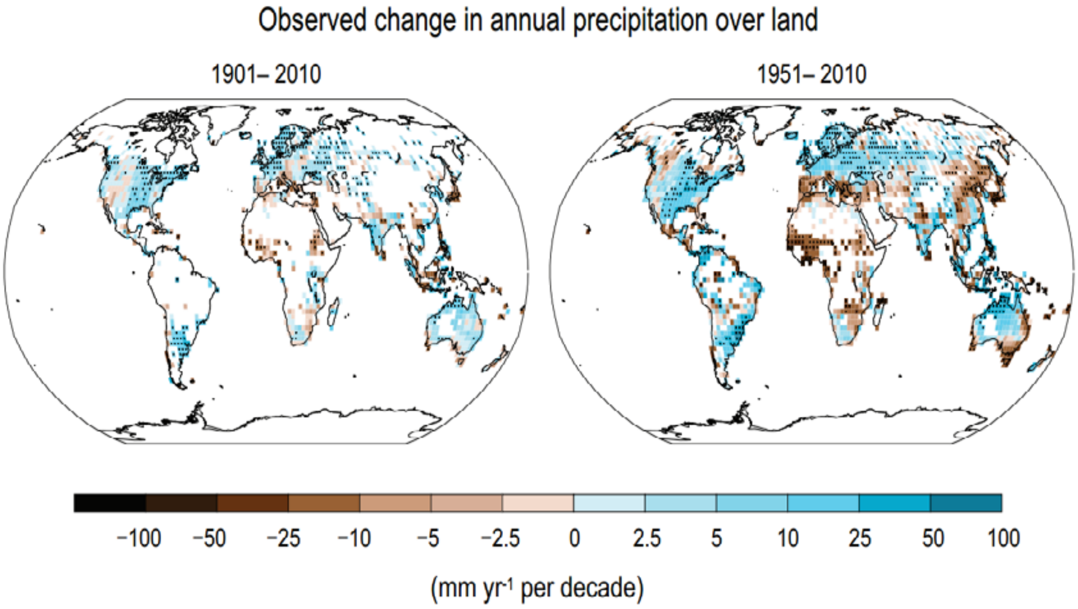


Figure 2-3 Surface rainfall observation 1901-2010(IPCC 2013)

2.1 Barley and the Scottish economy

Figures from the Scotch Whisky Association show that there was an increase in the value of whisky exports from Scotland in 2013 compared to the previous year with exports at £2 billion for the first half of 2013. This equates to the export of 563 million bottles ("Scotch Whisky Association - Home" 2014). In 2013 the annual value of exports of Scotch whisky were put at £4.5 billion. The largest export markets are the USA (£757 million), France (£434 million - although France imports a higher volume than the USA) and Singapore (£340 million) although Singapore acts as a distribution hub for much of Asia so figures do not apply to Singapore alone ("Scotch Whisky Association 2012 Statistical Report" 2014). In the UK 35,000 jobs are supported by the industry (Scottish Government 2012) (2011 figures) and there are 109 licenced distilleries in Scotland. The £4.5 billion in exports means that whisky accounts for 25% of UK food and drink exports. Indeed, such is the popularity of Scottish whisky that more is sold in France per month than Cognac per year.

Whisky is Scotland's second largest export after electronics and the geographic location of distilleries means that they are important employers in rural areas where there is an inherent lack of employment opportunities. For the last 20 years Scottish whisky exports have exceeded £2 billion annually (£2.09 billion in 1993, £4.27 billion in 2012) (Figure 2-4).

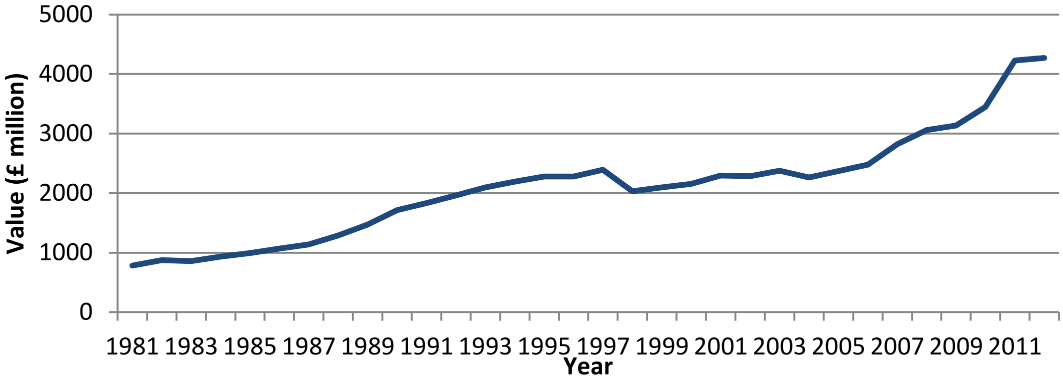


Figure 2-4 Scottish Whisky export value (£ million) 1980-2012

Because of the impact to the Scottish economy the efficient breeding of new varieties that account for changing climate conditions is increasingly important, as is the ability to supply the distilling industry with enough raw product to meet demand (Anderson 2014) as there are already supply problems. This is particularly important as Scotch whisky not only has protected Geographical Identification (GI) status (WIPO 2015) but is a high value item with worldwide recognition and quality acceptance.

The rapid advances in modern genomics is leading to an unprecedented ability to identify causal genes for economically important characters such as yield and malting quality. These advances facilitate the selection of pools of enriched germplasm in the laboratory and as a result improve the effectiveness of field trialling conducted in conventional breeding programmes. Whilst the costs of developing the resources to conduct such programmes have reduced, they are usually still much too resource intensive to be conducted by any one breeding company.

At this moment in time, the development of new barley varieties in the UK and Europe is predominantly carried out by private breeding companies such as, but not limited to, Syngenta, KWS and Limagrain.

In 2012-13 the amount of utilised arable land in the UK was 17.3 million hectares. There was a decrease in usage of 66% due to poor weather conditions, which meant that many farmers were unable to plant crops as usual. The largest crop in terms of production in the UK is wheat, followed by barley then oilseed rape (*Brassica napa*) (“National Statistics” 2014).

In 2014 the amount of barley grown in the UK increased by 22% to 8,174 million tonnes (“DEFRA Farming and Food Brief” 2014) large amount of this increase down to the competitiveness against wheat for animal feed. The use in 2014 with distillers and maltsters was 147 thousand tonnes. However, increases in cereal productivity are slowing from 2.3% a year in the 60s and 70s to 1% since 1990. There are also problems relating to degradation of crop land, water availability growing competition between fuel and food, climate change. Farmers will need to double output in the next few decades to meet both current needs and the increasing demand for food. Indeed, there are already reports of climate change adversely affecting current crop production (Intergovernmental Panel on Climate Change 2014; Goldenberg 2014).

One of the major issues facing modern agriculture is how the second ‘Green Revolution’ can be achieved. While the original ‘Green Revolution’ in wheat in the 1960’s reversed the problems associated with food shortages in India and Pakistan: there is a requirement for the development of precision and sustainable agricultural systems, better disease and pest management procedures, genetically modified crops and public and private research covering the development of new varieties and the exploitation of existing germplasm collections to be used in order to try and meet the problems that the world is facing with food security.

2.1 Genetic inheritance / transmission

Recombination is the process where, during sexual reproduction in eukaryotes, DNA molecules exchange genetic information resulting in the creation of new combinations of alleles. Recombination involves the pairing of homologous chromosomes which is then followed by the information exchange between chromosomes.

In meiotic (sexual) recombination the genetic composition of an individual is split, on average, 0.5:0.5 between its parents if the result of cross fertilisation or 1 if the parent is self-fertilised. Non-parental alleles can only result from either a misclassified genotype or the result of genetic mutation. Alleles must therefore be inherited from either parent. If this is not the case then it suggests there are problems with either the underlying genotypic data or misclassification of plant lines.

This is an important concept when using pedigrees as the presence or absence of a specific allele can be modelled and inference made based on parental contribution to offspring. We therefore know that if a child displays the presence of a particular allele it will have either inherited this from either of its parents (who must also have the same allele) in a process known as identity by descent (IBD) or through random mutation in the particular allele (Identity by association, IBA)). Both IBD and IBA can lead to the same potential outcome through two distinct biological processes.

The knowledge of this Mendelian inheritance allows potential problems with genotypic data to be identified based on the presence of genetic elements not in their parents and allows complex pedigrees to be constructed. This will be discussed more in Section 2.4.

2.2 Cultivars and inbreeding

A cultivar is a plant (or grouping/population of plants in a species such as maize (*Zea mays*)) which has been selected due to the presence of desirable characteristics or traits which can then be propagated to create plants with identical characteristics to their parents. Most commercially grown plants are cultivars and the selection and/or breeding process allows uniformity, leading to plants with predictable phenotypes suitable for commercial production. This stability, and indeed homogeneity and predictability, can be attributed to the inbreeding that routinely happens in commercial varieties whereby heterozygous individuals through a process of selfing, or through alternative genetic engineering processes such as the creation of doubled haploids (whereby haploid cells undergo artificial chromosome doubling) to create homozygous individuals with predictable characteristics. While inbreeding has detrimental effects in humans and can be the cause of many congenital birth defects due to the increase in chance of expressing recessive detrimental alleles, in commercial crop varieties the ability to predict phenotype with a degree of accuracy allows a product which can be sold commercially owing to its desirable characteristics. Indeed, since the majority of flowering plants are hermaphroditic, inbreeding or autogamy represents a breeding strategy that exists and is common in flowering plants. Subsequently, many crop species reproduce vegetatively and apomictically whereby the produced seeds have the same genotypic composition as the mother plant resulting in offspring that are genetic clones.

Problems such as inbreeding depression whereby there is a detrimental effect through mating with close relatives are also more easily tolerated in non-human systems and in plants tend to be most apparent in phenotypes such as pollen quantity, seed generation and growth rate (Keller and Waller 2002).

The experimental data used in this work deals with inbred barley varieties which are homozygous. Commercial varieties (and those developed as part of breeding programmes) are assigned a name to define a population of genetically identical plants derived from homozygous seed. These names are often the preferred names by the originating breeder and include examples in barley such as ‘Optic’, ‘Golden Promise’ and ‘Tipple’. As can be seen there is often a link between the targeted use of the plant line (commonly brewing and distilling) and its varietal name. It is important to remember that when a plant variety is referred to as a name it is in fact a population of genetically identical individuals as opposed to, for example in humans, an individual unique genotype.

2.3 **Pedigrees**

A pedigree (Figure 2.5) is a representation of how genetically discrete individuals are related (usually but not exclusively) in time to one another. It is therefore a representation of the genetic relationship between individuals, their parents and progeny (predecessors/ancestors and successors/descendants). Pedigrees are often used in human contexts to show the transmission of alleles responsible for genetic conditions of medical importance or for the display of traits and phenotypes of medical, or research importance.

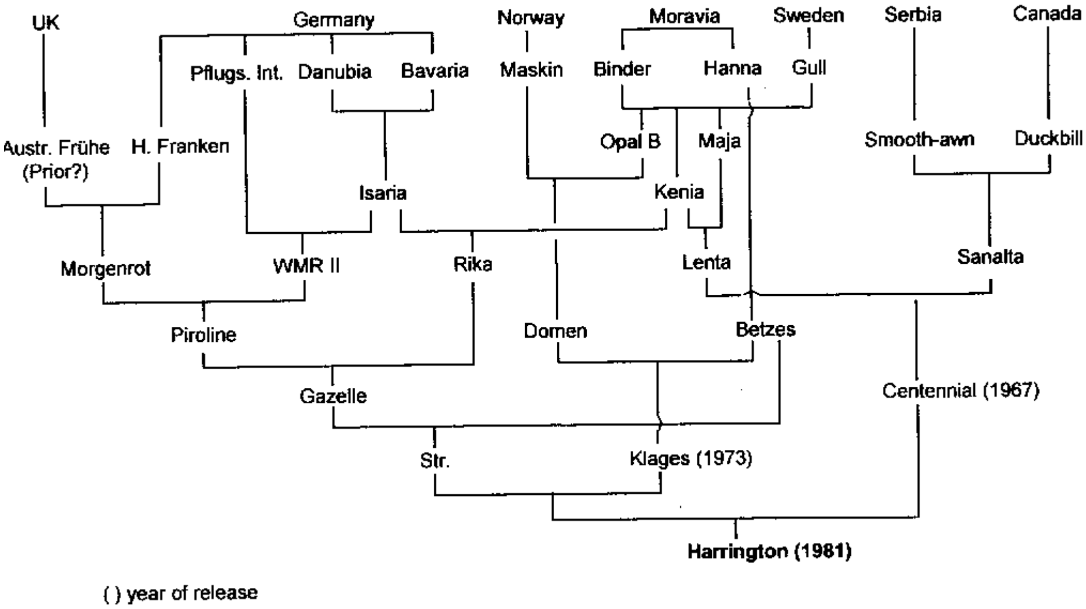


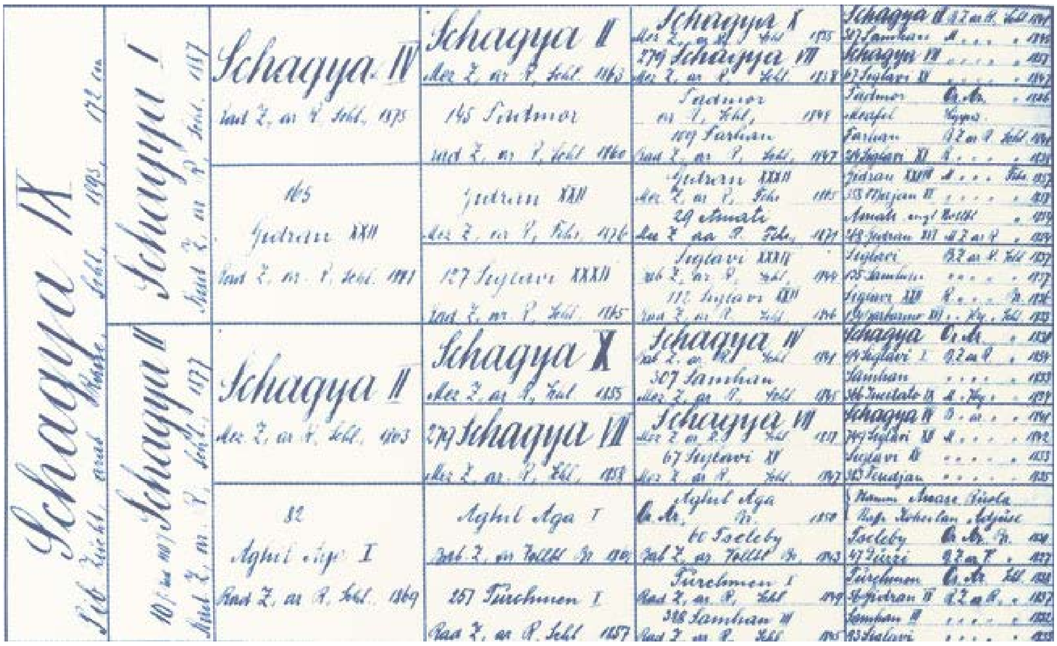
Figure 2-5 Typical barley pedigree (Fishbeck 2003)

2.3.1 *Pedigrees in humans and other animals*

Pedigree charts are most commonly used to show relatedness within species, in particular, and most commonly, in humans, dogs, birds and racehorses. The word pedigree is said to have come from a derivation of the French word ‘pied de grue’ which crudely translates to ‘cranes foot’ in recognition of the physical appearance of early pedigree diagrams (Oxford English Dictionary 2002).

Most people will be familiar with the concept of pedigree animals where individuals are of known genetic descent and breeding stock. In animals, pedigree charts have been used for over a hundred years to show ancestry of successful breeds. They are also used to select individuals with specific desirable traits which can be used in subsequent crosses. An early example showing the first known horse pedigree (Figure 2-6) shows the tree-like structure represented horizontally for six generations.

The main function of pedigree charts, as in any successful visualization is to take complex relationships and present them in a way that is easy to comprehend and understand. In humans, pedigrees use standard nomenclatures to show both male and female members. Generations are traditionally represented using Roman numerals (I, IV, V and so on) and individuals within a generation by standard numbering (1, 2 and 3). In human pedigrees the base individual, insofar as the individual which is selected as the focus or root node of any diagram is referred to as the proband; however this is not seen in plants. In medical circumstances pedigrees are commonly used to determine an individual's chances of showing a particular genetic disorder or working out the chances of progeny inheriting such conditions where the condition has a hereditary component such as diabetes or hypertension. In a human context pedigrees can also be used to determine the genetic basis for disease, autosomal or x-linked based on the percentage of individuals displaying the phenotype in the pedigree. Whether a condition is dominant or recessive can also be derived from such diagrams. It should be noted that in a human context the numbers of individuals in any chart is relatively low, again quite different from most plant pedigrees (Figure 2-7).



Schagya IX [illegible] 1895 [illegible]

Schagya I [illegible] 1887

Schagya II [illegible] 1877

Schagya IV [illegible] 1875

165 Gidran XXII [illegible] 1881

Schagya II [illegible] 1863

82 Aghil Aga I [illegible] 1869

Schagya II [illegible] 1863

145 Sadmor [illegible] 1860

Gidran XXII [illegible] 1876

127 Saglavi XXXII [illegible] 1865

Schagya X [illegible] 1855

279 Schagya VII [illegible] 1858

Aghil Aga I [illegible]

257 Turkmen I [illegible] 1857

Schagya X [illegible] 1855

279 Schagya VII [illegible] 1858

Sadmor [illegible] 1848

109 Farhan [illegible] 1847

Gidran XXXII [illegible] 1865

29 Amale [illegible] 1871

Saglavi XXXII [illegible] 1849

112 Saglavi XXII [illegible] 1846

Schagya VI [illegible] 1841

307 Samhan [illegible] 1845

Schagya VII [illegible] 1837

67 Saglavi XI [illegible] 1847

Aghil Aga [illegible] 1850

60 Toeleby [illegible] 1843

Turkmen I [illegible] 1849

318 Samhan III [illegible] 1845

Figure 2-6 first recorded horse pedigree

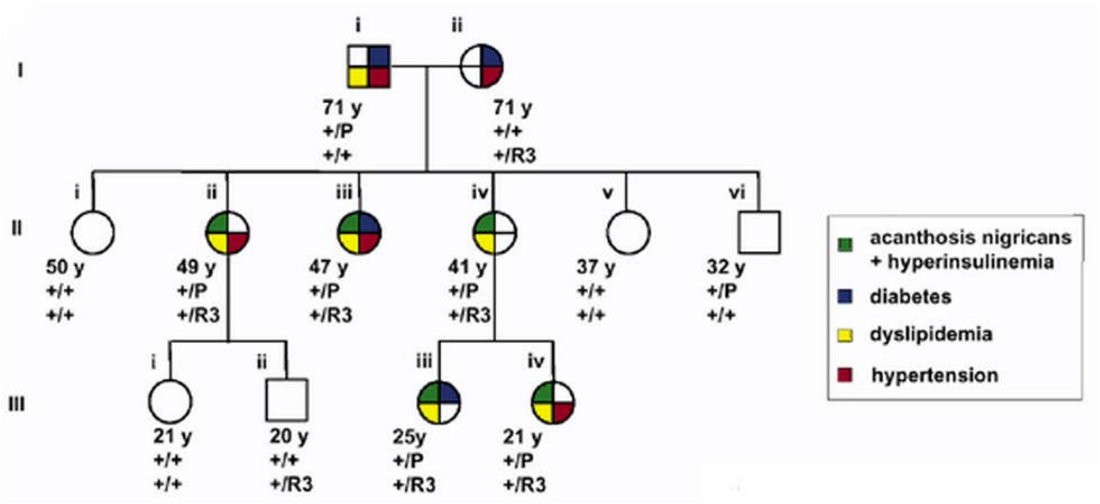


Figure 2-7 Human pedigree showing insulin resistance (Savage et al. 2002)

An adaptation of the common human family tree, the genogram, was described by Jolly (Jolly, Froom, and Rosen 1980) as a means of including additional information on to the family tree diagram. The genogram is used in a medical context and allows the inclusion of additional information such as relationship type (Figure 2-8) and the status of progeny. Additional information is also included on such charts such as dates of birth and death and disease states.

Genograms have a specific purpose in family medicine but detract from specific genetic relatedness and so can be discounted when looking at data in a specific genetic context. They aim to show the relationships that exist within human familial contexts where there may be same-sex couples or where divorced individuals have got back together. They also include dates of birth and death. In essence they are a description of the processes that go on in families that may or may not have a genetic basis.

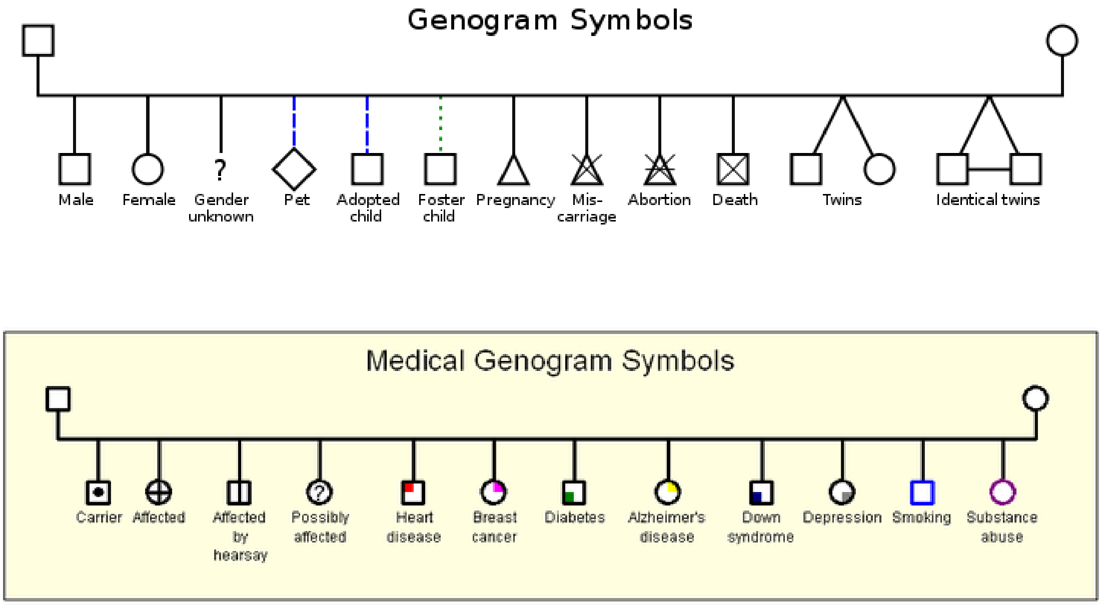


Figure 2-8 Genogram diagrams used in human pedigree representations

2.4 Plant Pedigrees

In plants, pedigrees can be used as a framework along with environmental data, on which statistical analysis can be used to determine factors such as mode of inheritance (Identity by Descent, IBD and Identity by Association, IBA). Additionally, they are often used to check for potential genotyping errors, since these errors, by the very nature of Mendelian inheritance, are constrained by the pedigree structure in which they exist (Paterson et al. 2011; Graham et al. 2011; Paterson et al. 2012). The accurate representation of pedigrees is therefore becoming increasingly important in plant breeding and genetics.

In the human context, pedigrees are a representation of a family tree. An individual has two parents. Each parent has two parents and so on. In this way a complex structure showing genetic relatedness can be built up. Unlike in humans where there are sensitive issues surrounding describing a family tree with regards to historical consanguineous mating this is not a problem in many commercial crops where differing varieties are genetically related to one another.

While there are defined standard nomenclatures for human pedigrees (Robin L Bennett et al. 1995; Robin L Bennett et al. 2008) there is no *single* formal system for plant pedigrees, however, there are moves towards defining standards. There are valid biological reasons for this including: the hermaphrodite nature of most plant species, the complexity of mating designs possible in plant genetics and, finally, the absence of any overseeing coordinating organisation.

While authors such as Purdy (1968) and Lamacraft (1973) suggested improvements to allow for better processing of pedigree strings by computers, these were in the late 1960's and early 1970's when computing was in its relative infancy (Purdy et al. 1968; Lamacraft and Finlay 1973). Since then, while there have been suggested improvements for human centred pedigrees (Robert L Bennett et al. 1995; Robin L Bennett et al. 1995), there has been no major improvements to Purdy and Lamacrafts' initial suggestions for a standard nomenclature in the plant domain. Figure 2-9a shows the Purdy Notation System (Purdy et al. 1968) established as a common format for representing small grain cereal pedigrees. Forward slashes '/' are used to delimit plant lines. In this case A is crossed with B which is then crossed with C whose progeny is crossed with D. Lamacraft and Finlay notation (Lamacraft and Finlay 1973)(Figure 2-9b) was presented as a format which could be more easily parsed by computers. The example here is the same as in the Purdy notation. Figure 2-9c shows a typical pedigree that can be found in old records where mixtures of notations are used. These mixed notation systems are common and most breeders will use shorthand that is unique to them. These records are sometimes difficult to read and would benefit from being represented in a more user friendly way.

a. A/B//C//D
b. ((A*B)*C)*D
c. [Ax[(BxC)*D]xE]*[FxA]xC

Figure 2-9 Text based pedigree records showing diversity of nomenclature.

While plant and animal breeding share routine breeding techniques such as standard crossing and back-crossing, pedigrees used in plant breeding display some subtle but important differences, often involving key shorthand conventions that are unique to plant mating designs leading to complex textual based records which can be difficult to read.

Named entities in plant pedigrees usually, but not always, represent a population of genetically identical individuals, not a single plant. While it is relatively simple to grow many plants from seed, potentially many decades after production, in humans and animals this is understandably not the norm. The generation of these genetically identical (homozygous) varieties is possible through doubled haploidy , inbreeding, or crossing of pairs of inbred plant lines to achieve what is termed an F1 (Filial 1) hybrid. Successive inbreeding by self-pollination of these F1 generation plants leads to individual plants that are close to homozygous across all alleles. The exploitation of homozygous plant lines in crop species such as barley is a powerful tool in genetic analysis, removing some of the genetic complexities associated with species (such as humans) where there is a high level of heterozygosity.

2.5 Problems with pedigree representations

Pedigrees are often referred to as trees but this is incorrect and something which should be avoided in future work or reference in this area. In order to maintain a *true* tree structure there can be no back-crossing or mating events between individuals and ancestors (known as consanguineous mating) which while relatively rare in humans does happen within several isolated human populations through either physical or cultural isolation. The problem with inbreeding in humans is that detrimental alleles which would be selected against under normal conditions persist in populations and lead to inbreeding depression and a subsequent reduction in genetic fitness of a population. In stark contrast to the human perspective inbreeding is normal and routinely exploited in commercially farmed animals and crops to achieve stability and consistency at a genetic and expressed phenotype level.

Standard nomenclatures are required in the plant community to describe plant pedigrees. There needs to be a paradigm shift away from breeders using bespoke nomenclatures wherever possible. Human nomenclatures are unsuited as they assume a logical progression through generations and through normal sexual reproduction, where inbreeding is extremely rare. Plant pedigrees however can be the result of standard reproduction, derived from a specific filial generation or the result of doubled haploidy. Each of these is conceptually very different and important in the classification of genetic material.

2.6 Experimental data

This section examines the data types that were used in this work and explains the data sources and the format in which they were presented. It describes the complex relationships that exist between the multiple data sources and types and details the problems associated with historical data and the benefits that an integrated database system brings to experimental plant genetics. Finally, it discusses the problems that exist when handling large volumes of experimental data and suggests ways in which the process is improved and developed to offer greater and more stable functionality for future work.

2.6.1 *Data types used in this work*

There are a number of main broad ranging categories of data that were used in this work: continuous (quantitative), ordered and categorical (qualitative). The continuous data refers to measurements that are quantitative in nature, for example plant height or a leaf surface area where measurements are non-discrete/continuous and follow a normal distribution. Ordered data refers to data which has a defined order, whether this is in relation to, for example, time readings or the classification of a trait on a known and non-quantitative scale (ordinal). Categorical or nominal data refers to a descriptive classification of a data point. Barley winter and spring ecotypes are an example of this where a plant line can be classified as one or the other with no associated quantitative or qualitative measures. Another example would be the alternate type or hairiness of leaf sheaths where hair is either present or absent. Additionally, there is also interval data, mainly used to classify DUS characteristics whereby the classification of a phenotype is represented as a number within a scale, an example would include the categories for intensity of anthocyanin colour which range from absent to very weak to very strong with 7 divisions in between. The number of categories within the DUS data ranges from 2 (present, absent) to over 12. The upper end of this is very much an exception and most data falls within the 3 to 7 category range. There is no temporal aspect to these recordings as a variety should have stable readings year on year.

Furthermore, the data can be classified as parametric whereby it is assumed the data adheres to some sort of probability distribution or non-parametric where the data can only be classified into groupings.

Within these categories of data there are a number of different data types along with meta-data describing the data itself or how that data was collected or generated.

The main data types that were used in this research are; passport data, pedigree data, phenotypic data, genotypic data and finally annotation data derived from experimental work on barley. Pedigree and limited phenotype data exists for wheat (*Triticum spp.*) and Asian rice (*Oryza sativa*) which was used to test the system for alternate species.

2.6.2 *The Germinate data warehouse*

The Germinate data warehouse ("Germinate 3" 2014) was developed in order to hold passport, genotypic, pedigree and phenotypic data which formed the foundations of this work. Germinate was developed using MySQL and uses Perl to interface with a custom web interface a data services application programming interface (API) which was used to retrieve data.

The development of the Germinate data warehouse facilitated the integration of heterogeneous datasets. In addition to the barley data used in this work, Germinate instances have been implemented to store genotypic, phenotypic and passport data for a range of other species including pea, rice, ryegrass, maize, potato and wheat.

The reason that Germinate was used was to ensure that researchers working on the barley data would all be using the same data from the same source. This is particularly important as a means of trying to help reduce errors and to ensure researchers are working on the most up to date datasets available; something which is an important and often overlooked problem in modern experimental plant science.

2.6.3 *Experimental datasets*

Data was required in order to efficiently model and test any resulting database schema. The datasets available covered both state of the art and legacy datasets. The availability of large amounts of high quality data was important in the development of any system as it provided the facility to accurately gauge performance when dealing with real-world data and data volumes. Data provenance is critical in experimental genetics as the source of genetic material is important for subsequent experimental work and sourcing the exact germplasm used in experiments. These datasets ranged from thousands of data points to tens of millions in the case of high-throughput genotyping dataset and are detailed below.

2.6.3.1 Pedigree definition data

Pedigree definition file for 803 UK Elite barley cultivars (both spring and winter ecotypes) most of which having gone through National List trialling in the UK at some point over the past 20 years and are therefore grown commercially in some capacity. This data was collated from a number of data sources including published pedigrees, expert in-house knowledge and from online resources including the Science and Advice for Scottish Agriculture's barley variety database ("The Scottish Barley Variety Database" 2014) the HGCA (HGCA 2014) and Lfl Pflanzenbau in Germany ("Bayerische Landesanstalt Für Landwirtschaft" 2014). This data was collated as part of this work and new data is available every year as new varieties are released and put through the trialling process.

The nucleus of pedigree data are a series of parent/child relationships defined as encoded strings (Lamacraft and Finlay 1973; Purdy et al. 1968). Data was atomised into simple parent/child definitions which were used to dynamically reconstruct the pedigree. In addition there may also be information identifying whether the parent was male or female and the type of genetic cross performed. Something unique in plant breeding is where a plant can be both male and female parents in the same cross so parental assignation can be important.

There were however complications which arose from older pedigree data which was error prone and is difficult to verify without expert guidance. These problems included the re-use of names to describe varieties which resulted in the creation of false relationship joins and typographical errors. It is not uncommon for a breeder's favourite name to be used multiple times until a plant line is adequately different, and has sufficient performance to be accepted for wider distribution into the UK recommended list programme. The current accepted way of using these names to name samples can therefore lead to confusion and a better system needs to be adopted to try and prevent errors being introduced. It is highly probable that there are errors in datasets where a plant lines name has been used more than once and the incorrect version used by accident. While this problem is diminished if the variants are morphologically different this is not always the case and staff handling samples may not be trained in the identification of subtle morphological variation between varieties.

The data was also inconsistent. Depending on the source of the information the formats varied which meant that each data source needed to be interpreted then represented in a standard format for inclusion into the database.

The pedigree data used in this work was broken down or atomised into its constitutive components, insomuch as each plant line exists as its own entity with two parents. One of the major underlying concepts behind this work was to ensure that the system and methods chosen to store pedigree data was both as flexible as possible and simple as possible in order to diversify and increase its potential use and acceptance in the community. At the most basic level an entity has one or more different parents. Pedigree strings were split into the equivalent of key/value pairs as follows;

Assume the pedigree for the 2-row spring barley Quench. Quench has the pedigree ***Sebastian * Drum*** (in Lamacraft notation). In this work, the pedigree for the plant line Quench is stored in the pedigrees table in the Germinate data warehouse as:

Quench	Sebastian	F
Quench	Drum	M

A slightly more complex example for the 2-row spring barley Puffin shows this in more detail and introduces the concept of what has been called '*intermediate plant lines*' for the purpose of this work. Puffin has the pedigree **Maris Otter * (Athos * Igri)** which is stored as:

Puffin	Maris Otter	F
Intermediate_1	Athos	F
Intermediate_1	Igri	M
Puffin	Intermediate_1	M

Intermediate plant lines (Intermediate_1 in, and unique to the above example) are plant lines for which it is known there has been a crossing of two parent plant lines but the resulting progeny that was used as a parent in a subsequent cross is not known. I.e. there is not a defined name nor identification code for it and only the parents are known.

It is standard nomenclature for the female parent in any cross to appear first. However, this information is not always reliable especially with older datasets. Using the system detailed here highly complex pedigrees can be built up from repeating child/parent pairs. This structure also ensures that the database can easily scale vertically as the datasets increase in size over time.

The use of this simple plant line/parent method of storing pedigree definitions is the simplest format into which a pedigree can be broken down and provides the greatest flexibility for the storage of pedigree data in a plant context.

A simple Java command line application was written which takes data in Lamacraft and Purdy notations and parses this into the defined pedigree notation. The pedigree parser takes a string and tries to atomise it into the format described above. Any pedigrees that do not match this format are highlighted and can be dealt with manually. The program automatically creates intermediate crosses in a format compatible with Germinate.

2.6.3.2 Phenotypic data / Nominal / Categorical DUS Data

A DUS (Distinctiveness, Uniformity and Stability) dataset comprising detailed records for 34 phenotypes used to determine if varieties are sufficiently different from one another to be allowed into the UK recommended list system (RL) was used in this work. This data was obtained from the National Institute of Agricultural Botany (NIAB) in Cambridge and has been used in association analysis (Cockram et al. 2010; Wang et al. 2012). These characters, many of whose genetic basis have been experimentally derived, are used to differentiate new plant lines and are used by breeders as a reference to ensure new varieties are different from currently accepted varieties. This data is also used to maintain reference stocks and verification of VCU (Value for Cultivation and Use) submissions, which are entered into the National List (NL) and Plant Breeders Rights (PBR) schemes. As breeders use these datasets to breed new plant lines they are important in plant breeding in the UK.

The DUS data was comprised of 12 nominal and 22 ordinal data categories (Appendix 1) across 581 plant lines/varieties from the test pedigree resulting in 16,211 data points in total.

Phenotypic data relates to measurements or morphological or physiological/biochemical characteristics that can be identified in a single plant line. These are genetic characteristics/traits under the control of genes or genes and the environment. Phenotypes take the form of integer, float based measurements or a text description of the character being defined.

The phenotypic data in this study has been either collected in field experiments or by molecular testing. Though many of the agriculturally important traits are controlled by many genes of small effect (quantitative traits) for simplicity this work concentrates mainly on traits under simple genetic control. Examples of such traits include some of the DUS characteristics which are used in the varietal registration and seed certification process and allele data on disease resistance genes such as *Mlo* (Jorgensen 1992; Thomas et al. 1998; Buschges et al. 1997) and *Mla* (Mahadevappa, Descenzo, and Wise 1994; Wei, Wing, and Wise 2002).

2.6.3.3 Plant passport and background data

Plant lines often have passport data associated with them but it is frequently sparse and incomplete. Passport data describes information relating to the storage and naming of a plant line. This can include data such as alternative identification credentials, information on the breeder of the plant line and information about the gene bank from which the seed was sourced. It can also include collection site data including geographic coordinates of where the original germplasm was collected if available. Passport data is relatively well defined in terms of the definition of characters recorded but is often scant in nature and there is a greater acknowledgement and work focussing on how to try and improve the quality of these sorts of data in gene banks (van Hintum, Menting, and van Strien 2011; Thormann et al. 2012) both by performing data cleansing on existing data and by inclusion of older paper-based records into digitised form to complement data already held.

The current standard for gene bank data is the guidelines set out by the FAO/IPGRI (Food and Agriculture Organization of the United Nations / International Plant Genetic Resources Institute) and adapted by projects such as Eurisco ("EURISCO" 2014), Genesys ("Genesys PGR" 2013) and Germinate ("Germinate 3" 2014) and takes the form of a series of what are termed Multi Crop Passport Descriptors (MCPD). The idea behind the standard was to allow a defined set of data to be stored for all crop species. The latest update to the standard was in June 2012 ("FAO/Bioversity Multi-Crop Passport Descriptors V.2 [MCPD V.2]." 2015) but this was mainly to clarify how missing data should be stored and to change the curators of the standard from the International Plant Genetic Resources Institute (IPGRI) to Biodiversity International ("Biodiversity International" 2015). Other updates such as the removal (and addition) of countries which have now been succeeded such as East and West Germany and Yugoslavia. A more detailed description of the descriptors is covered by Faberova (2010) although this predates the v2 2012 release (Faberova 2010).

2.6.3.4 Genotypic data

Genotypic data defines the genetic basis of a plant line. Depending on the technology these can take the form of integer or float based values such as amplified fragment length polymorphism (AFLP) or microsatellite based technologies. Others are text strings representing a nucleotide base in high-throughput single nucleotide polymorphism (SNP) based assays. In this work only marker based assays and not genomic sequence data are utilised.

A given plant variety will have an allele call for each of a series of loci represented as a pair of nucleotide bases e.g. AA, GG (which are homozygous) or AG (which are heterozygous), for a locus (one from each chromosomal strand). Due to the inbred nature of the barley germplasm there are low levels (less than 0.5%) of residual heterozygosity present, this is not the case in humans and most domesticated animals.

Current genotyping technologies output data, based on raw base calls at a genetic locus. This can either take the form of A, C, G, T or a failed call which is usually, although not exclusively described with the symbol '-'. Older technologies also include the generation of fragment lengths, which take the form of an integer based number, usually of known sizes that lie within defined parameters. Data can also be stored in AB format, which is coded based on the call against reference genotypes.

The primary genotypic data set is composed of a large barley pedigree data set for 803 UK Elite cultivars as well as Single Nucleotide Polymorphism (SNP) genotypic data for 750 of these plant lines across 4,769 genetic markers.

The SNP markers were mapped to known chromosome positions in the barley genome. Each plant line within the test set has been genotyped for a set of 7,842 SNP assays of these markers. 2,832 of these were derived from previous oligo pooled assays and 5,010 derived from next generation sequencing data (Cockram et al. 2010; Comadran et al. 2012).

This set of SNPs was derived from Illumina ("Illumina BeadArray Microarray Technology" 2015) RNAseq reads mapped onto Harvest35 reference sequences ("HarvEST" 2014). The raw Illumina reads were a mixture of 54 nucleotide and 76 nucleotide reads and derived from the cultivars Barke, Betzes, Bowman, Derkado, Intro, Morex, Optic, Quench, Sergeant, and Tocada.

Due to inherent failures resulting from failed mapping, failed SNP assay chemistry and markers which are not polymorphic, the actual total is reduced to 4,769.

2.6.3.5 Annotation data

Annotation data is information that is assigned post data collection by domain specialists and is important in the curation of large data sets and resources. Annotation data is text based and can be added by users as required. It can be viewed as meta-data used to add knowledge to an individual data item. Annotations can be defined against each of the main data classes, genotypic, phenotypic and passport data. They mainly take the form of comments made by breeders about a variety, which frequently are not part of the trialling or listing processes or requirements. While the volumes of this sort of data is currently low it is envisaged that the development of tools to aid in the visual analytics of these sorts of data will lead to increased volumes over time.

2.7 **Initial requirements gathering**

In order to establish the requirements for this work a questionnaire was used to obtain feedback from potential users of a pedigree visualization system to try and understand the kinds of tools and features that geneticists and plant breeders would find useful. The aim of the questionnaire was to determine the kinds of questions that a user would want to be able to answer. The results of this preliminary requirements gathering exercise was used to help develop a plan for the implementation of a visualization system.

The initial questionnaire (Appendix 2) was carried out with 6 individuals. Three of these defined themselves as geneticists and 3 plant breeders, one of whom was a representative for a large plant breeding company. The questionnaire was designed to establish the volumes of data that would need to be routinely dealt with by any visualization system that was developed and important features required by target users of a visualization system. The questionnaire was split into two sections; a background section which was used as a profiling tool to categorise the respondent and establish data volumes that they routinely deal with and a functionality section which was used to identify the main questions that the test subject thought were important to their work. The questionnaire was also followed up with face-to-face meetings to obtain general feedback and ideas about the problems that exist with their pedigree data.

2.7.1 *Initial requirements gathering results*

The results obtained showed that 4 out of the 6 respondents indicated that they use more than 100 but less than 1000 lines. One used more than 10,000 plant lines routinely.

The results also showed that 5 out of the 6 respondents indicated that they use in excess of 1000 but less than 10000 genetic markers. The other response was that they used less than 100. Results to determine the number of phenotypic scores that the test users thought they would be required to handle were 3 responses indicating up to 1000 and 3 indicating between 10,000 and 100,000.

Feedback from users showed that they wanted to be able to trace lineage of specific plant lines which were deemed important and be able to quickly determine specific characters of these plant lines. They also wanted to merge data so that links between varieties which may not be obvious are easily identified. Users additionally wanted to be able to overlay phenotypic data so that they could quickly tell plant lines which exhibited a specific character of interest. They wanted to be able to quickly identify a plant lines lineage and show both ancestral and descendant lines for a chosen plant line.

There was also an indication that users wanted to be able to perform more advanced statistical analysis on the underlying data, such as looking at genetic similarity or identifying haplotypes which may be responsible for an expressed phenotype. They also indicated that being able to access background information on plant lines and phenotypic characters was important as these are often complex and having them in a single location that could be used as a reference would be beneficial.

The testing showed us that all the users thought that data visualization (the term data visualization is used here but can equally apply to information visualization as well due to context) was important to the work they do and all believed that community interaction and engagement was useful in their work. It was also clear from the commercial side that breeders wanted simple easy to use tools that isolated them from the complexity of underlying stats but presented them what they needed to know in a simple and intuitive way.

Other results showed that while a public repository of pedigree data would be a useful feature there were reasons why data needed to be kept private.

Finally users identified the ability to export data in formats suitable for further analysis was important in any analysis system.

The data types used for this study are complex. By their nature biological systems are complex and it is important that users’ are allowed to view as much background information as possible along with the pedigree definitions to aid in data analysis. Any visualization tool needs to allow users to quickly and accurately identify plant lines which have particular characters which are unique within a dataset.

2.8 Discussion

Changing environmental conditions mean that plant breeders are looking towards the development of new varieties which have agriculturally important characteristics such as drought tolerance and pest resistance. In order to develop these new varieties existing germplasm can be examined to identify potential plant lines which have desirable characteristics which can be used in breeding programmes.

There are a number of data types which are routinely used including genotypic and phenotypic data and the ability to explore this data in a pedigree framework, which details the genetic relationship between individuals, will be a valuable tool for the development of new varieties. There is however no single standard nomenclature for plant pedigree data.

An initial user survey was carried out to identify two things. Firstly it aimed to identify the volumes of data that both breeders and plant geneticists were using, and expected to be using, and finally the sorts of functionality that they would like to see in such a tool.

The results showed that the volumes of data that were expected to be used were in the thousands with 4 out of the 6 respondents indicating that they routinely use less than 1000 plant lines and 2 indicating that they use more than 1000 and one more than 10,000 plant lines. The respondent that indicated more than 10,000 plant lines was referring to every plant line used in a breeding programme and not just the ones which were selected for use, or to be kept going forward. For this reason it was decided to focus on ensuring that anything that was developed was able to handle plant lines in the thousands of individuals scale. It was also clear that there was a need to allow the overlaying of data and the storage of additional meta-data on plant lines and phenotypic characters that could be easily retrieved in the context of the pedigree that users were working with.

The feedback from the initial user survey indicated that users wanted to be able to browse complex pedigrees and identify lineage (both parents and children) and overlay additional data so that the structures could be examined to identify potential patterns or to identify plant lines with unique characteristics.

Chapter 3 Biological data visualization

3.1 Overview

This chapter details why visualization is important in modern biological science, with particular reference to plant pedigrees, phenotypic and genotypic data. It then discusses the problems associated with visualizing large biological datasets and surveys the tools that have been developed to visualize biological data, with particular reference to pedigree visualization; evaluating why these are not appropriate for the needs of this research.

3.2 Information overload

While the term information overload is perhaps now seen as a cliché, in the domain of modern plant genetics there has never been a more appropriate time for its use. Modern high-throughput technologies are routinely producing many millions of data points for individual experiments, a data quantity which has been, until recently, unheard of.

Modern molecular biology is advancing at an ever increasing pace. The data being produced by next-generation sequencing and genotyping technologies is advancing faster than Moore’s Law (Sansom 2007; Moore 1965) and is presenting scientists with major challenges. As new technologies become available the types and volumes of data available are in a state of constant flux. Institutes such as the European Bioinformatics Institute (EBI) in Hinxton UK, one of the world’s largest biology related data repositories, as of 2013 had in the region of 20 petabytes of data of which 2 petabytes is genomic information. This figure is predicted to now double every year (Marx 2013).

Lathe (Lathe, Williams, and Karolchik 2008) showed that while DNA (Deoxyribonucleic Acid) sequence databases were increasing in size, in terms of data volumes, at an ever increasing pace the number of databases available for researchers was also increasing at a similar rate. One reason for this increase in data is the cost of its production (Hayden 2009). Large volumes of data are much cheaper to generate now than they have been with the development, and constant refinement, of third generation high-throughput DNA sequencing and developments in phenomics. Because of these increasing data volumes, techniques need to be identified which allow researchers to make informed choices when dealing with their data. These techniques include methods to examine data quality and looking for unusual data or patterns which can be either an indication of problems or a pointer to interesting biological findings.

The ability to visualize large and complex data sets and the genetic relationships between germplasm, such as those discussed earlier, will allow meaning to be gained from the underlying data which can otherwise go undetected. While there are tools to identify patterns and potentially interesting nuances with large datasets, there is often no substitute for expert user interaction and visual analytics particularly in biology where exceptions to rules and patterns are often the most interesting results.

3.3 **Visualization**

Visualization has been described as a way of transforming the symbolic into the geometric enabling researchers to observe simulations and computations and a way of seeing the unseen and a way of gaining information from a sea of data (McCormick, Defanti, and Brown 1987).

Visualization communicates information by abstract representation of data. The use of visualization to convey information is not a new idea (Marchese 2011). Its use dates back thousands of years in domains such as cartography, cave paintings and more recently William Playfair’s first use of what are now common chart types. These included line (first seen in 1786) (Figure 3-1B), bar (1786) (Figure 3-1A) and the hotly debated pie chart (1801) (Figure 3-1C) which were used to convey information on trade and shipping. Playfairs’ charts brought visualization to the wider public audience.

It is widely accepted that the advent of computer graphics has revolutionised visualization in the science, computing and engineering disciplines.

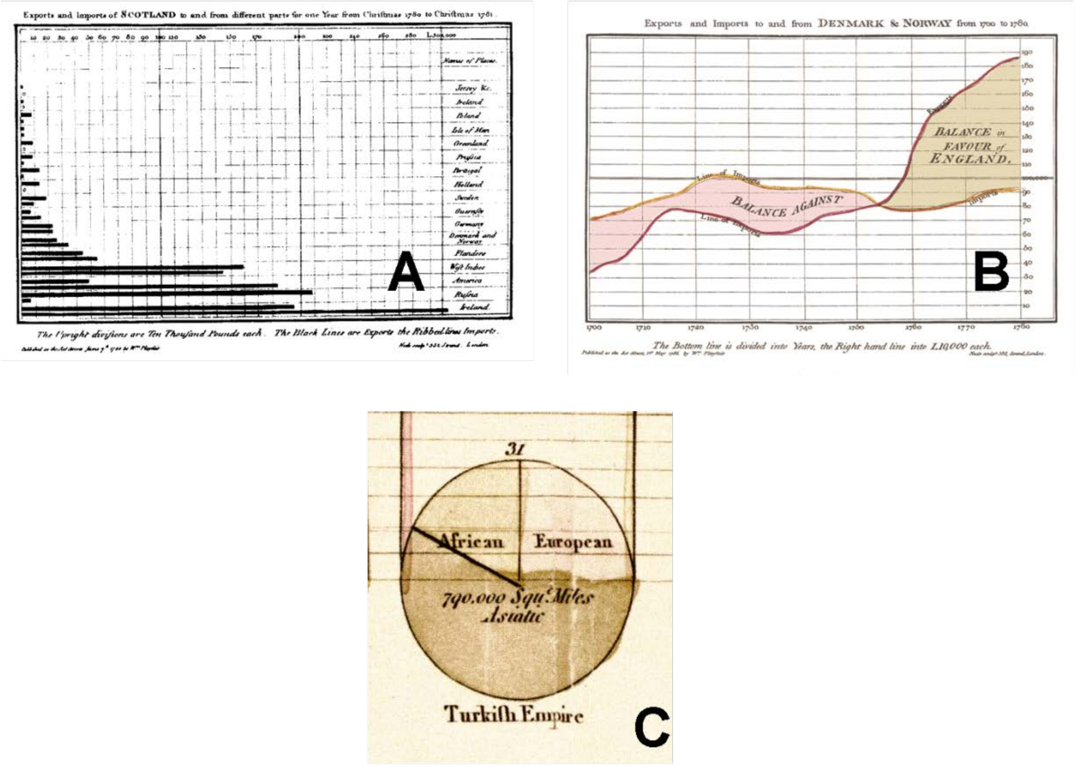


Figure 3-1 William Playfair charts

3.4 Visualization definitions

There are a number of main visualization sub-categorisations. These include, but are not limited to; scientific visualization, information visualization and finally biological visualization. The following sections describe these discrete areas and how they are related to one another.

3.4.1 *Scientific visualization*

It was not until 1987 (McCormick 1988) that the mainstream use of the new area of scientific visualization using computational derived graphics was accepted. This is reported to have brought the techniques that could be offered by computer graphics experts to the attention of the mainstream scientific computing and scientific organisations as a tool to aid in the understanding of data.

Scientific visualization was as late as 2004 still being described as a relatively new discipline (Johnson 2004), this is somewhat surprising. Johnson suggests that because of this, the discipline had a number of assumptions and adopted practices that needed to be inspected in order to bring it under the same scrutiny and rigour afforded by the more traditional scientific research areas. He also suggests that one the main problems currently unsolved in scientific visualization is a lack of understanding of the underlying science, something which is true of bioinformatics in general, and integrating scientific and information visualization.

One of the first scientific visualizations using modern 3D rendering was performed in 1981 by Nelson Max from the Lawrence Livermore University where a series of images of the molecular structure of DNA were collated into the short film '*DNA with Ethidium*' ("DNA with Ethidium" 1978). The ability to view structures in 3D has led to increased biological knowledge in a number of areas in the life sciences.

While Michael Friendly (Friendly 1995) describes scientific visualization as being primarily concerned with the visualization of three dimensional phenomena and tries to create realistic renderings there appears to be dispute over what the term it actually embraces.

Friendly suggests that statistical graphics applications such as scatter plots should indeed be termed data visualization and not scientific visualization, this however is challenged by definitions by Mann (Mann et al. 2002) and Johnson (Johnson 2004) who appear to move in-and-out of the data visualization definitions with their descriptions of scientific visualization. This clouding of the definition between scientific and data visualization is commonly seen.

3.4.2 *Information Visualization*

Information visualization (also termed InfoVis) was first described in the seminal paper (Card, Robertson, and Mackinlay 1991) in 1991 and was described by Card et. al. as 'methods and machines that would allow people to bring to bear on a task of interest more information more quickly than otherwise possible.'

Other definitions include describing information visualization as the visual representation non-spatially defined data, such as the representation using visual metaphors of phylogenetic trees (Ruths, Chen, and Ellis 2000), biological networks (Pavlopoulos et al. 2008) or human networks derived from social networking sites. What's important is the representation of the relationships connecting each entity, and the information that brings.

Card (Card, Robertson, and Mackinlay 1991) puts forward the idea that the goal of all information processing systems is to minimise the cost structure of information processing. Card also shows that information systems consist of multiple levels of information storage or abstraction which are available to a processing system for generation of user visualization. The paper describes how maintaining levels of abstraction allows users with varying technical abilities to use the data based on their existing knowledge. The more abstract the information visualization the simpler it should be for cognitive processing of the data or information (Figure 3-2).

What is important in the examples above is the high level of user interaction required in order to manipulate the visualization to gain a deeper understanding of the structure of the abstraction and visual metaphors it contains. The visualization of additional data that is available as a user interacts and analyses specific regions in the visualization means that only directly relevant information is presented to the user thus removing problems associated with information overload.

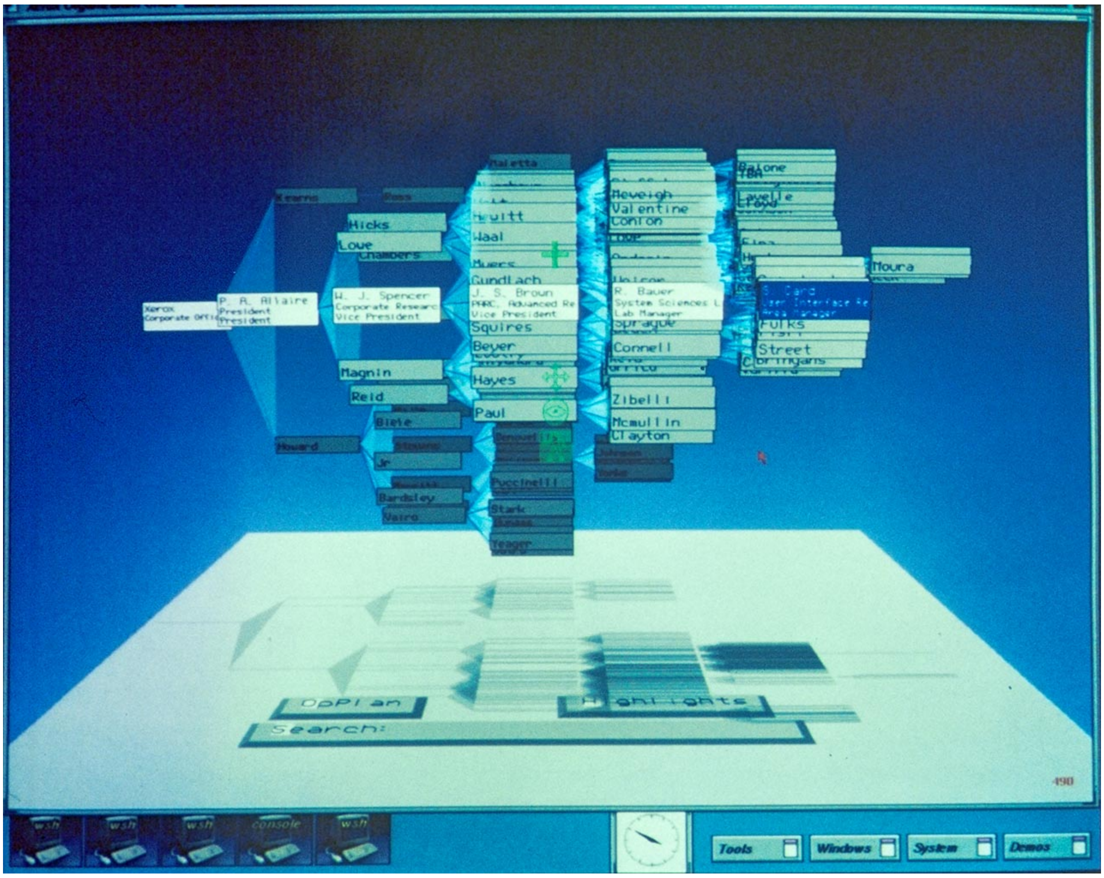


Figure 3-2 Information visualization by Card (Card, Robertson, and Mackinlay 1991)

Examples of information visualization with regards to biological data include Mizbee (Figure 3-3) and Circos (Figure 3-4). Both applications show similar information visualization approaches by two different research groups. These two visualizations create metaphors of chromosomes to show synteny, which is defined as the conservation of blocks of genetic elements between chromosomes which can be compared to one another, often between species. The physical representation of a chromosome is arranged radially and synteny visualized using connecting arcs between chromosomes or regions.

The representation of a chromosome as a linear line along which blocks of colour are used to define regions or loci is an over simplification of a chromosome which removes much of the complexity of the biological molecule but it is a metaphor which is easily understood by biologists. Also, the circular nature that both Mizbee and Circos use to represent chromosomes is not how such a DNA molecule exists in nature but instead allows a greater density of data to be represented in a page.

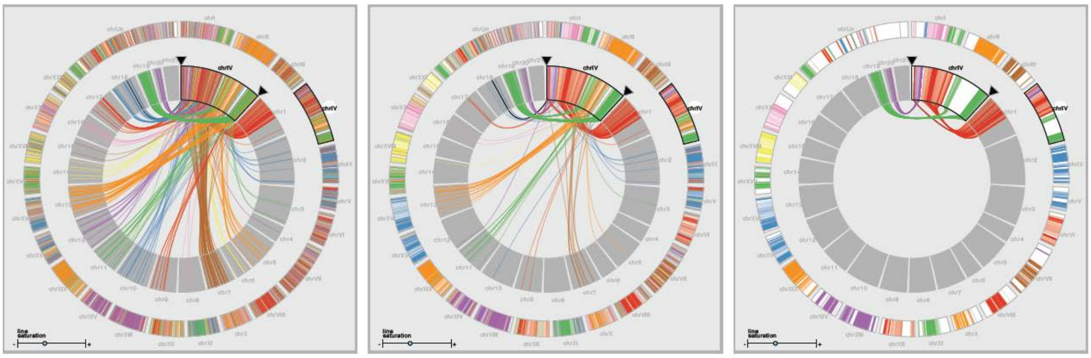


Figure 3-3 Mizbee (Meyer, Munzner, and Pfister 2009)

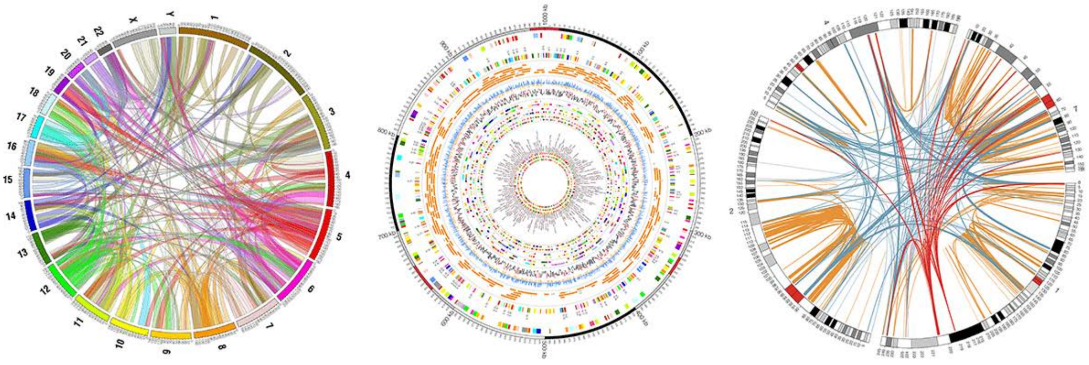


Figure 3-4 Circos (Krzywinski et al. 2009)

3.4.3 *Differences between scientific and information visualization*

Rhyne (Rhyne 2003) asks why information and scientific visualization must be treated as separate entities and discusses whether seeing them as separate entities provides a mechanism to advance both fields or cause confusion; especially in bioinformatics. The reference to the two as ‘separate but equal’ by Rhyne lends itself to the argument that the dichotomy between them can be relatively narrow. This idea is strengthened by Tamara Munzner who hypothesises that the terms may merge within the decade (Rhyne et al. 2003), although this has not happened.

Scientific visualization (Figure 3-5A) can be distinguished from information visualization (Figure 3-5B) in a number of ways. Scientific visualization deals with quantitative data which has some representation of location in space and tends to follow a normal distribution while information visualization is more qualitative and abstract in nature. Information visualization represents non-numerical information where the links that exist between data points are critical in defining the quality of the underlying data. Without the information contained in the links between data points the data points are of little value. The same cannot be said for scientific visualization systems whereby individual data points might not be linked to each other but offer insight in to processes and trends. While spatial position is already determined in scientific visualization, in information visualization it needs to be decided.

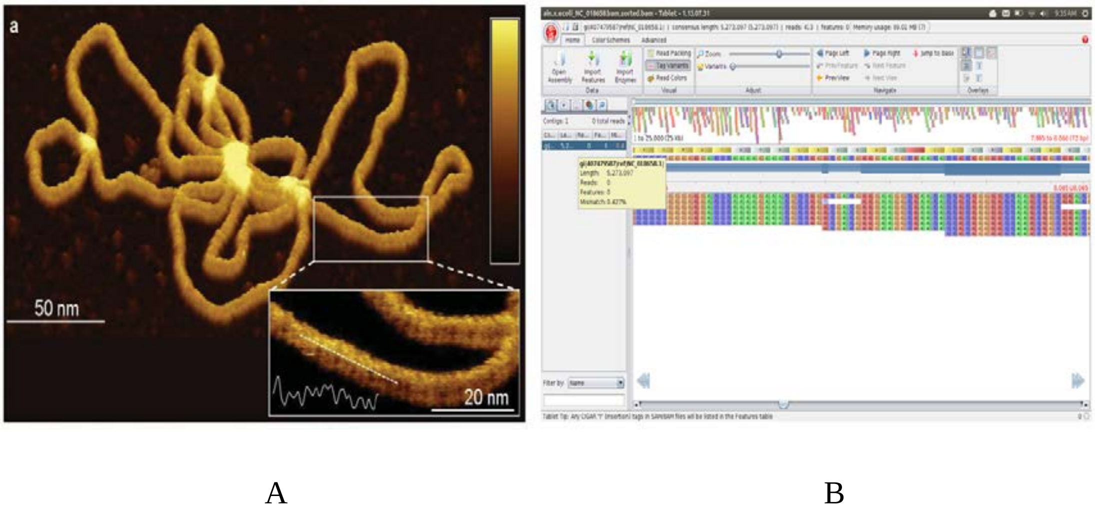


Figure 3-5 Comparison between A) scientific (Pyne et al. 2014) **and B) information visualization** (Milne et al. 2009)**.**

Whereas scientific visualization deals with raw data, precise and unambiguous in nature, information visualization is defined by Gershon (Gershon, Eick, and Card 1998) as combining aspects of scientific visualization, human-computer interfaces, data mining, images, and graphics and deals with data which is often abstract in nature.

Where these two techniques do find parity is in their ultimate goal of manipulation of the visual representation of data so that new or increased information can be obtained through visual data mining techniques.

Although there is clearly overlap between information and scientific visualization the paper by Robertson (George G. Robertson, Jock D. Mackinlay, and Stuart K. Card 1991) is regarded as one of the first to try and identify the differences between them.

3.5 Biological data visualization

Biological visualization straddles both the information and scientific visualization domains. The degree of overlap depends on the exact problem being investigated. Recent work (O'Donoghue et al. 2010; Gehlenborg et al. 2010; B. Wong 2012) has begun to firmly place biological visualization as a defined field within a specific problem area, albeit a very large, ill-defined and diverse problem space.

Biological data visualization encompasses information visualization, scientific data visualization and visual analytics; defined as 'the science of analytical reasoning facilitated by interactive visual interfaces' (Thomas, J.T., Cook 2005). It is a wide and diverse genre with examples ranging from the representation of the neural map of *Drosophila melanogaster* where networks are presented on top of a 3D volume rendering of the fruit fly brain (Sorger et al. 2013) to the representation of microarray time series data in the MaTSE application (Craig et al. 2013). These examples show not only the diversity of techniques in biological visualization but also the different visualization domains that both draw from.

While the modelling of shape of for example brain or a molecular model can be achieved in 3D space and appear recognisable the depiction of complex biological processes at the cellular level are often abstracted and have no semblance to the actual real life process.

Biological visualization therefore brings together scientific visualization, information visualization and visual analytics into an area to address the specific problems associated with visualization and analysing biological data.

3.1 Visualization techniques

3.1.1 *Detail and overview*

Detail and overview systems are defined as systems whereby the concurrent display of both an overview and detailed view of a visualization are present with each representing a distinct (concurrent but spatially segregated) representation of the same underlying data (Andy Cockburn, Karlson, and Bederson 2008). Common desktop examples of such systems include applications such as Microsoft PowerPoint that show both the selected slide and a series of thumbnails giving an overview of the slides within the presentation and Google Maps which shows both the map area of interest plus a thumbnail representation in the corner, often using an alternative map overlay system. Cockburn also addresses some of the issues with such systems such as the acceptance that there will be a reduction in visual clarity as overviews are scaled in size. There is only so much information you can display on a small thumbnail. It is common for synchronisation between the overview and detail views to be one-way. This prevents problems arising from people clicking on overview areas and changing the data contained in the detail panel.

3.1.2 *Zooming*

Zooming is a common feature of visualization systems and allows the user to navigate through large information spaces which would not fit easily within the comfortable desktop area. Whereas the detail and overview type systems overview gives a ‘zoomed out’ representation of the detail window, zooming techniques usually allow the user-defined zooming of the information space. Zooming often goes hand in hand with scrolling and panning techniques (van Wijk and Nuij 2003) to allow the user to explore the visualized data. Work by Cockburn (A. Cockburn and Savage 2003) has examined the problems associated with zooming in user interfaces by utilising automatic zooming techniques with maps.

3.1.3 *Filter*

Filtering is an important technique in visualization as a means to try and remove data which may not be immediately required in the information space. Shneiderman (1994) suggests that with the increase in information and data being visualized filtering techniques can aid in finding information of interest in large visualizations, allow efficient traversal of large datasets and increase comprehension (Ahlberg and Shneiderman 1994).

Filtering can be achieved in a number of ways including the use of pointing systems whereby areas of interest are selected and highlighted with non-selected data removed or made less prominent or the creation of filter interfaces whereby data is selected based on queries across the dataset using techniques such as slider bars which can be altered in real-time.

3.1.4 *Focus and context systems*

Focus and context systems differ from overview and detail techniques in that the information is shown within a single information space. They offer selective zooming of data using techniques such as fisheye lens views whereby there is a continuous and seamless integration of zooming effects which makes the selected areas more prominent and make information outside of the focus less prominent. Focus and context techniques are often referred to as distorted views.

3.1.5 *Coordinated multiple views*

Coordinated multiple views (Andrienko and Andrienko 2007) is a visualization technique whereby different techniques are employed to create a variety of alternate views on to the underlying data in order to try and gain a deeper insight into the dataset. These views are linked so that a change in one leads to the appropriate changes in the other views. In the biological context, examples of this could include the representation of a pedigree image with views showing phenotypic and genotypic data for a specific plant line then how this data sits in relation to other plant lines in the dataset by means of a scatter plot. Changing a selected plant line would result in a change of the other information views to reflect this.

3.2 Visualization process

The process of visualization, which has formally been described and represented as a state reference model (Chi 2000) describing a logical flow from data to visualization, typically involves the filtering of raw data to select points of interest then rendering this data in a suitable format whether that is a static image, animation or any other means by which a greater understanding of the data can be achieved by abstraction. Erroneous data points, trends and data features which would otherwise be invisible within the raw data sets can be subsequently identified.

Fry (2007) proposes 7 stages to data visualization: acquire, parse, filter, mine, represent, refine and interact (Fry 2007). Fry also suggests that when each component is undertaken by a different person, problems from a 'Chinese whispers' type effect can surface with information being lost at each stage. This shows the importance of an integrated system. Fry also shows that while data is important, what is also equally important is what is *left out* in a particular visualization in order to gain clarity in data. It is also important to note the differences between the stages of information and data visualization as both perform different functions. Fry's stages can be represented as a pipeline and not stages of visualization in the strictest sense while Shneiderman states that the stages of information visualization are overview first, zoom and filter, and details on demand (assuming data will be loaded once into memory then used, this is also known as Shneidermans' Mantra) (Shneiderman 1996). This however is not always appropriate and any visualization may have to involve the use of drilling-down and associated interactive techniques. Interaction is imperative in visual analytic tools as they are not just static infographics; which are limited in their ability to visualize and filter large datasets, but changeable visualizations handling different data types and combinations.

One apparent weakness of these formalised definitions of visualization process is that they are linear in nature, it would be reasonable to assume any visualization development process would require development loops.

Modern biological data can be obtained both quickly and in large volumes, but is not *necessarily* of high quality. This in turn requires efficient visualization in order to maintain data quality by means of visual data-screening and to present data to users in an abstract but meaningful way. While automated methods exist to perform such functions they cannot replace human interaction and intervention. There has never been a more appropriate time in biology than now for the development of new techniques to visualize and allow the analysis of data and bring as much biological understanding as possible from the ever increasing volumes of data being generated.

3.2.1 *Visualization in molecular biology*

The role of visualization in the understanding of scientific and in this case biological data is well documented (Pook, Vaysseix, and Barillot 1998; McCormick 1988; Domik 1991; Mann et al. 2002; Merico, Gfeller, and Bader 2009; Eavenson et al. 2008) in areas ranging from genomic data (Pritchard et al. 2006; Carver et al. 2009), sequence assembly (Milne et al. 2009; Milne et al. 2013), expression (Kestler et al. 2005), and phylogeny (Ruths, Chen, and Ellis 2000; Sanderson 2006). It is crucial to further understanding in areas where data volumes and complexities continue to advance. While each of these tools vary in the level of abstractness that they represent from scatter plots showing raw data to DNA assembly software using coloured blocks or visual encodings to represent individual nucleotide bases they all are common in their focus to better represent the complex data underlying them by creating a recognisable visual metaphor.

The visualization of molecular data uses a wide variety of diverse techniques. Hahne (2012) describes various visualization methods in the understanding of the underlying data (Gentle and Hardle 2012). These techniques range from scatter plots to the use of density estimations to show data concentration and describing the issues surrounding colour in the representation of data. In Figure 3-6 each plot uses a different technique to visualize data from black dots through kernel density hexagon binning, smooth point density and smooth point with black dots where low density data exists.

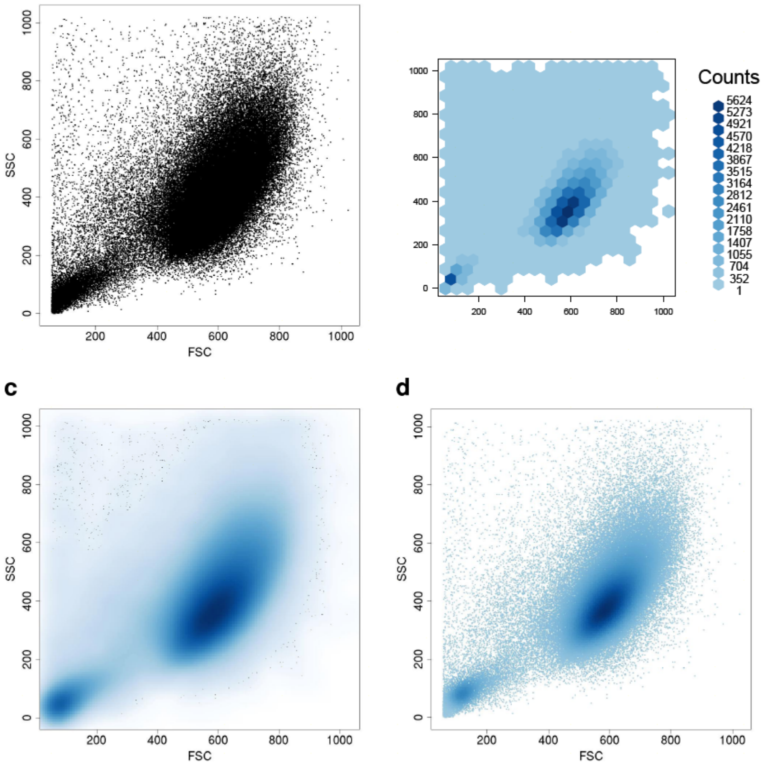
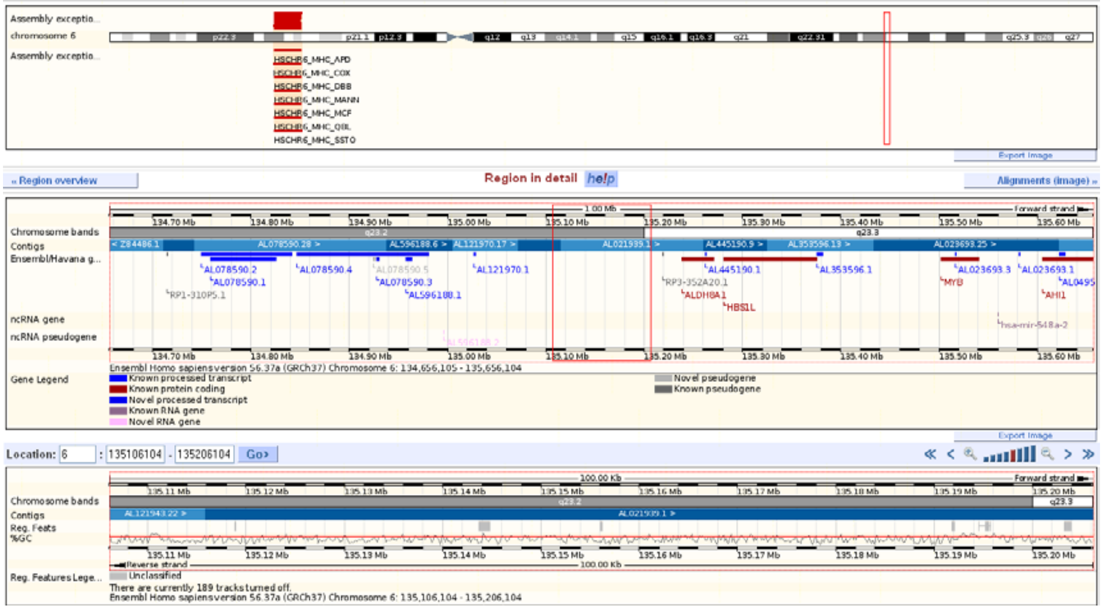


Figure 3-6 Hahne's genomic visualization (Gentle and Hardle 2012)

Other examples of visualization tools within the molecular biology domain include the genome browsers Ensembl (T. Hubbard et al. 2002; Kersey et al. 2010; E. Birney et al. 2006)(Figure 3-7), Genome Browser (D. Karolchik et al. 2003; Rhead et al. 2010) and Viz Genome (Jakubowska et al. 2007) (Figure 3-8) where scientific data is merged with the creation of visual metaphors of genomic regions of a chromosome to aid in the conceptual positioning of the data, in a way that is not representative of how it would appear in real life.



Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 6: 134,656,105 - 135,656,104

Location: 6 : 135106104 - 135206104 Go»

There are currently 189 tracks turned off.
Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 6: 135,106,104 - 135,206,104

Figure 3-7 Ensembl visualization of a region of human chromosome 6

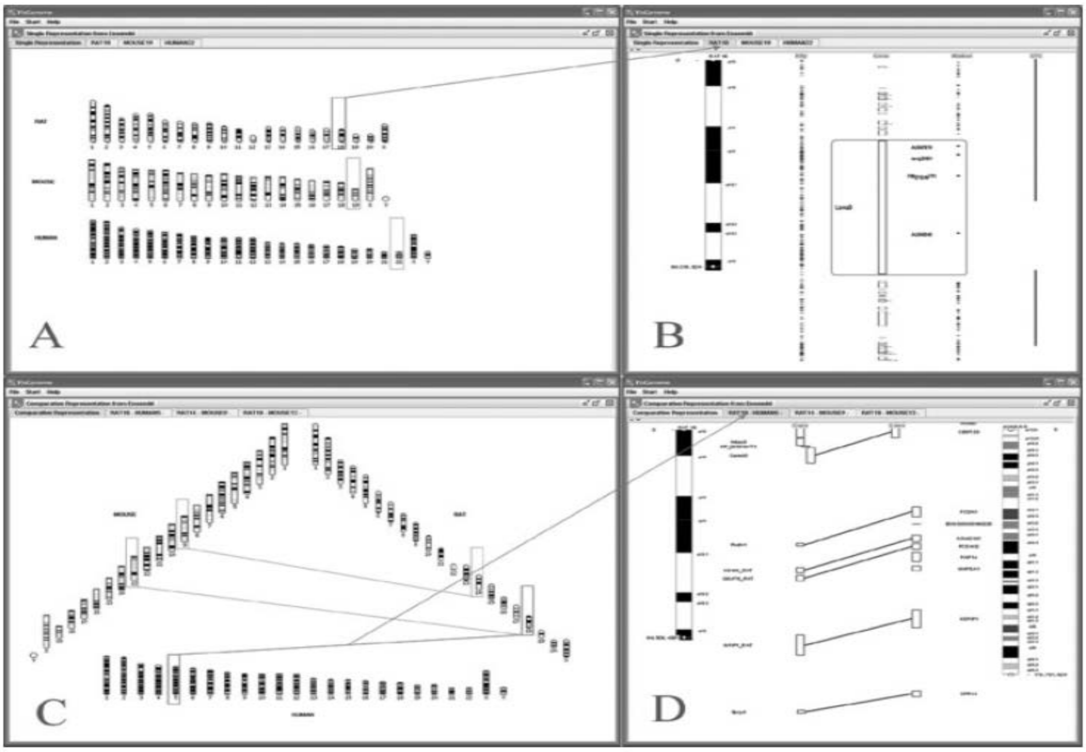


Figure 3-8 Vis Genome visualization representing karyotype image and zoomed genomic regions

3.2.2 *Genotype visualization*

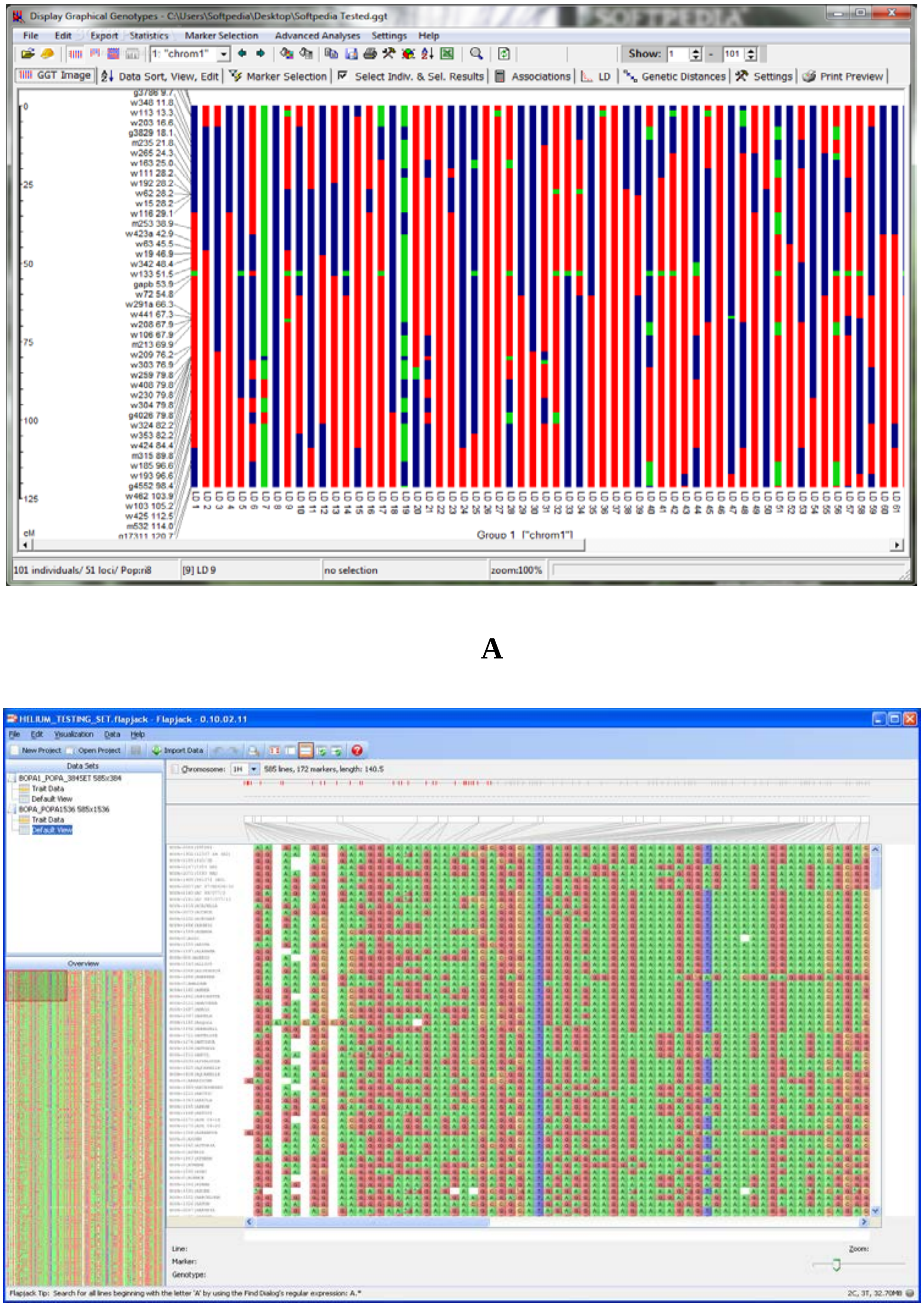
The use of graphical genotyping to visualize diversity at either the SNP or haplotype level has been widely used since it was first seen in the context of restriction fragment length polymorphism (RFLP) analysis (Young and Tanksley 1989). There are a number of tools that are utilised for the visualization of genotypic data. While within the last ten years data volumes have been low, in the low thousands of genetic markers, recent advances in molecular marker technologies have meant that volumes have been increasing at increasing rates both in terms of the number of physical genetic markers (millions of SNPs are not uncommon) being examined and the number of plant lines being genotyped. This trend is likely to continue and eventually be replaced by direct sequencing once costs are lowered and data handling improved.

The use of visualization or what is termed graphical genotyping has been used for a number of years. Tools such as GGT (Figure 3-9A) (Ralph van Berloo 2008) have allowed researchers to view graphical genotypes but are limited in the number and density of genetic markers they can visualise. Such tools have been important in the identification of haplotypes and features however they are often limited by the number of data points they can comfortably handle. Current generation genotyping platforms due to the volumes of data they can produce have limited the use of such tools to specific applications and small data volumes.

Recent graphical genotype visualization tools such as Flapjack (Milne et al. 2010) (Figure 3-9B) have addressed many problems associated with visualizing a large number of genotypes by employing back buffer techniques to paint the representations without a requirement for a large amount of memory (Demange et al. 2013), indeed such tools are currently capable of representing in excess of 250 million genotypes in real time with little if any performance lag.

Flapjack addresses some of the problems associated with visualizing large datasets and is optimized for efficient sorting and querying of genotypic and phenotypic data, but currently lack the ability to display data on a pedigree-based scaffold.

Flapjack offers high performance visual genotyping for up to 250 million genotypes in real time (memory limited). Genetic markers are represented by coloured blocks which can either be single DNA bases (adenosine A, cytosine C, guanine G or thymine T) or haplotypes (which are larger regions of commonality comprised of identifiable orders of SNP calls or any other genetic element) that form a matrix with plant lines on the y-axis and genetic markers on the x-axis (or vice versa).


Display Graphical Genotypes - C:\Users\Softpedia\Desktop\Softpedia Tested.ggt
File Edit Export Statistics Marker Selection Advanced Analyses Settings Help
1: "chrom1"
Show: 1 - 101
GGT Image
Data Sort, View, Edit
Marker Selection
Select Indiv. & Sel. Results
Associations
LD
Genetic Distances
Settings
Print Preview
cM
Group 1 ["chrom1"]
101 individuals/ 51 loci/ Pop:ri8
[9] LD 9
no selection
zoom:100%
A
HELIUM_TESTING_SET.flapjack - Flapjack - 0.10.02.11
File Edit Visualization Data Help
New Project
Open Project
Import Data
Data Sets
BOPA1_POPA_384SET 585x384
Trait Data
Default View
BOPA_POPA1536 585x1536
Trait Data
Default View
Chromosome: 1H
585 lines, 172 markers, length: 140.5
Overview
Line:
Marker:
Genotype:
Zoom:
Flapjack Tip: Search for all lines beginning with the letter 'A' by using the Find Dialog's regular expression: A.*
2C, 3T, 32.70MB

B

Figure 3-9 Graphical genotyping using GGT (A) and Flapjack (B). Both applications perform similar functionality in different ways and varying data scales. [117]

3.2.3 *Phenotypic visualization*

There are a number of systems that have been developed to try and address the problems associated with handling large volumes of phenotypic data. While in the plant breeding community commercial applications such as Agrobase ("Agronomix

Software - AGROBASE Generation II® Plant Breeding Software" 2014) are used to track both field layout and phenotypic data they are heavily focussed towards plant breeding. Other databases such as GrainGenes (Matthews et al. 2003; Carollo et al. 2005), Grameme (D. H. Ware et al. 2002; Ni et al. 2009), PhD (Li et al. 2005), PhenDisco (Doan et al. 2014) and resources the Scottish Barley Variety Database (SASA 2014) all hold various types of phenotypic data they offer limited functionality for basic phenotypic visualization tools in the form of data plots. They are also more suited to the research track and not for commercial breeding operations. There are also phenotype databases which aim to categorise data using controlled ontologies and visualization interfaces to browse the ontologies (by browsing tree representation of ontology) although there are a number of concerns within the community about this (Akiyama et al. 2014; Gkoutos et al. 2005).

Database systems such as Germinate (Lee et al. 2005) which were originally developed as small scale database to meet the demands of small lab based research projects have now been reengineered and implemented and has been used in a number of online data resources housing plant phenotypic data including the John Innes Centre *Psium* collection (Jing et al. 2010), The AGOUEB project ("AGOUEB -Association Genetics of UK Elite Barley" 2014), *Lolium* and *Festuca* Diversity Array Technology (DArT) markers (Kopecky et al. 2009; Jaccoud et al. 2001), DUS database encompassing phenotypic and genotypic data for UK Elite barley cultivars (Cockram et al. 2010; Wang et al. 2012) and flowering time data in barley (Comadran et al. 2012) have all been implemented in Germinate 3.

There is clearly a recognised disparity between phenotypic databases and integrated visualization tools with references to the issues and problems associated with this and a suggestion for more integrated resources in the future between databases and visualization and analysis tools being presented (Thorisson, Muilu, and Brookes 2009).

There a clear need is for is an integrated platform that allows the storage of phenotypic data (both quantitative and qualitative data) along with either integrated visualization tools or the ability to easily export data to alternate tools for subsequent analysis.

3.2.4 *Pedigree visualization*

Plant pedigrees can be complex. Efficient and intuitive tools are required to visualize (and interact with) complex pedigrees.

The development of pedigree visualization tools has primarily been carried out in humans and mammals. These include farmed (such as sheep (*Ovis aries*), cattle (*Bos taurus*), domesticated pig (*Sus scrofus*)), research based such mouse (*usually* but not exclusively *Mus musculus*) or domesticated such as cats (*Felis catus*), horses (*Equus ferus caballus*) and domesticated dogs (*Canis lupus familiaris*). There is often a large number of plant lines involved in any pedigree, many more so than in a traditional human pedigree. Much of the visual analysis used by breeders and plant scientists using plant pedigrees is still carried out using large print outs (Figure 3-10).

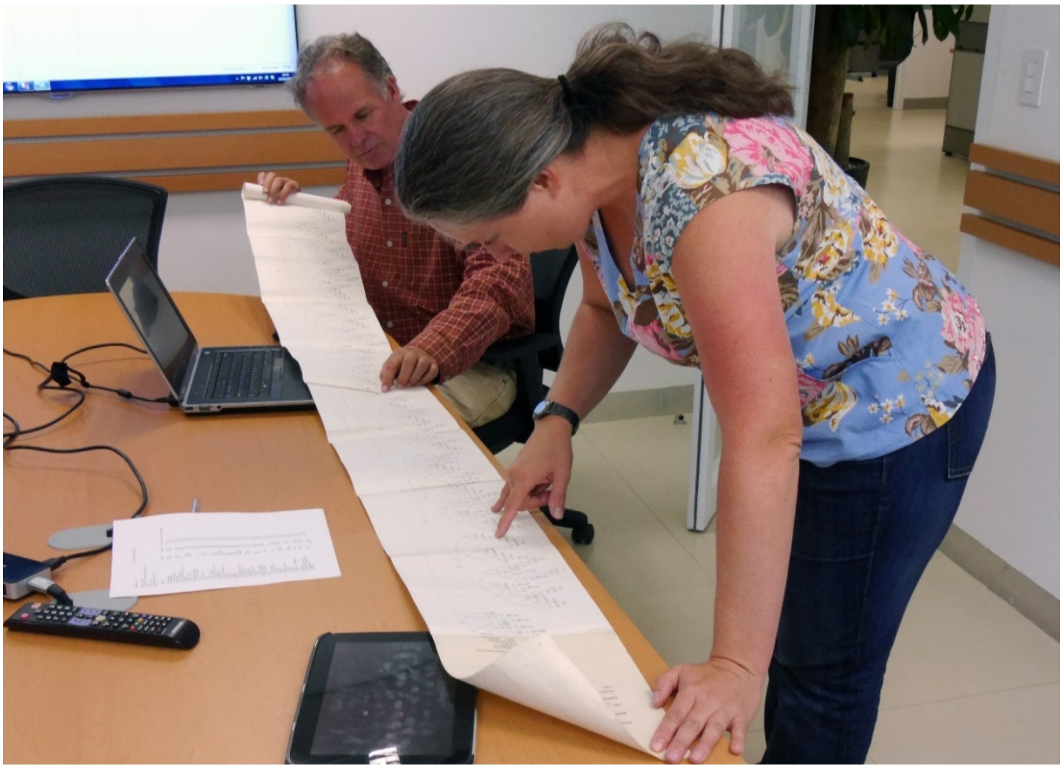


Figure 3-10 Examination of wheat pedigree records in wheat at CIMMYT (Centro Internacional de Mejoramiento de Maiz y Trigo) in Mexico

3.2.4.1 Previous work in pedigree visualization

Until now, pedigree visualization, with few exceptions (Voorrips, Bink, and van de Weg 2012; R van Berloo and Hutten 2005) has primarily been focussed on work carried out in the human genetics domain. Because plant breeding programmes involve phenomena not normally seen in human populations, there are additional visualization challenges that need to be overcome. There are often large numbers of plant lines involved in any pedigree, many more so than in an average human pedigree due to factors such as generation time/time to sexual maturity which is far lower in most plant species than that of their mammalian counterparts.

This section will look at the various visualization techniques used to represent pedigree based data and highlight the problems and strengths that these techniques exhibit.

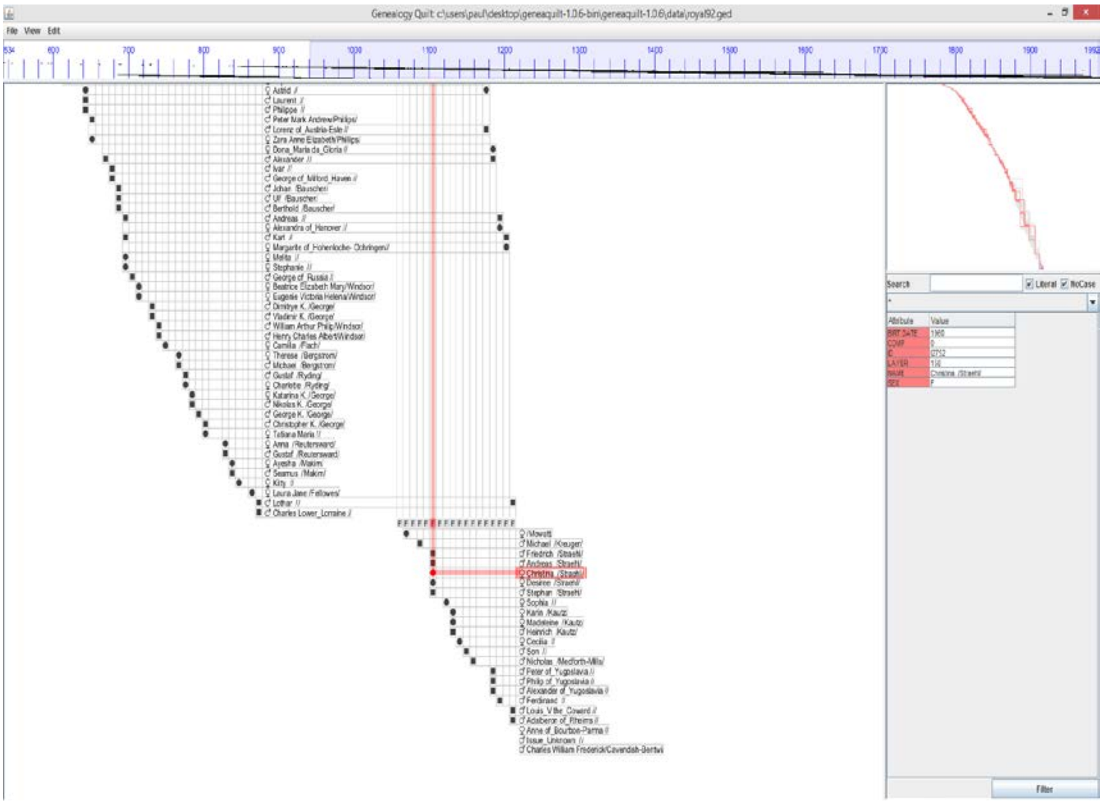
3.2.4.2 Table-based approaches

Tools such as PedStats (Wigginton 2005) offer statistical validation of users' pedigree data without visualization of the actual pedigree structure. It is difficult, if not impossible, to conceptualize pedigree structure for complex data sets without some visual representation.

Matrix-based visualizations to represent pedigrees use the intersection of the x and y edges to define relationships. Matrix-based visualizations have advantages over node-link or graph-centred layout approaches including the ability to create compact graph representations and the ability to remove edge overlapping. However, tests generating matrix visualizations using pedigree data as part of this work have shown that the data density is so low the resulting representations are not particularly insightful. The ability to quickly and easily track genetic flow and identify paths is also removed.

Tools such as GeneaQuilts (Bezerianos et al. 2010) offers a new visualization technique suitable for use with thousands of individuals (Figure 3-11) but offers limited scope for addition of complex genotypic and phenotypic data in its current form. In addition with large pedigrees it is difficult to view lineage without considerable panning across the screen. Discussions with users showed that they found it difficult to easily interpret such representations. The techniques described by Bezerianos have also been implemented in commercial software for drawing family trees and renamed Trellis Charts™ (“Trellis - The Chart With Everyone On It” 2014).

The physical layout of the Geneaquilts layout is a considerable paradigm shift from usual user expectation of a family tree with the top-to-bottom or left-to-right orientations normally seen. In addition it’s relatively difficult to quickly see complex relationships due to having to track horizontally and then vertically to find predecessors and successors. While highlighting as shown in Figure 3-11 uses coloured bars to show relationships goes some way to address this there is still a considerable amount of vertical scrolling required to view complex genealogies.

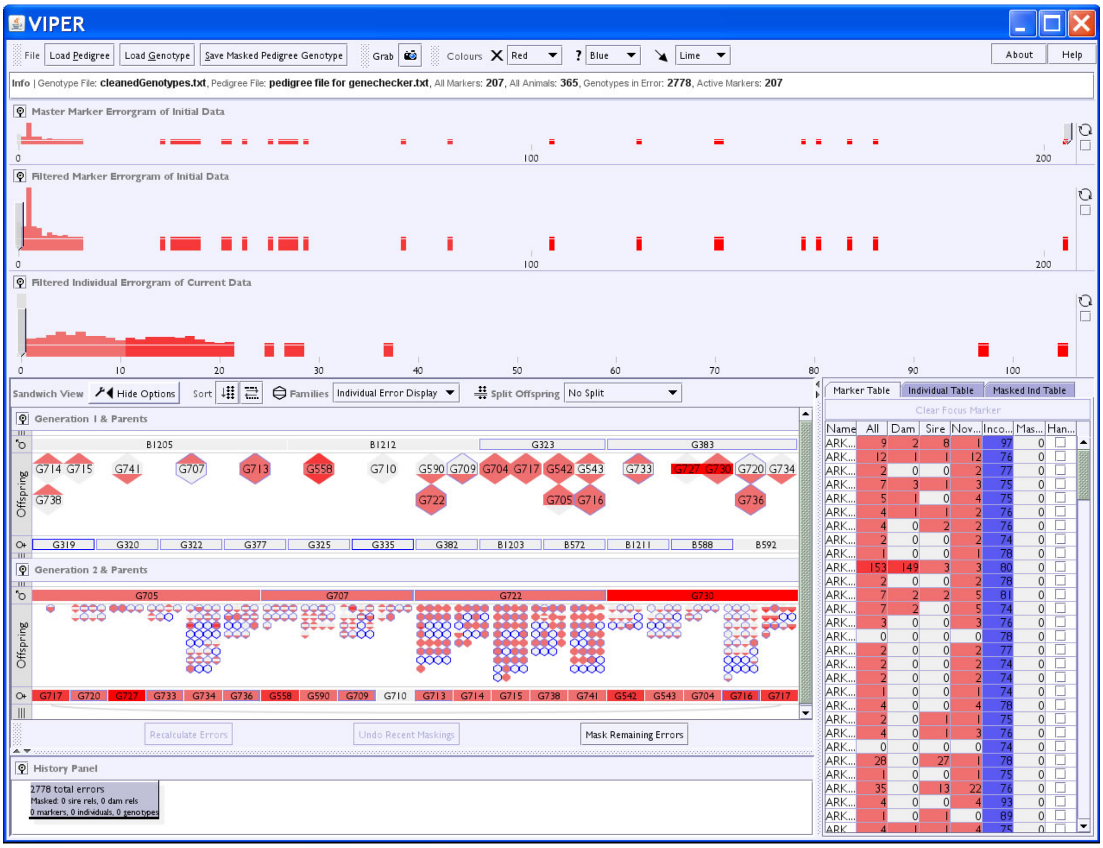

Genealogy Quilt: c:\users\paul\desktop\geneaquilt-1.0.6-bin\geneaquilt-1.0.6\data\royal92.ged

Search ☑ Literal ☑ NoCase

Attribute	Value
BIRT.DATE	1966
COMP	0
ID	2712
LAYER	150
NAME	Christina /Straehl/
SEX	F

Figure 3-11 Geneaquilts table style pedigree representation

Finally, tools such as VIPER (Paterson et al. 2012; Paterson et al. 2011), offer novel pedigree visualization and genotypic error checking capabilities but doesn’t allow for the inclusion of phenotypic information. VIPER is essentially a stack of nested table representations of generations where rows represent sires (male parents), dams (female parents) or children and columns represent individuals which can span multiple columns where they are parents. VIPER's primary use is in identification of genotyping problems in farmed animals and would be unsuitable for visualizing the complex crossing relationships that exist between crops where selfing is not uncommon. VIPER requires both separate male and female parents which is the norm in any applications handling animal or human data, but not always the case in plant breeding where the male and female parent can be the same individual. While VIPER does show structure in terms of male and female parents and their offspring what it doesn't highlight is the complexity of crosses, the cross type, nor the overall structure (Figure 3-12).



Info | Genotype File: **cleanedGenotypes.txt**, Pedigree File: **pedigree file for genechecker.txt**, All Markers: **207**, All Animals: **365**, Genotypes in Error: **2778**, Active Markers: **207**

Name	All	Dam	Sire	Nov...	Inco...	Mas...	Han...
ARK...	9	2	8	1	97	0	□
ARK...	12	1	1	12	76	0	□
ARK...	2	0	0	2	77	0	□
ARK...	7	3	1	3	75	0	□
ARK...	5	1	0	4	75	0	□
ARK...	4	1	1	2	76	0	□
ARK...	4	0	2	2	76	0	□
ARK...	2	0	0	2	74	0	□
ARK...	1	0	0	1	78	0	□
ARK...	153	149	3	3	80	0	□
ARK...	2	0	0	2	78	0	□
ARK...	7	2	2	5	81	0	□
ARK...	7	2	0	5	74	0	□
ARK...	3	0	0	3	76	0	□
ARK...	0	0	0	0	78	0	□
ARK...	2	0	0	2	77	0	□
ARK...	2	0	0	2	74	0	□
ARK...	2	0	0	2	74	0	□
ARK...	1	0	0	1	74	0	□
ARK...	4	0	0	4	78	0	□
ARK...	2	0	1	1	75	0	□
ARK...	4	0	1	3	76	0	□
ARK...	0	0	0	0	74	0	□
ARK...	28	0	27	1	78	0	□
ARK...	1	0	0	1	75	0	□
ARK...	35	0	13	22	76	0	□
ARK...	4	0	0	4	93	0	□
ARK...	1	0	1	0	89	0	□
ARK	4	1	1	4	75	0	□

2778 total errors
Masked: 0 sire rels, 0 dam rels
0 markers, 0 individuals, 0 genotypes

Figure 3-12 VIPER interface for genotype checking in farmed animals

3.2.4.3 Tree and Graph-based layout approaches

Visualization techniques such as sunbursts (Stasko et al. 2000) which are space filling versions of a node-link diagram have the advantage that a node's position in a hierarchy is maintained. Additionally, Fan Charts (Draper 2008) and H-trees (Claurissa Tuttle, Nonato, and Silva 2010) have also been described as a means for recounting human genealogy; these techniques however assume no inbreeding (they are trees and not graphs) and thus rule themselves out for use with plant pedigrees. They are also not suited when dealing with large numbers of individuals such as in the experimental datasets used in this work which have over 500 'nodes' and 1000 edges. With these volumes of data the visualizations become complex and cluttered. Additionally tools such as PedVis (C Tuttle, Nonato, and Silva 2010) offer alternate space filling layouts to techniques such as fan charts. Figure 3-13 shows the difference between a traditional tree-type left to right hierarchical (Sugiyama layered) layout (Figure 3-13A), a fan chart (Figure 3-13B) and finally the space filling PedVis layout (Figure 3-13C). Each of the images show the same data so a comparison can be made between them. While there is clear structure in the traditional model and fan chart the PedVis layout is somewhat more difficult to conceptualise initially and thus requires familiarity to start to identify features that are clearly seen in traditional layouts. Baring this in mind Tuttle states that most users (12 users) preferred the H-tree layout compared to a traditional (4 users) and fan chart layout (3 users).

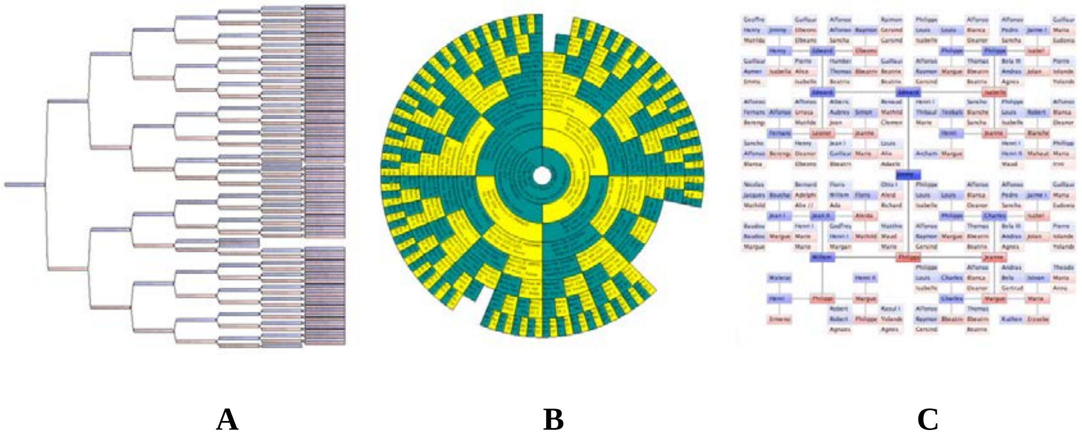


Figure 3-13 Traditional layout (A), fan chart (B) and PedVis space filling layouts (C) (C Tuttle, Nonato, and Silva 2010)

Another problem with these layouts is that they also require the duplication of nodes in order to maintain pedigree structure. The duplication of nodes while it can be argued increases the readability and simplifies layout; simplifies the complexity that is represented in a complex pedigree which may not be desirable and conceptually confusing in that you would not expect to see multiple entries for the same plant line on the same chart.

While the main problems with these additional techniques are that they are not appropriate for observing a pedigree in its entirety (indeed the complexity of the data rules many of them out), they may however be useful when trying to visualize a sub-section of data such as a sub-pedigree for specific plant lines where the pedigree complexity is reduced, and extraneous plant lines removed.

The problem of very large pedigrees in humans has been identified and solutions proposed in tools such as PViN (Wernert and Lakshmipathy 2005) which looks at windows/viewports on large datasets but only offers pedigree drawing with no scope for addition of other information onto the visualization. PViN allows a large and complex pedigree to be more easily viewed by showing an expanded view for a specific section of the complete pedigree when selected by the user while maintaining the full pedigree and showing, by means of focus selection the area being observed. Figure 3-14 shows PViN running on a large display screen. The screen in this case is split in to two distinct regions. The uppermost showing the entire pedigree and the lower portion showing a selected region from the upper display. This allows the entire pedigree to be visualized while showing detail. While tools such as this allows large pedigrees to be displayed they are reliant on expensive visual display equipment which is not available to the average scientist or plant breeder.

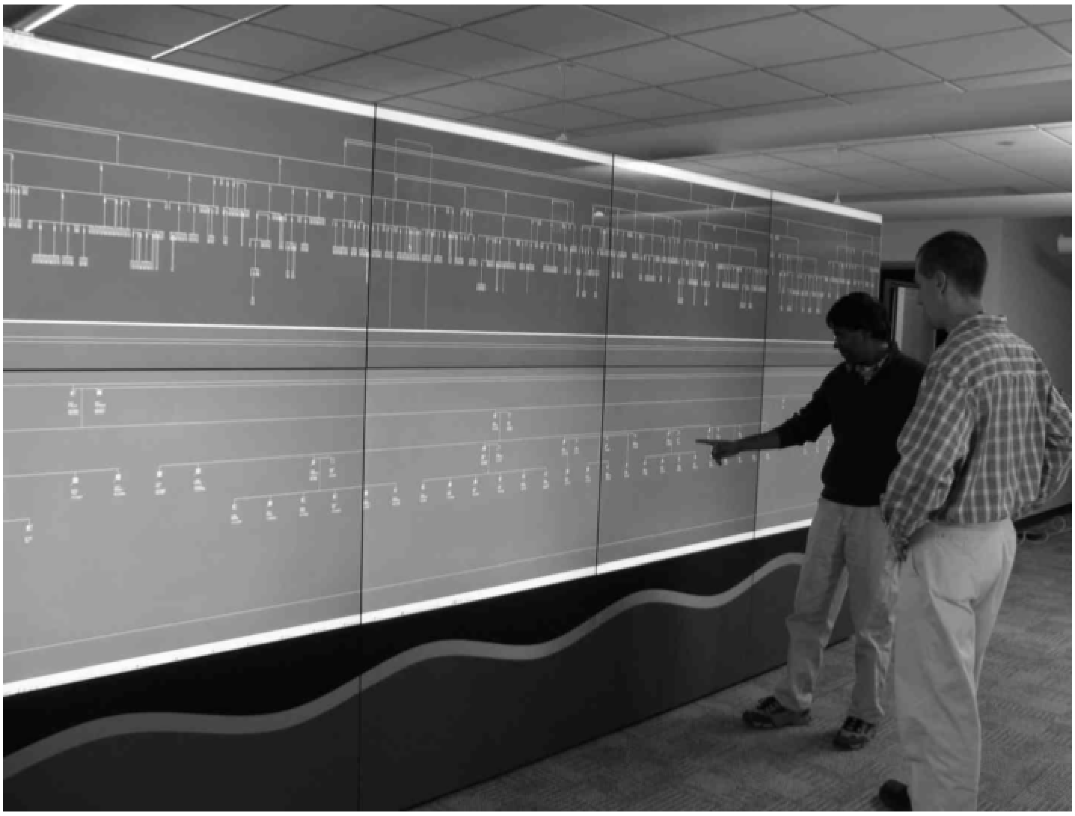
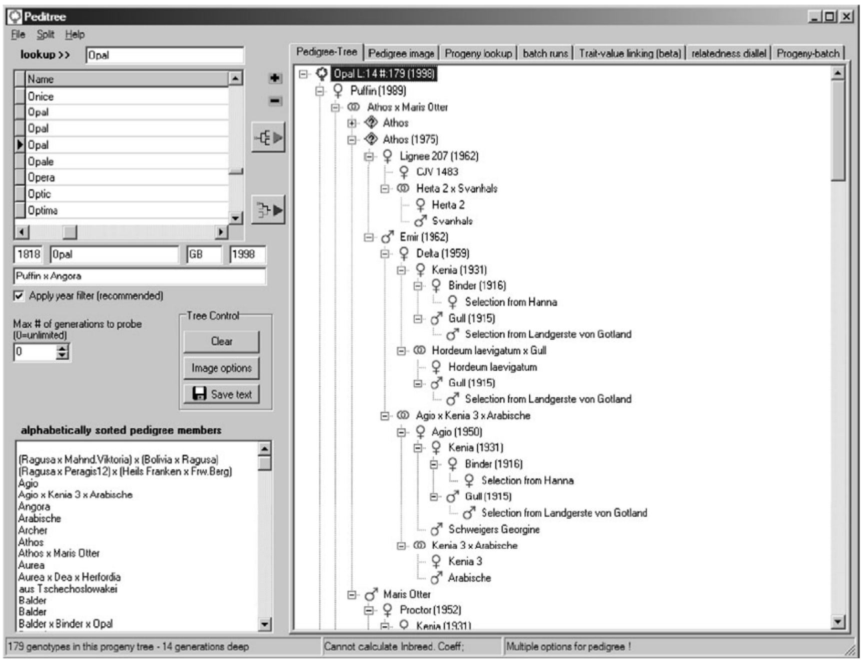


Figure 3-14 PViN (Wernert and Lakshmipathy 2005)

Another example of tree-based pedigree visualization software is Peditree. Peditree exhibits the problem that it only offers a tree-based view of data in a pedigree but this is not necessarily suited, as previously discussed, to plant pedigrees due to inbreeding and the use of older plant lines in more modern crosses which prevents us from treating them as such.

Unlike trees, graphs allow for the precise modelling of the complexity of a plant breeding programme. *Pedigrees are not trees*, although they are often presented as such (R van Berloo and Hutten 2005) which is misleading in the context of this work. While van Berloo's work with Peditree (Figure 3-15) represents pedigrees as a hierarchical tree structure and therefore the visualization presents potential comprehension problems.

Techniques such as node link diagrams have long been used as a way of representing graph-based data and recent work has examined how effective the node-link model performs representing graph data when compared to matrix-based visualizations (Ghoniem, Fekete, and Castagliola 2004). Work carried out by Purchase (H C Purchase, Cohen, and James 1995; H.C. Purchase 2000; H. Purchase, Carrington, and Allder 2002) and Bennett (C. Bennett et al. 2007) also indicate that while graph layout played an important part in a user's understanding, it was not the major focus; this focus perhaps being the use of other aesthetics relating to node colour and shape.



lookup >> Opal

Name
Onice
Opal
Opal
Opal
Opale
Opera
Optic
Optima

1818 | Opal | GB | 1998

Puffin x Angora

☑ Apply year filter (recommended)

Max # of generations to probe (0=unlimited) 0

alphabetically sorted pedigree members

```
(Ragusa x Mahnd.Viktoria) x (Bolivia x Ragusa)
(Ragusa x Peragis12) x (Heils Franken x Frw.Berg)
Agio
Agio x Kenia 3 x Arabische
Angora
Arabische
Archer
Athos
Athos x Maris Otter
Aurea
Aurea x Dea x Herfordia
aus Tschechoslowakei
Balder
Balder
Balder x Binder x Opal
```

```
Opal L:14 #:179 (1998)
  ♀ Puffin (1989)
    Athos x Maris Otter
      Athos
      Athos (1975)
        ♀ Lignee 207 (1962)
          ♀ CJV 1483
          Herta 2 x Svanhals
            ♀ Herta 2
            ♂ Svanhals
        ♂ Emir (1962)
          ♀ Delta (1959)
            ♀ Kenia (1931)
              ♀ Binder (1916)
                ♀ Selection from Hanna
              ♂ Gull (1915)
                ♂ Selection from Landgerste von Gotland
            Hordeum laevigatum x Gull
              ♀ Hordeum laevigatum
              ♂ Gull (1915)
                ♂ Selection from Landgerste von Gotland
          Agio x Kenia 3 x Arabische
            ♀ Agio (1950)
              ♀ Kenia (1931)
                ♀ Binder (1916)
                  ♀ Selection from Hanna
                ♂ Gull (1915)
                  ♂ Selection from Landgerste von Gotland
              ♂ Schweigers Georgine
            Kenia 3 x Arabische
              ♀ Kenia 3
              ♂ Arabische
      ♂ Maris Otter
        ♀ Proctor (1952)
          ♀ Kenia (1931)
```

179 genotypes in this progeny tree - 14 generations deep | Cannot calculate Inbreed. Coeff; | Multiple options for pedigree !

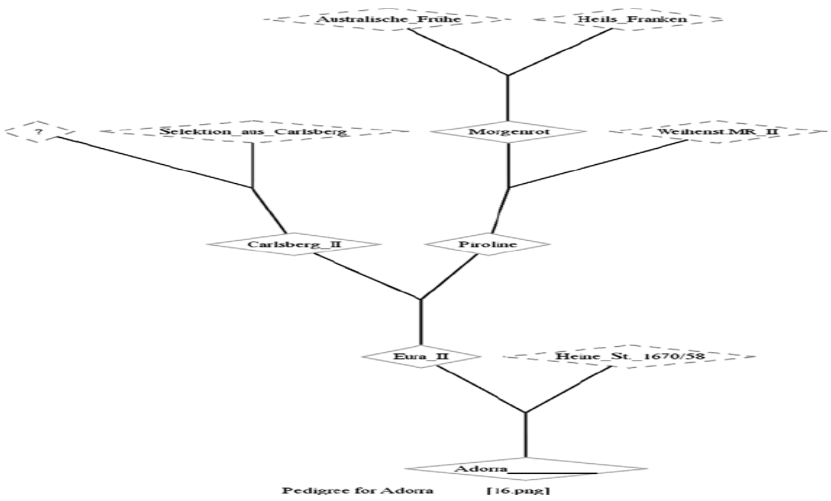


Figure 3-15 Peditree user interface including simple tree representation of pedigree

The assumption that pedigrees take the form of trees rules Peditree out for our test datasets. There are however other pedigree visualization tools which facilitate incestuous relationships and consanguinity such as Madeline (Trager et al. 2007). Madeline allows cyclic graphs in its pedigree layouts to allow for consanguineous mating Figure 3-16. Madeline allows the inclusion of categorical information by encoding this in what it terms a quadrant. Each quadrant represents an individual in the pedigree and is limited to 8 states per sex (circular and square quadrants). Madeline uses grayscale or bichromatic colour palettes to colour individuals.

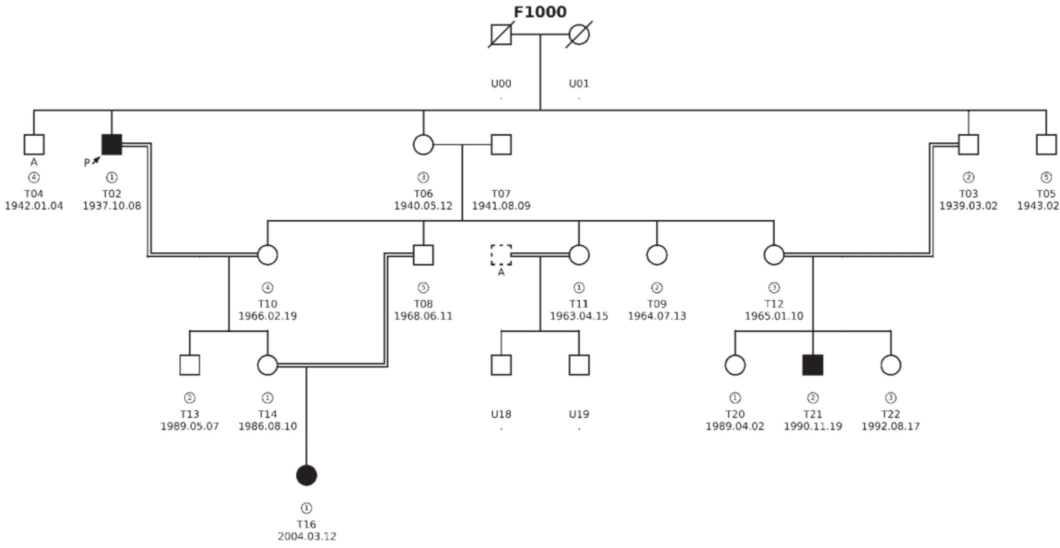


Figure 3-16 Madeline showing consanguineous mating events and duplicate entries.

Madeline does however have problems with the layout it uses in terms of the number of individuals that can be presented on the screen at any time but this is a consequence of its target domain of human genetics where population numbers are frequently much less than in crop plants. While the images it produces are very clear more efficient use of space would increase node density and the use of duplicates to reduce edge overlap means that that while the images it produces have a high degree of clarity the overall picture of how individuals are related is simplified and therefore detracts from the actual pedigree complexity. These problems mean that Madeline would be unsuitable for the density of individuals that large plant pedigrees contain.

Other tools such as Cranefoot (Mäkinen et al. 2005) report the use of mathematical graph structures to deal with between-relative mating but the approach is limited in its current form in the amount of information that can be attached to a node (Figure 3-17). Cranefoot uses the same standard human nomenclature for males and females (square and circular nodes) and like Madeline only offers the ability to include categorical data within a quadrant. There are no interactive features making exploration of large pedigrees difficult with such tools.

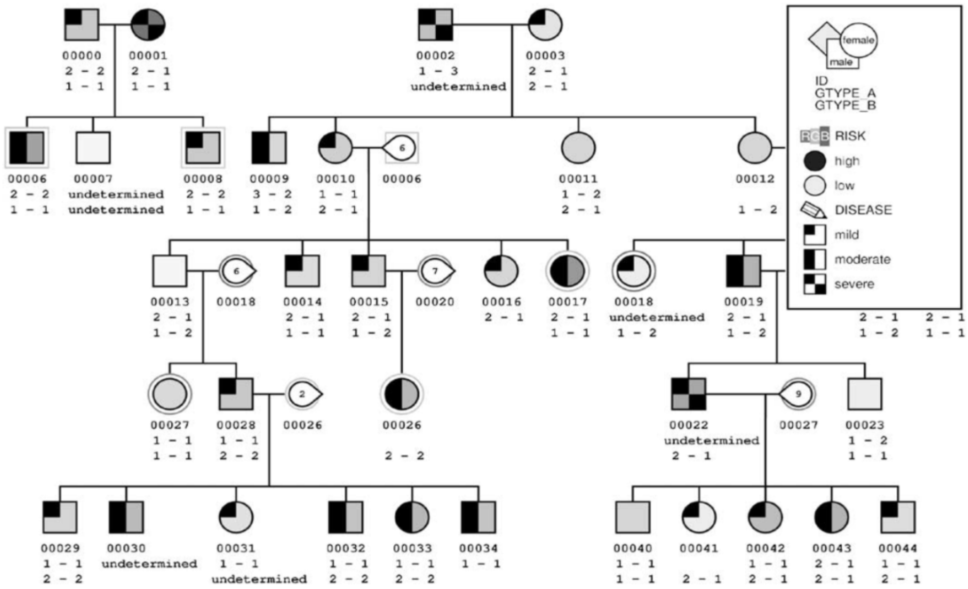


Figure 3-17 Cranefoot

Finally, HaploPainter (Thiele and Nürnberg 2005) allows the drawing of genetic haplotypes, but suffers from being restricted in the number of individuals it is able to display (Figure 3-18).

HaploPainter also allows the traditional representation of pedigrees (Figure 3-19) and allows users' to change node colour based on a defined character but again like the other node-link style applications no interactive features for visual exploration of data.

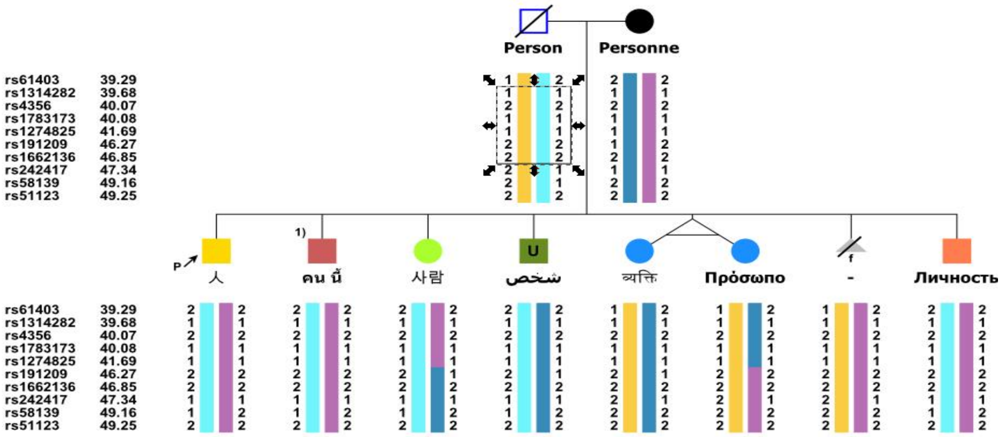


Figure 3-18 HaploPainter haplotype visualization

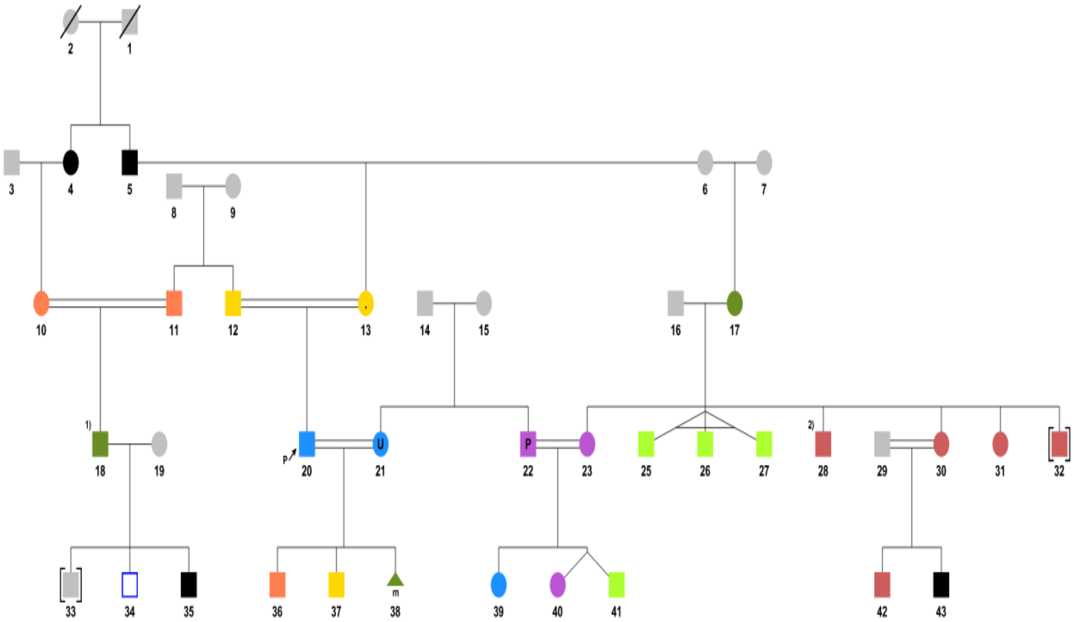
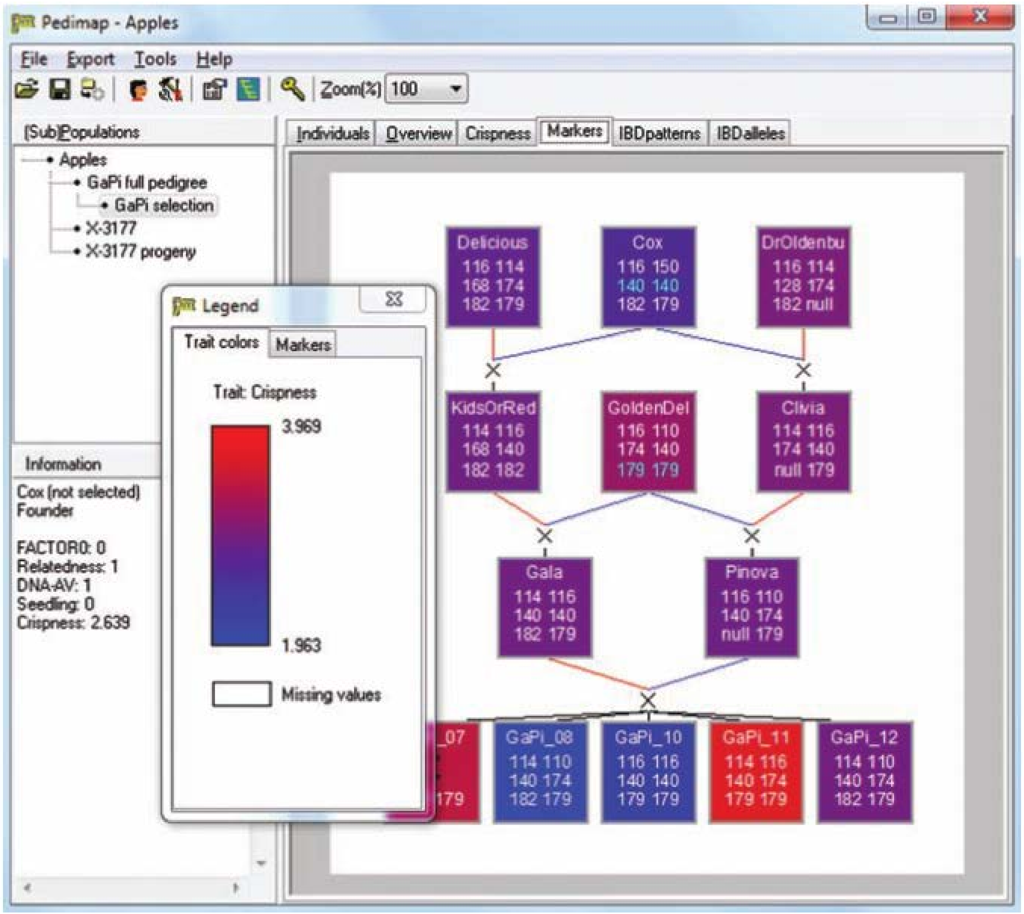


Figure 3-19 Haplopainter pedigree drawing application.

Newer applications such as Pedimap (Voorrips, Bink, and van de Weg 2012) offer the ability to colour nodes based on phenotype (Figure 3-20) and also allow the overlaying of genetic information but performs no calculations, these must all be handled by external applications then imported as text files into Pedimap. The density of plant lines that this tool can handle is also relatively small and more suited to smaller breeding programmes. The representation of the pedigree as a collapsible tree means it has the same problem as tools such as Peditree which don't truly reflect a pedigree structure. Pedimap also allows the overlaying of genetic data, in the case of Figure 3-20 SSR (Simple Sequence Repeat) loci along with the phenotype 'crispness'.

Other tools such as the web-based Pedigree Visualizer (Figure 3-21) by Wong (L. Wong 2000) offer alternative layout algorithms. Wong suggests introducing duplicate 'alias' entries in representations with multiple matings from the same individuals, phenomena that are commonplace in plant data. This however has a major drawback in that the complexity of the pedigree cannot be accurately visualized. Other tools such as PyPedal (Cole 2007), which is a Python module, not only offers rudimentary graph drawing tools, restricted to changing node shape to represent male and females (Figure 3-22), but also error checking algorithms to try and identify potential pedigree errors where appropriate genotypic data exists. It can also be used to calculate statistics such as coefficient of inbreeding which may be useful in animal contexts.

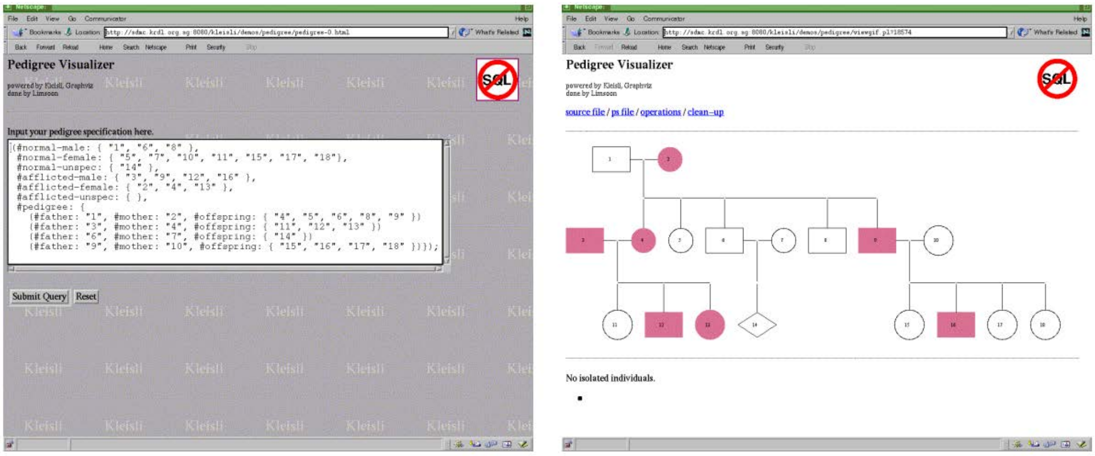


Information

Cox (not selected)
Founder

FACTOR0: 0
Relatedness: 1
DNA-AV: 1
Seedling: 0
Crispness: 2.639

Figure 3-20 Pedimap phenotype colouring


Netscape:
File Edit View Go Communicator Help
Bookmarks Location:
What's Related
Back Forward Reload Home Search Netscape Print Security Stop
Kleisli
SQL
Submit Query
Reset
source file / ps file / operations / clean-up
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

http://sdmc.krdl.org.sg:8080/kleisli/demos/pedigree/pedigree-0.html

Pedigree Visualizer

powered by Kleisli, Graphviz
done by Limsoon

Input your pedigree specification here.

```
(#normal-male: { "1", "6", "8" },
 #normal-female: { "5", "7", "10", "11", "15", "17", "18"},
 #normal-unspec: { "14" },
 #afflicted-male: { "3", "9", "12", "16" },
 #afflicted-female: { "2", "4", "13" },
 #afflicted-unspec: { },
 #pedigree: {
   (#father: "1", #mother: "2", #offspring: { "4", "5", "6", "8", "9" })
   (#father: "3", #mother: "4", #offspring: { "11", "12", "13" })
   (#father: "6", #mother: "7", #offspring: { "14" })
   (#father: "9", #mother: "10", #offspring: { "15", "16", "17", "18" })});
```

http://sdmc.krdl.org.sg:8080/kleisli/demos/pedigree/viewgif.pl?18574

Pedigree Visualizer

powered by Kleisli, Graphviz
done by Limsoon

No isolated individuals.

Figure 3-21 Pedigree Visualizer

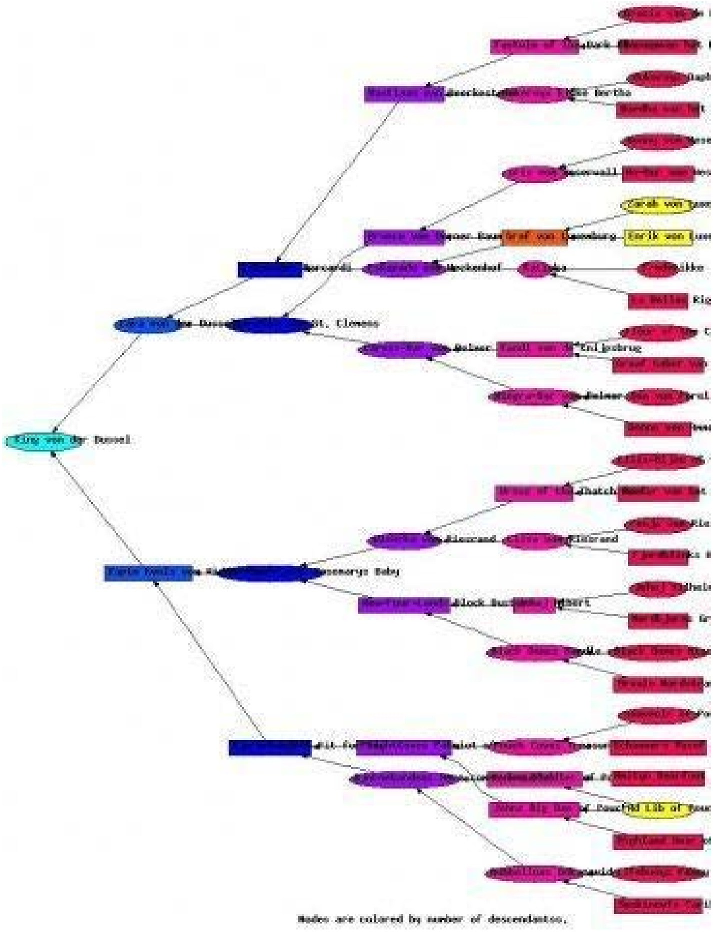


Figure 3-22 PyPedal pedigree visualization

Although there are problems associated with 2D node-link layouts such as a lack of horizontal space and problems with crossing of edges (Loh et al. 2008) they are still well suited to displaying pedigree data. While 3D based tools do exist, they display problems including visual occlusion and that they tend to visualise high-level features and not specifics, so while some trends are easy to spot, the actual detail is hidden from the user. From this point of view they are limited in use and offer no advantages over their 2D counterparts. Notable examples of such tools are Walrus (“Walrus - Graph Visualization Tool” 2011) and Celestial3D (Loh et al. 2008) (Figure 3-23) who identify the problem with large animal pedigrees where numbers increase generation on generation. There are however well documented issues with using 3D representations of what in reality is 2D data with regards to data occlusion, something which may have major negative impacts on visual analytics functionality of these tools.

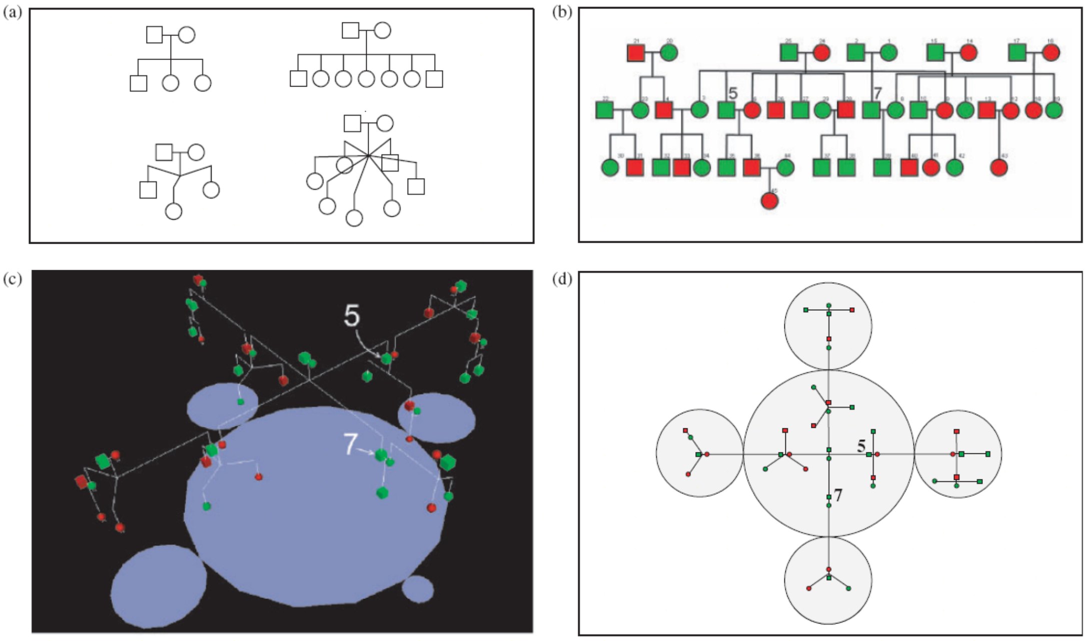


Figure 3-23 Pedigree representation using Celestial

3.3 Screen availability, distorted and non-distorted views

The amount of data routinely available very often is far greater than the amount of space available on screen for visualization (Shneiderman 2008). This is a problem for a number of reasons. Firstly, while more abstract views can be constructed on the data in order to fit more information on screen at the same time, this leads to the loss of factual data which may be required by any user of such a system in order to make informed decisions.

While screen resolutions are increasing there is still a limitation to how much information can be displayed at any point of time and be comprehended and digested by any user.

Leung's (Leung and Apperley 1994) taxonomy of presentation techniques for large graphical data spaces shows that it is possible to categorize techniques in two different classes: non-distorted and distorted views. Examples of non-distorted approaches include scrolling and zooming, and hierarchical views, where part of the information is hidden. While this is ok for small datasets the use of such techniques in larger datasets can lead to problems of losing position within the visualization. Taking this into consideration and the fact that related data may not sit close to each other the use of distorted methods such as extreme fish eye effects are not suited to this type of information visualization (Bartram et al. 1995). It is therefore important that the data overview window is maintained and only change the position and layout of nodes within the detail window, if at all.

3.4 Discussion

There are two main classes of pedigree visualization tools; table based and node-link/graph based. Of these the table based solutions can be discounted due to the low density of individuals they can effectively show and the problems they have showing complex structure involving selfing and inbreeding; they do not effectively represent hierarchy that is seen in pedigrees. An overview of the different visualization techniques, features and their presence in the visualization tools is shown below. The last column identifies techniques and features required by this work.

Visualization Feature	PedStats	PedVis	GeneaQuilts	Viper	PViN	Peditree	Madeline	Cranefoot	Haplopainter	Pedimap	Pedigree Visualizer	PyPedal	Walrus	Celestial	Required for This Work
1. Detail and Overview			Y	Y	Y				Y						Y
2. Zooming, Scrolling and Panning			Y	Y	Y	Y			Y	Y			Y		Y
3. Filtering	Y		Y	Y		Y				Y					Y
4. Focus and Context															
5. Graph Based							Y	Y	Y				Y		Y
6. Tree Based		Y			Y	Y				Y	Y	Y			
7. Table Based / Matrix	Y		Y	Y											
8. Selection mechanisms		Y	Y												Y
9. Data Density	L	L	M	M	L	L	L	L	L	L	L	L	H	L	H
10. Phenotypic Data						Y	Y	Y		Y	Y	Y		Y	Y
11. Genotypic Data				Y					Y	Y					Y

Table 3-1 Comparison of pedigree visualization tools and their features. Y indicates presence of feature. In “10. Data Density” the scale of L – up to 100 plant lines, M – up to 1000 plant lines and H – over 1000 plant lines is used.

Of the tree and graph based solutions only Madeline, Cranefoot, Walrus and HaploPainter offer the features that are required to accurately model plant pedigrees, namely the ability to account for inbreeding. Their use of graphs instead of tree structures facilitates this. These tools however have a number of problems which make them unsuitable for use with the plant pedigree datasets used in this work. Firstly they are limited on the number of nodes that can be displayed (with the exception of Walrus) and while clear, do not make good use of available screen space for larger pedigrees. Secondly, there are limitations with the types of phenotypic data that can be assigned to each node, mainly limited to categorical/qualitative data. Lastly these tools offer limited, if any, interactive features which makes the visual exploration and interaction with large pedigrees difficult.

All of the graph based solutions use Sugiyama-style (Sugiyama, Tagawa, and Toda 1981) top to bottom (or left to right) layout algorithms to maintain the concept of generations. These layered layouts preserve the topological structure that is typically seen in pedigree representations.

There is no compelling reason to use 3D layouts for this work. The problems with occlusion of data and the data not having a 3rd dimension that would be useful precludes 3D techniques from this work. While there is a time component to many pedigrees in terms of the year of release of particular cultivars the vertical stacking of nodes in the traditional layouts used for pedigrees is both well tried and tested and more suitable for the time component of datasets.

Finally, there are issues surrounding the inclusion of other data types within the visualization. While tools such as Geneaquilts, Haplopainter and Pedimap allow the inclusion of additional information in the form of basic phenotypic (mainly categorical data) the other tree and graph based approaches do not offer this functionality. It must be noted that none of these tools offer the ability to overlay anything but the simplest descriptive phenotype data, something which would not meet the needs of this work. Data can be overlaid on to the pedigree representation but there are no features to allow data to be retrieved from databases and included into the visualization within detail panels, nor exploration of this information.

The genotypic data capabilities are also very rudimentary in most of the tools which support it. While applications like Viper allow for sophisticated error checking based on pedigree it lacks features to display the complexity associated with complex plant pedigrees without introducing duplicate plant lines. While other tools such as Haplopainter and Pedimap allow for the inclusion of a small number of SNP's or the display of haplotype information they lack the ability to overlay complex marker based data from large datasets and features such as using this data to display genetic similarity, again a feature that was requested during user interviews.

The number of plant lines that was identified in user interviews (Section 2.7.1) mean that any tool that is used or developed needs to be able to handle upwards of a thousand plant lines. To this end all but Geneaquilts, Walrus and Viper (which have all previously been discounted due to the inability to incorporate specific additional data types) are suitable for anything approaching this volume of plant lines. Therefore there are no currently available tools that would handle the volumes of data that form the test datasets used in this work unless compromise was met with regards to displaying only sections of a pedigree, something which is undesirable.

It is clear that these techniques and tools contain many features that are useful, but none meet the *exact* requirements (including data abstraction) of the defined problem of being able to overlay genotypic and phenotypic data onto a large and complex plant pedigree structure. These pedigree visualization tools detailed here are more generally suited to human pedigrees where the number of individuals is comparatively low when compared to even a relatively small scale plant or animal breeding experiment with much more complex relationships between individuals. Because of this the classical pedigree chart is unsuited for use in plant research in its current form.

There is therefore a need for the development of tools that are tailored for the unique needs of plant breeding with the ability to explore pedigree structure, and paint additional genotypic and phenotypic data on top. This will allow breeders to make informed decisions and visualize the way in which alleles for agriculturally important traits are transmitted through previous and subsequent generations. Such tools do not currently exist.

Through the examination of previous visualizations to display pedigree data it was hypothesised that the best method to visualize plant pedigree data was a node-link based approach. Not only does this allow the accurate mapping of the exact specifics of how breeding programmes run (including inbreeding) but also provides a well-established visual metaphor onto which a visualization tool could be built. The layered layout representation also brings a coherent structure to sparse relationships and generations and topological layout are clearer when compared to matrix style layouts.

This is not the case with animal and human pedigrees whose top-down fan type shape is not well suited to a layered layout as they quickly become very large, consuming large volumes of horizontal space (Paterson et al. 2012).

Tools that allow exploration of data to try and bring a greater understanding of complex relationships between individuals should bring greater insight into how plant breeding programmes operate at the genetic level and how to bring maximum potential benefit from them. The ability to detect patterns and associations (or even anomalies) within these datasets such as; the identification of problems with inheritance of alleles, the identification of plant lines from which additional information would allow inference of data on large parts of the pedigree, simple typos and errors, or looking for plant lines which are similar to unknown plant lines.

It is clear that such a solution will need to draw on visualization techniques such as detail and overview/zooming, filter, focus and context techniques to try and pull together data into a pedigree structure without overwhelming any visualization with screen clutter.

The volumes of data that are now used in many visualization systems mean that not only are single visualizations sometimes inappropriate, but the volumes and complexity of data means that a simple visualization is not sufficient to convey all information that is required to make decisions. Because of this many tools use what has been coined ‘Detail and overview systems’ (Section 3.1.1) and allow users to switch between linked, contextual views of data allowing a much greater density of data to be viewed without having busy visualizations that are too information rich to be of use to domain experts.

While other techniques such as zooming, panning (Section 3.1.1) and paging allow large volumes of data to be moved into conceptual focus (Section 3.1.4) and techniques such as introducing new overlapping windows into visualization displays allows new displays with additional information the ability to have linked windows side-by-side removes some of the discontinuity of jumping between windows and occlusion of other visualization windows (Andy Cockburn, Karlson, and Bederson 2008).

With the development of the Germinate data warehouse the data that is required for the visualization of pedigrees and associated data is held in a standard single location in a format suitable for pedigree analysis.

Visualization tools which can use this data to present users with useful views on their data were then required.

Chapter 4 Pedigree visualization prototype

4.1 Overview

This chapter discusses the development of an initial paper-based pedigree visualization prototype. It starts by detailing the process behind the decision to develop a desktop application instead of alternative technologies then moves on to detailing the iterative development process behind the creation of visual metaphors to represent plant lines and relationships between plant lines and thus the representation of pedigree structure. The chapter discusses the visual variables used to encode information using these visual metaphors, layout considerations and concludes by discussing the feedback that was gained when talking to domain experts while they interacted with the paper-based prototype.

4.2 Developing a prototype system

Taking the results obtained from the initial requirements gathering (2.7 Initial requirements gathering) the foundations that would form a prototype pedigree visualization system were put in to place. In order to model the abstract representation of a plant pedigree visual metaphors to represent the main concepts that were trying to be conveyed were developed.

Software tools for the visualization of data need to be able not only to handle current data requirements but also those which will arise in the future.

An iterative design pattern was used in this work. Feedback would be obtained from users which would then be used to drive the continued development of the visualization tool.

4.2.1 *Data pre-processing*

A simple Java command line application was written which takes data in Lamacraft notation and parses this into the defined pedigree notation. Any pedigrees that do not match this format are highlighted and can be dealt with manually. The program automatically creates intermediate crosses in a format compatible with Germinate.

4.3 **Delivery of visualizations**

There are two main formats used for data visualization that were considered. Online or web-based visualization using the web browser as the delivery medium or application based utilizing the features offered by desktop programming languages and toolkits. Each approach has advantages and disadvantages.

4.3.1 *Web based delivery*

4.3.1.1 Advantages of web based visualization approaches

The main advantage to a web-based approach is in its cross platform compatibility. In addition additional software usually will not need to be installed prior to use if the application sticks to basic core web-standard technologies such as Cascading Style Sheets (CSS) and JavaScript with pre-processing carried out on a central server whose makeup can be as simple or complicated as the task requires. Using these web standards ensures, so some extent, cross platform compatibility. Web based approaches also mean that the specifications of the users client machine are not as important as most data processing will be done at the server end.

Recent years have also seen the increase in the use of web-based applications and the use of computing clouds to perform large data analysis and manipulation tasks such as Amazons EC2 service.

Web based approaches also mean less complicated configuration for client users to use your applications, while network traffic may be an issue the increased performance of networks and internet bandwidth to the average user continues to increase.

4.3.1.2 Disadvantages of Web Based Visualization Approaches

The main disadvantage of data visualization via a web based medium is ultimately performance. While web standard technologies such as the HTML5 specification may offer visualization capabilities for small scale datasets they tend to scale badly when dealing with larger volumes of data such as those generated through current high throughput technologies. While many of these issues can be addressed by pre-processing of data they still impose limitations that are difficult to deal with.

4.3.2 *Standalone application based delivery*

4.3.2.1 Advantages of standalone application based approaches

The use of high performance programming languages such as Java, C# or C++ allows the development of tools for the visualization of large data volumes. These tools can make use of local hardware resources such as graphical processing units (GPU) where available to increase visualization performance. This ability to offload data processing to the user's machine instead of the server as would be the case in many web-applications is an important benefit of this type of approach, especially with the volumes of data being used here.

4.3.2.2 Disadvantages of standalone application based approaches

As an application is run on the user's computer and not on a server this brings the overhead of having to deal with an infinite number of both computer hardware configurations and host OS (Operating System). While programming languages such as Java try to overcome these issues by using a common runtime environment there are nevertheless still instances where differences exist between machines.

In addition, where large volumes of data are available in for example a database it is not always practical, or indeed beneficial to store this on a user's machine in its entirety. This means that there are situations where the use of a combined model is the most appropriate choice for such applications, this is described below (Table 4-1).

	Client Based	**Web Based**	**Hybrid Approach**
Processer Availability	User control over resources	Determined by server load	Determined by both server and users processing power
Speed	Determined by users machine	Determined by users machine and server	Determined by users machine and server
Data	Local data repository required	Data held on server	Data held on server
Interfaces			
Languages	Java, Air, C++, etc.	HTML, JavaScript, Flash	Hybrid of client and web based technologies

Platform Independency	May need to be compiled for operating system or reliance on Java Virtual Machine cross platform compatibility	Reliance on browsers standards compliance not operating systems	May need to be compiled for operating system or reliance on Java Virtual Machine cross platform compatibility
Software Updates	Cannot guarantee all users are using the most up to date software version even with automatic updates	Updates to server only ensures users are all using current version	Cannot guarantee all users are using the most up to date software version even with automatic updates
Portability	Software needs to be installed on all computers it is used on	Only a web browser is required with no additional software requirements*	Software needs to be installed on all computers it is used on
Availability	No restrictions on availability	Available only when user has internet connection	Available only when user has internet connection
Security	As secure as users machine	Transmission of data over internet may have security implications	Transmission of data over internet may have security implications
Deployment	Deployed as installer for users operating system	Deployed to controlled server	Hybrid of client and web based approaches

* Except where Flash or other browser add-ons are required.

Table 4-1 Comparison between Web and Desktop Application Based Approaches

4.3.3 *The combined model*

The combined model takes the benefits of a web-based approach in that data is held on a central server maintaining data security and access rights but utilises application based tools and utilities for viewing and visualization of complex data. The combined solution can also use the central server for generation of visualization components out with the client application if required.

This has the advantage of allowing interactive visualization within a standalone application with subsequent reduction on server load with the ability to restrict access to data using a web based technique. Where appropriate smaller scale data visualizations can also be carried out server-side directly into the web browser for users whose requirements may not require the standalone application.

Disadvantages do however exist in that a constant web connection is required to be maintained and one that is responsive enough to carry data to the client interface.

4.4 Design methodology

The software tool dot which is part of the Graphviz application suite was used in order to develop a simple prototype to examine how pedigrees would best be represented and to develop ideas on the best solution to meet the research questions detailed in Section 1.1. The following work in designing a paper based prototype was done in close collaboration with expert users adopting a flexible based approach which involved creating a prototype visualization then seeking input from the expert users which would in turn guide the next iteration of the prototype. This bottom up methodology meant that the prototype started as a simple representation and was tweaked based on the user feedback in order to move towards a solution to the problem of visualizing large plant pedigrees.

While the availability of expert users to interact with was one of the many positives to this work it was sometimes a challenge to identify what was indeed important to all users and what were requirements that would only benefit individuals or a small number of users. This was particularly evident with some users being more vocal and active than others leading to a tendency to prioritise, or pay higher regard to ideas and suggestions they offered.

Concentration on a static paper-based prototype at this stage would mean that concepts could be implemented and presented in order to both evaluate their acceptance and gain a better understanding of what end users would want from a more advanced prototype, and whether a more advanced solution was actually required for this work.

4.5 Design of initial low fidelity prototype visualization

Pedigrees are hierarchical in nature so it makes sense to select a layout algorithm for the graph data structure which maintains a sensible topology for the underlying data. There are two main graph layout types which could be suitable for this work, Sugiyama or layered layout (Figure 4-1A) and force directed (Figure 4-1B). However, force-directed layouts are unsuitable for use with pedigrees and are non-intuitive due to the lack of a temporal aspect to the layout, a layered layout is much more suited to pedigree visualization and so was used from this point forward in this work. The representation maintains arrow heads on edges which have been included for clarity (H.C. Purchase 2002).

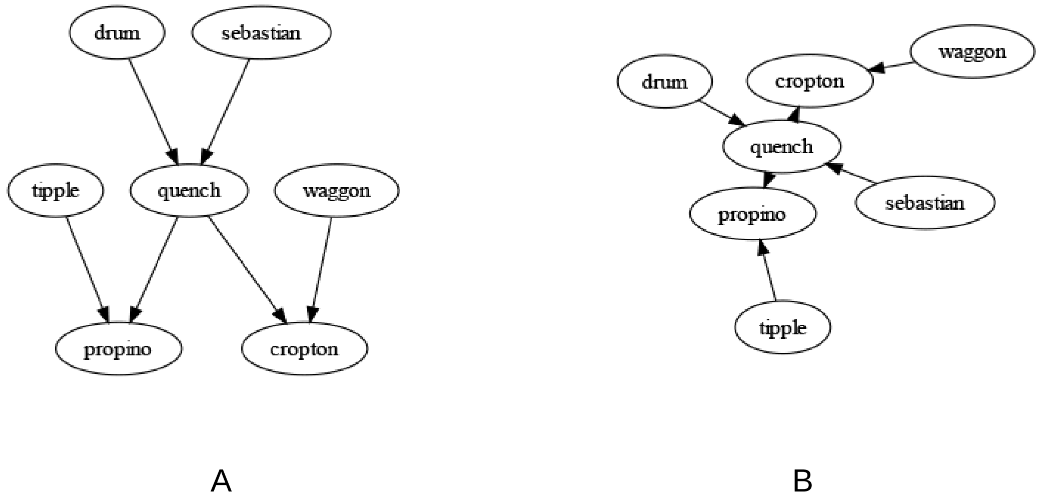


Figure 4-1 Sugiyama (A) and force directed layout (B). Force directed are unsuitable for pedigree layouts.

These abstract representations shown here include a time component in the form of generations, but due to the viability of seed, and the existence of varieties and landraces that can be many hundreds of years old, there is the potential to use these older varieties in modern crosses. This situation leads to nodes at the top of the graph having edges connecting to nodes at the bottom - this is not common in animals and would be extremely unlikely in humans. The existence of a time component means that the use of a layout algorithm that preserves topology (top-down generations) is nonetheless important as most (but not all) crossing will be between newer varieties. Because of this, layout methodologies such as force-directed algorithms (Figure 4-1B) would not offer the ability for us to arrange the pedigree based on time.

Taking the results from the initial requirements gathering, and drawing on the conclusions from the literature, a simple static prototype system was developed modelling and storing data in a DAG (Directed Acyclic Graph) format and layout using Sugiyama based graph layouts.

Graph visualizations (as opposed to the data structure) are used to encapsulate the relationships that exist between objects and the layout facilitates the visualization of these relationships and optionally additional overlaid information. When using layouts to present pedigrees, there are aesthetics which can be used to help aid the comprehension of the graph. Examples of these include layout simplicity, insomuch as the reduction of overlapping edges which leads to potential confusion over relationships between nodes and the ability to layer the nodes in some sort of metric such as time. This is particularly important in genealogical studies where the time aspect is of the essence. In addition the placement of objects on a page carries significant influence to how a user comprehends visualization. The use of layouts must therefore be carefully selected. Graph layouts allow the visualization of the connectivity that exists in datasets; however there is often a need to deal with problems relating to the lack of known parentage within the datasets.

In order to test if the use of a DAG based data structure and layered layout approach would work with the barley pedigree test datasets a paper-based layout was implemented which overlaid basic character data on to graph nodes. These character classes were represented by colour in the initial visualization. The main purpose of this initial static prototype was to establish if the tools that were selected to carry out this work were appropriate for the volumes of data and that users were happy using the Sugiyama layout to represent pedigrees.

The dot library from GraphViz (www.graphviz.org) was selected to perform the layout of the visualization and a Perl application was written in order to parse the pedigree definitions in Germinate (Section 2.6.2), create the graph structure (using Graph-0.96) and create the dot input files, run dot and finally show the output. In its initial form the program that was developed only handled the creation of simple diagrams (Figure 4-2).

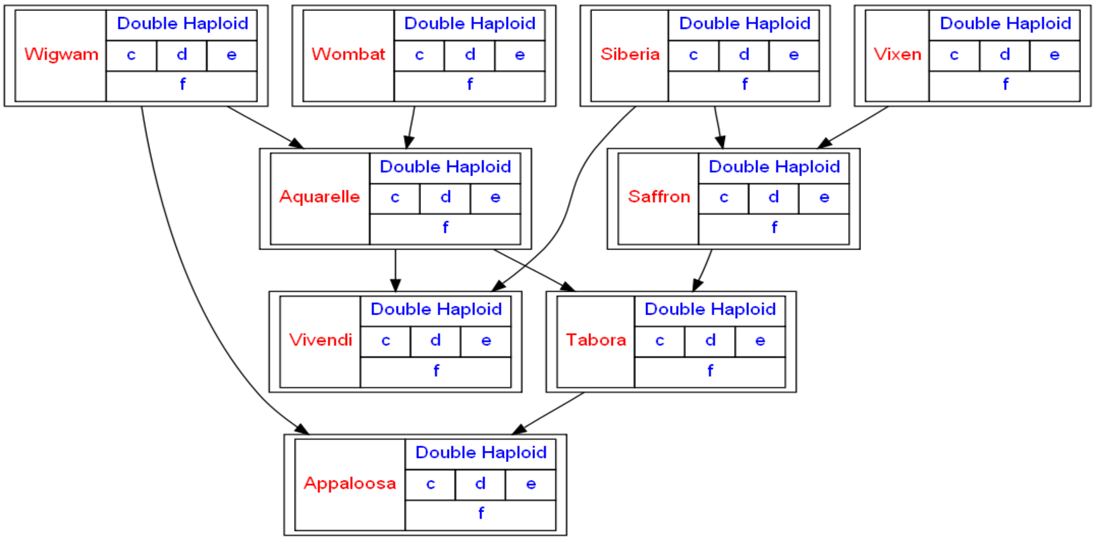


Figure 4-2 Initial Sugiyama style layout

The initial prototype involved the creation of rectangular boxes in which the plant line name, how it was generated and other information were added but on discussion with users it quickly became apparent that this was actually quite complicated and the general consensus was that it was confusing. Taking this into consideration, it was decided to simplify the graphic that represented a single plant line into a single oval shaped graphic which would contain only the plant line name. Initial impressions with end users suggested that this representation was both more visually appealing and easier to comprehend (Figure 4-3).

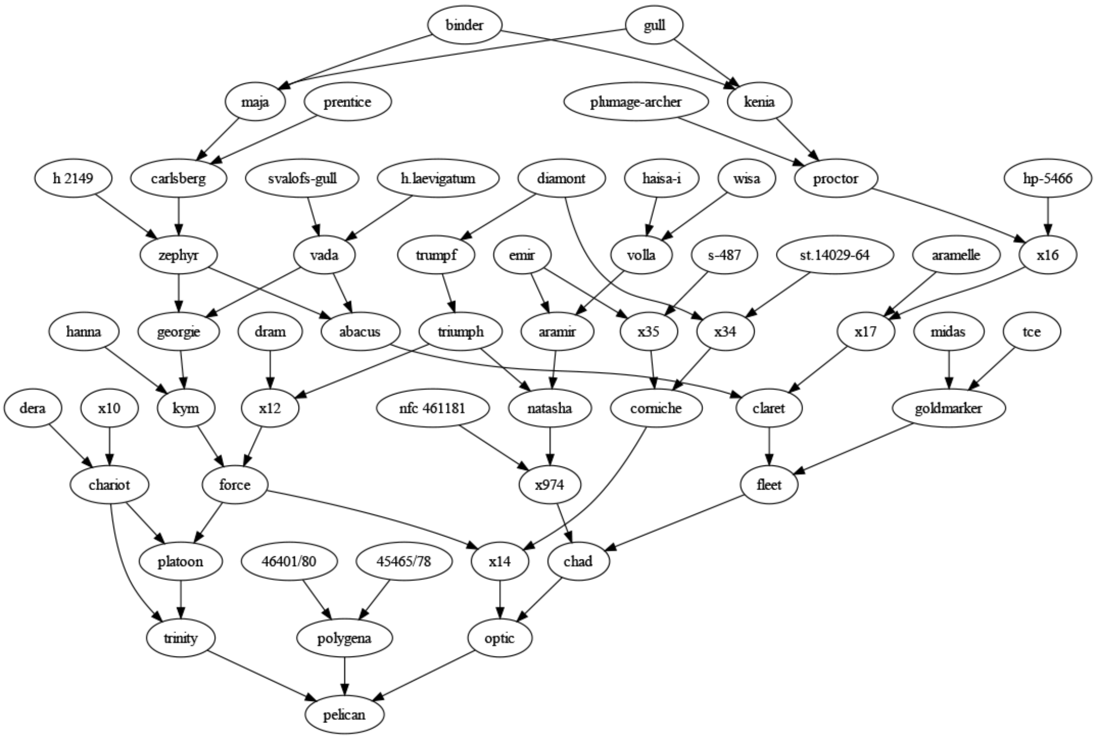


Figure 4-3 Simple Sugiyama style graph layout of example barley dataset.

Using these examples, informal discussions were held with potential users of this visualization and feedback gained after showing them visualizations such as those seen Figure 4-1, Figure 4-3 and Figure 4-3.

4.5.1 *First static prototype*

Users said that they preferred the round graph layout nodes to the square boxes and that the Sugiyama layout was easier to understand than that of a force directed layout. Figure 4-4 shows an early example using both round graph nodes as plant lines and uses colour (in this case red) to represent pedigree end points.

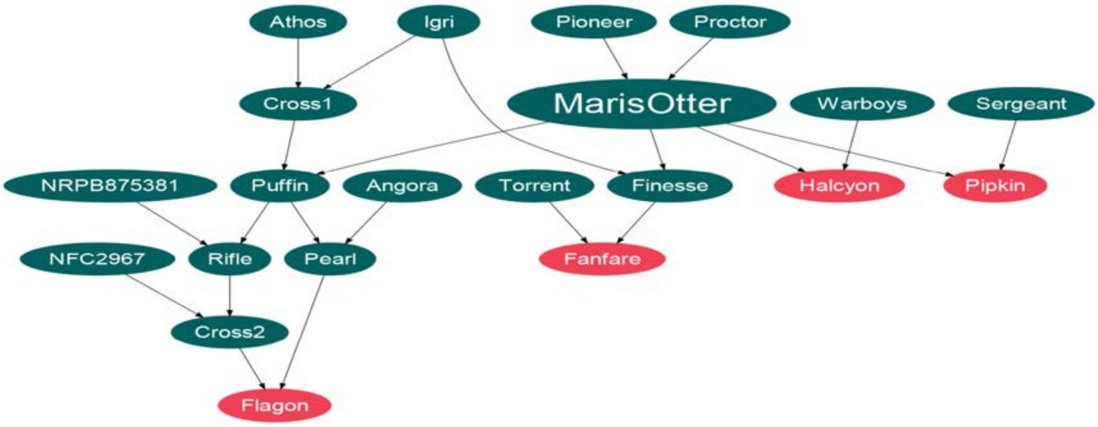


Figure 4-4 Early pedigree layout using colour coding to show plant lines (graph nodes) and relationships (graph edges)

The use of visual metaphors in this case relates to information relayed by graph nodes and edges. It can, and should be argued that the choice to keep these metaphors as simple as possible will help in the discovery of visual patterns more simply than by packing each visual element with as much information as exists. In most cases the simplification of output using accurate and concise data representations should lead to increased clarity of the high level processes that are being examined. This is something that would be tested with end users.

One technique in which the information that is being relayed to the user can be simplified is to reduce the amount of visual clutter using visualization techniques such as data-on-demand or selective highlighting to remove, or at least make less prominent, information that is not required to answer the questions that are being asked. This is important as there is a general trend in science (and in some infographics) to increase the amount of data to fill out white space creating visual impact through information overload in preference to lucidity.

4.5.2 *Graph nodes – the plant line*

Each graph node represents a population of a single plant line/variety. In the barley test dataset this is a representation of a population of plants which are inbred and genetically identical. These are usually named with either a varietal name; examples include ‘Prisma’ and ‘Golden Promise’ or breeders codes such as ‘NSL 98-4087’. Breeder’s codes are codes given to varieties before they are assigned a name used in promotion of the material commercially. The graphical variables that are available that can be changed for a graph node are colour, size, shape and position of nodes (Table 4-2).

Graph Component	Visual Variable	Encoding
Node	Position	Position in pedigree
Node	Size	Number of times used as parent
Node	Shape	Restricted to round nodes only
Node	Colour (Value)	Saturation – Quantitative value for selected trait
Node	Colour (Hue)	Qualitative value for selected trait
Edge	Size	Highlighting flow or edge selection
Edge	Direction	Indicating predecessor or successor

Table 4-2 Graph attributes (Ordered by salience)

The visual variables used in these graph representations are detailed below.

4.5.2.1 Node position

Described by both Bertin (1983) and Mackinlay (1986) as the most salient of the visual variables the position or more explicitly (in this case) the layer in which a node is found gives an indication as to its position within a pedigree context while this relates to a time series it does not relate to the age of a specific variety but only when that variety was used in time in mating events. While in general the higher up the pedigree the older the variety this is not always the case due to older varieties sometimes, but relatively rarely, being crossed with newer varieties and not appearing towards the bottom of the visualization.

4.5.2.1 Node size

Node size can be used to show a number of variables such as the number of times that a plant line has been used in subsequent mating events. What has been described as ‘direct sizing’, sizes nodes based on the number of children that are derived from a plant line and ‘overall-successor sizing’ sizes nodes based on their overall pedigree contribution which will include children, grandchildren and all subsequent generations for which data exists. Node sizing is user definable to allow definition of cut-off values which meet particular users’ requirements but will adopt default values based on analysis of the loaded pedigree file.

Node size was used as it is known that the sizing of visual elements is a good indicator for quantitative data types (Bertin 1983). Bertin’s well documented and accepted systematic system for utilising 7 visual variables (Position, size, shape, value, colour, orientation and texture) to aid a users’ perception of differences was later added to by Mackinlay (1986).

4.5.2.1 Node shape

While described by Bertin (1983) as one of the most salient visual variables, in this work, all individuals are represented only as round nodes within the graph representation. There are no other changes to overall node shape in the graph visualizations produced.

4.5.2.2 Node colour

Colour means many things. It not only allows us to tell the difference between graphical elements but can also highlight these elements as important, or not so. It can be used to draw or attract a user’s eyes to areas of a visualization that are deemed most important or move them away from areas deemed not so. There is a great deal of research published on visual salience (which does not just include colour) (Treisman and Gelade 1980; Wolfe and Horowitz 2004).

The use of appropriate colour palettes that are not only suited to the visualizations that are being implemented but also taking into account problems that some users have with colour perception difficulties, however, it is important that sacrifice is not made to the ability of the system to deliver results to end users. The appropriate use of colour palettes, such as not using the clichéd red-green gradients should help users with colour perception problems.

Node colour was used to show phenotypic characters (nominal and ordinal data types) and genetic similarity data.

Phenotypic characters were retrieved on demand from the Germinate 3 database or from text-based input files. For this part of the work they were all nominal data types with a defined number of descriptive identifiers. The pedigree visualization coloured all the graph nodes based on a single character to maintain consistency and visual comparisons across entire datasets. This would in return try to help avoid confusion with inconsistent colour schemes.

4.5.3 *Graph edges – mating events and showing genetic flow*

Graph edges represent mating events. Each node or plant line has a male and a female parent. The male and female parent are either different plant lines, or they are the same plant line. In the prototype, edges from nodes that have a lot of children are darkened to help show the relative importance of the plant line in breeding programmes (Figure 4-5).

Edges can display additional data such as highlighting predecessors and successors or coding male and female parents if such data exists. The use of edge bridges will improve clarity and reduce confusion of direction and edge tracing. The use of edge weighting was also used to emphasise plant lines which had been frequently used in crosses to make these plant lines more prominent in the visualization (Figure 4-5).

4.5.4 *Information layout - putting everything in its place*

Due to information density issues when working with large pedigrees it is not always feasible to display all nodes and connecting edges in such a way that the specifics of genetic linkage between the nodes are clear. However, at this overview level, this is acceptable as it is not the case of trying to give users the exact specifics of the pedigree but rather a representation of its overall structure. Major data trends within the data can however be easily and clearly shown within the datasets. Specifics relating to edges are displayed when zoomed in on the data of interest.

4.5.5 *Orientation assumptions*

Much as in human family trees, it can be assumed that visualizations will run from top to bottom. This is the standard way of representing genetic relationships in humans and farmed animals. However, unlike in family trees where each level represents a particular generation, in plant pedigrees, while this is in the most part also true, it can be argued that even a radial representation of the pedigree would be appropriate as the time aspect may not always be critical in the establishment of new genetic crosses. It is easier when dealing with plants to maintain genetic stock, something which is limited in humans and animals by the relatively short time window in which they can reliably generate viable gametes. It can be strongly argued that it is more important to easily identify plant lines that have desirable alleles as opposed to their positioning in time. Taking this into consideration, under certain circumstances, the use of radial layout algorithms would reduce the visualization space needed. The use of arrow heads would maintain direction in this scenario. This is not to say that the use of a radial layout is appropriate in all circumstances, but where the number of plant lines is limited (for example less than a dozen) it would offer an alternative view on the data.

4.5.6 *Interactions; accepting that scrolling is a reality*

When represented with a vertical time axis (the newer plant lines appear towards the top and older plant lines towards the bottom of the visualization), the dataset used in this work amounts to a size of around 70 nodes wide on the horizontal axis by 14 nodes deep. This equates to a 5:1 aspect ratio which clearly is very different from the 4:3 and 16:9 aspect ratios commonly seen on desktop computers and mobile devices. When taking into account the display of plant line names on nodes this ratio increases dramatically resulting in a visualization of the pedigree which is far wider than it is tall. This is problematic for displaying using standard desktop monitors. You cannot see everything at once (with anything but the smallest of pedigrees), or more accurately, you cannot see everything at once in meaningful detail which further highlights the requirement for overview and context techniques. This is why an overview and context system is necessary for these sorts of data.

The physical size of the example pedigree datasets (650 nodes) means that there was an inevitable degree of scrolling required around any application and visualization that is developed if it is represented on a computer monitor. While this is not ideal the combination of a scrollable area and the generic overview will allow users to maintain knowledge of where they are in the dataset and prevent problems leading from losing your position in the dataset. The scrolling problem should also only exist in the high-level overview as there should be sufficient screen space when looking at more detailed portions of the pedigree graph.

4.6 Automated prototype pedigree visualization using Perl and Graphviz

An application was written in Perl using the Graphviz libraries and Graph.pm module ("Graph - Graph Data Structures and Algorithms - Metacpan.org" 2014) to generate dot input files from text files containing pedigree and phenotype definitions.

The application takes the form of a series of simple Perl modules for dealing with the input data types and generating the DAG structure used for modelling the pedigree data. In addition the application contains methods for generating colour schemes and simple routine analysis on the input data files such as node numbers and identifying singletons where no parents, nor progeny, existed.

This application was named 'Orb'; a reference to the circular like appearance of the early prototype visualizations.

4.7 Using more advanced datasets with Orb

Using the Perl application detailed above a more advanced dataset comprising 650 barley varieties was used to generate larger, static pedigree visualization. Figure 4-5 below shows the output from this data with the barley ecotype winter/spring colour coded. Nodes with a high number of outgoing edges (this means that the plant line is used more frequently in crosses) were increased in size and the outgoing edges were darkened to make them stand out more from the surrounding varieties.

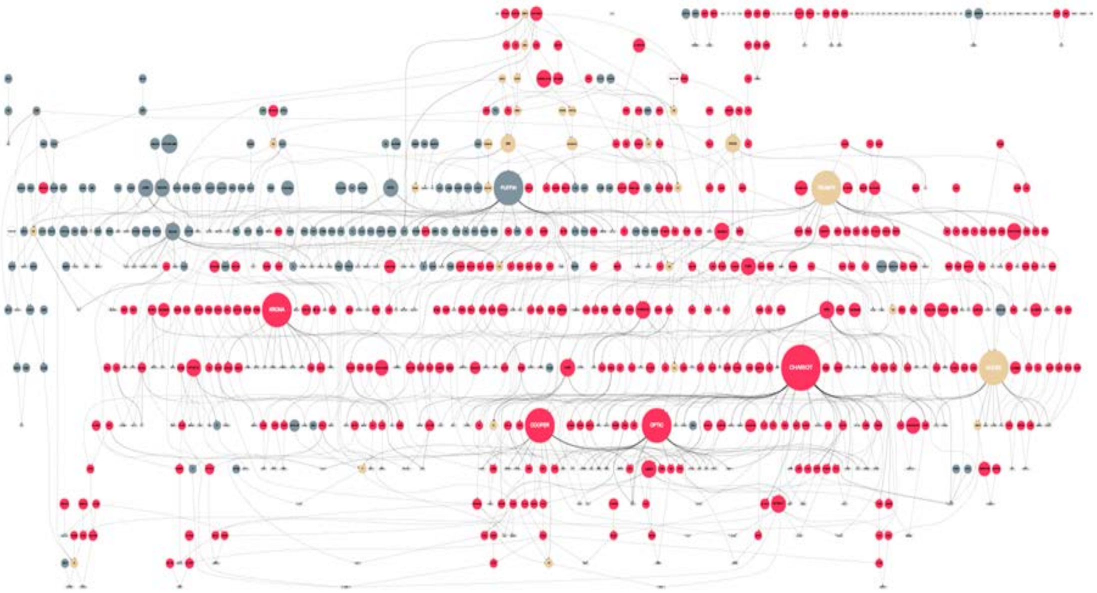


Figure 4-5 Pedigree visualization static prototype. Colour is used to distinguish the winter (blue)/spring (red) barley ecotype with cream nodes being plant lines in both winter and spring pedigrees. Size is used to show the number of times a plant line has been used in subsequent crosses.

This was one of the first attempts at visualizing the entire barley pedigree. Node size was used to show the number of times the plant line has been used in crosses that have given rise to progeny that have been successful in National List trialling in the UK.

This is the first time that a pedigree involving this number of commercially released barley plant lines has been brought together in one place and sparked interest with commercial plant breeders when they were presented with it.

Additional examples were also created using the Orb application to show genetic relatedness (Figure 4-6). The darker the green the more genetically similar a plant line to a chosen 'base' or proband plant line.

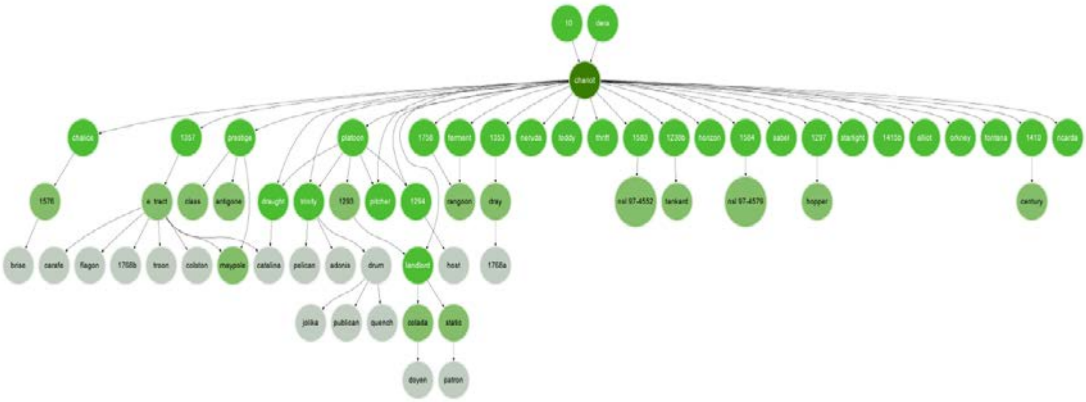


Figure 4-6 Overlaying genetic similarity with Orb. Colour is used to show genetic relatedness based on a similarity matrix generated from underlying SNP data. The focus node (proband) is Chariot and dark green. The darker the green the more similar the plant line to the proband.

4.7.1 *Using Orb to generate visualizations for other species*

Using data obtained from the International Rice Research Institute (IRRI) in the Philippines and The Wheat and Maize Improvement Centre (CIMMYT) in Mexico the Orb software was used to generate diagrams similar to those in barley for both wheat and rice. This showed that Orb was useful across species highlighting the generalization of such a visualization tool.

What is interesting about these visualizations (Figure 4-7 and Figure 4-8) is that the structure is very different to that of barley with backcross ‘chains’ easily recognisable (Figure 4-7). These types of breeding features are not seen in the barley datasets used in this work.

In addition to the wheat and rice data it was also possible to run Orb across pedigree data from a pig (*Sus domesticus*) F2 population (Figure 4-9). This is the *only example presented in this work in mammals* which highlights the very different topological structure compared to plant data.

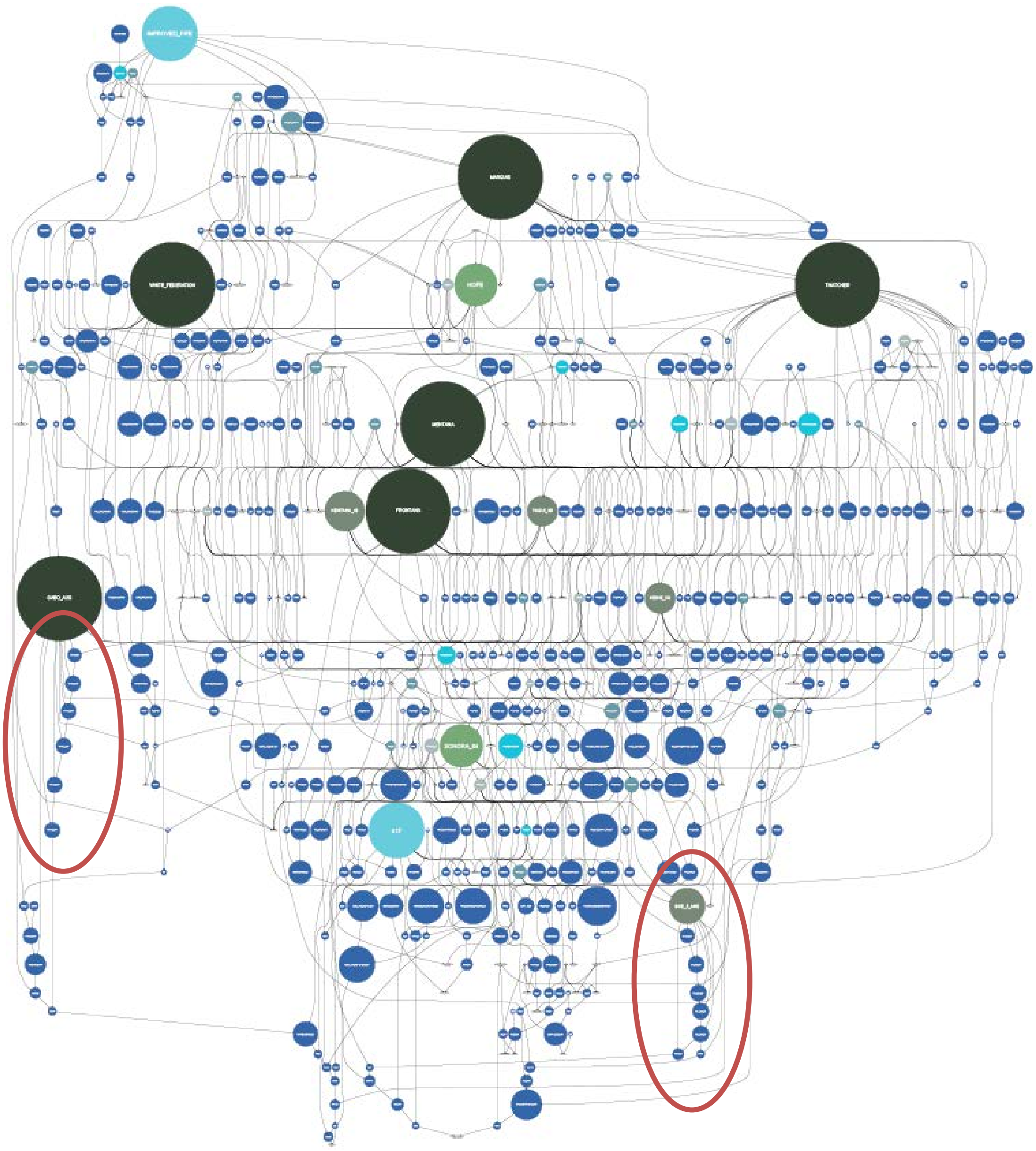


Figure 4-7 Wheat pedigree visualization. Backcrossing ‘chains’ highlighted in red.

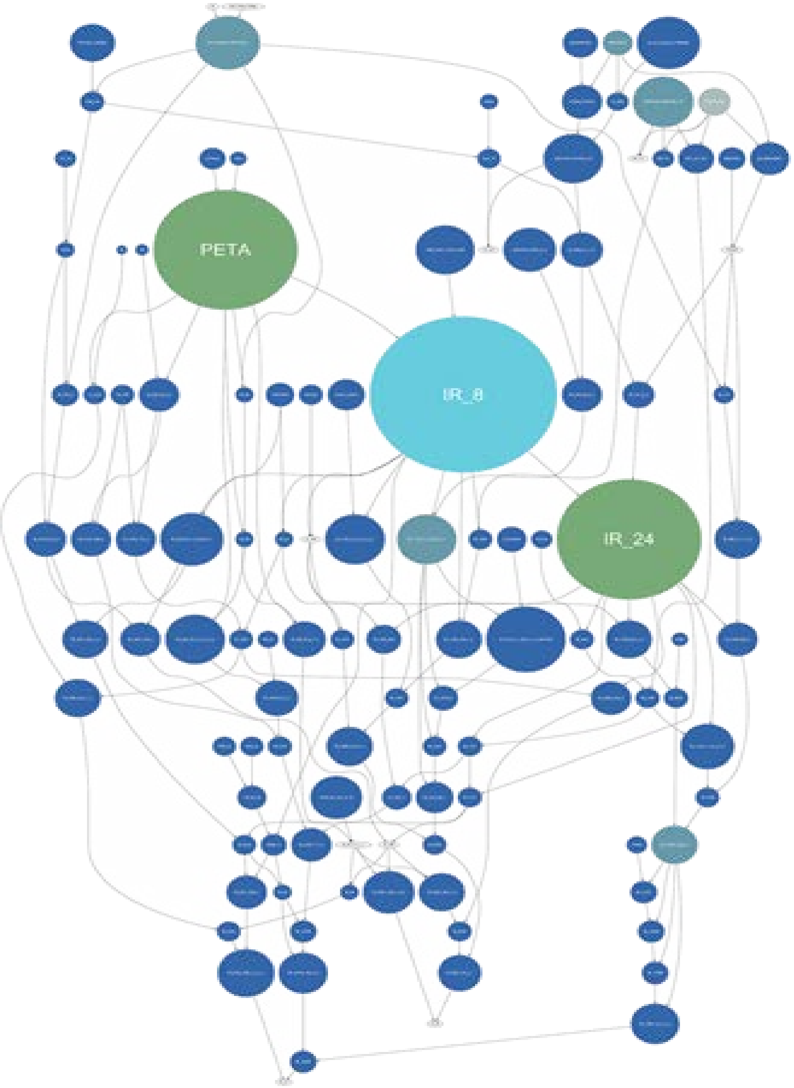


Figure 4-8 Rice pedigree

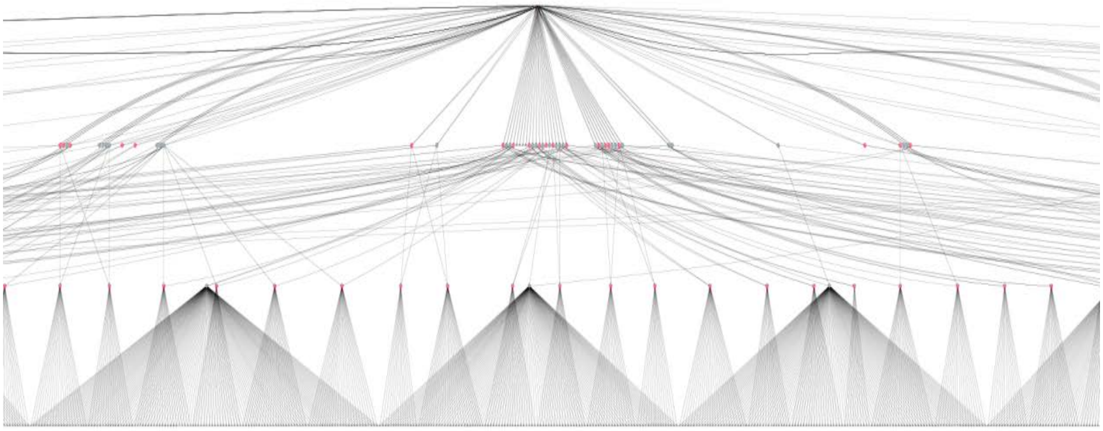


Figure 4-9 Pedigree visualization for pig F2 population

As described, Figure 4-9 shows the representation of a typical F2 (second filial) population in pig. The pedigree shows that the approach used in the visualization of plant pedigrees are not suitable, in its current form for animal populations due to the horizontal space that such representations use. The visible representation here in Figure 4-9 is only 20% of the total pedigree that was available. This representation for pig, which would yield a similar structure for other domesticated animals shows that the layouts used for plants are not suitable in this domain. The shape of these animal pedigrees have been coined ‘*pyramid type*’ or ‘*delta*’ pedigrees due to the top-down triangular shape to the data when laid out using Orb.

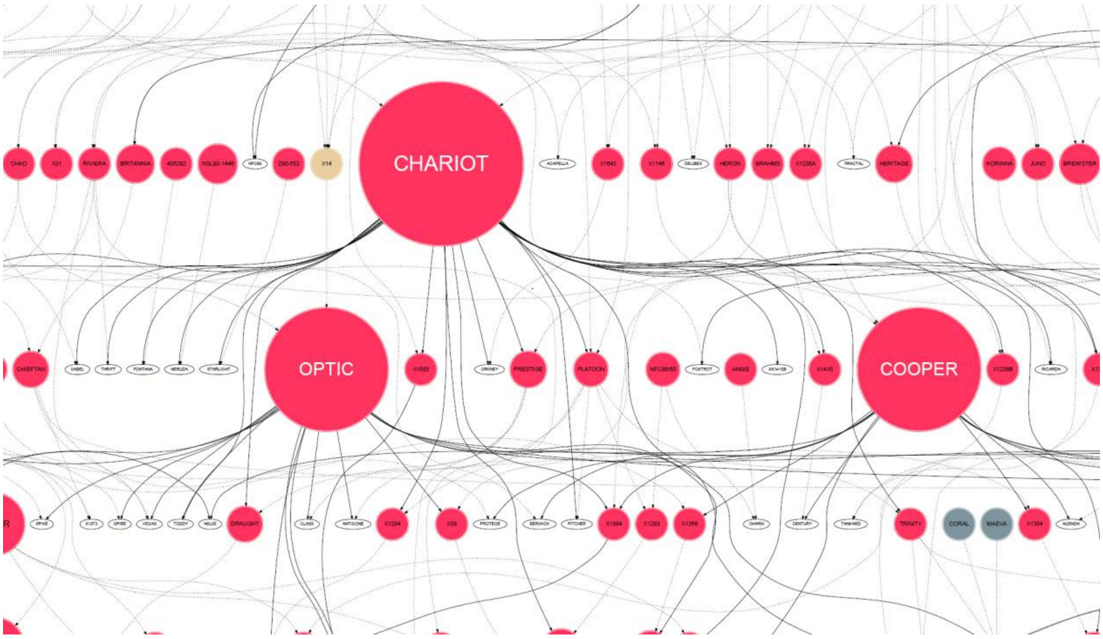


Figure 4-10 Zoomed pedigree to show complexity

While these pedigree visualizations were visually appealing and well received by users when viewed on computer monitors it was decided that printing this static representation at a size of 2.5m x 1m (Figure 4-11 and Figure 4-12) would allow domain experts to better interact with the visualization. The complexity of the layouts is better highlighted in the zoomed in representation shown in Figure 4-10. Information was overlaid, by means of colouring nodes, the winter/spring ecotype category on this dataset as (along with the 2-row/6-row ecotype) it is the most commonly used physiological means of differentiating barley varieties, and one that all of the test users were familiar with. This tool was also implemented as a web-service which allowed us to include static (but dynamically generated) pedigree representations within the internal Germinate barley database (Figure 4-13).

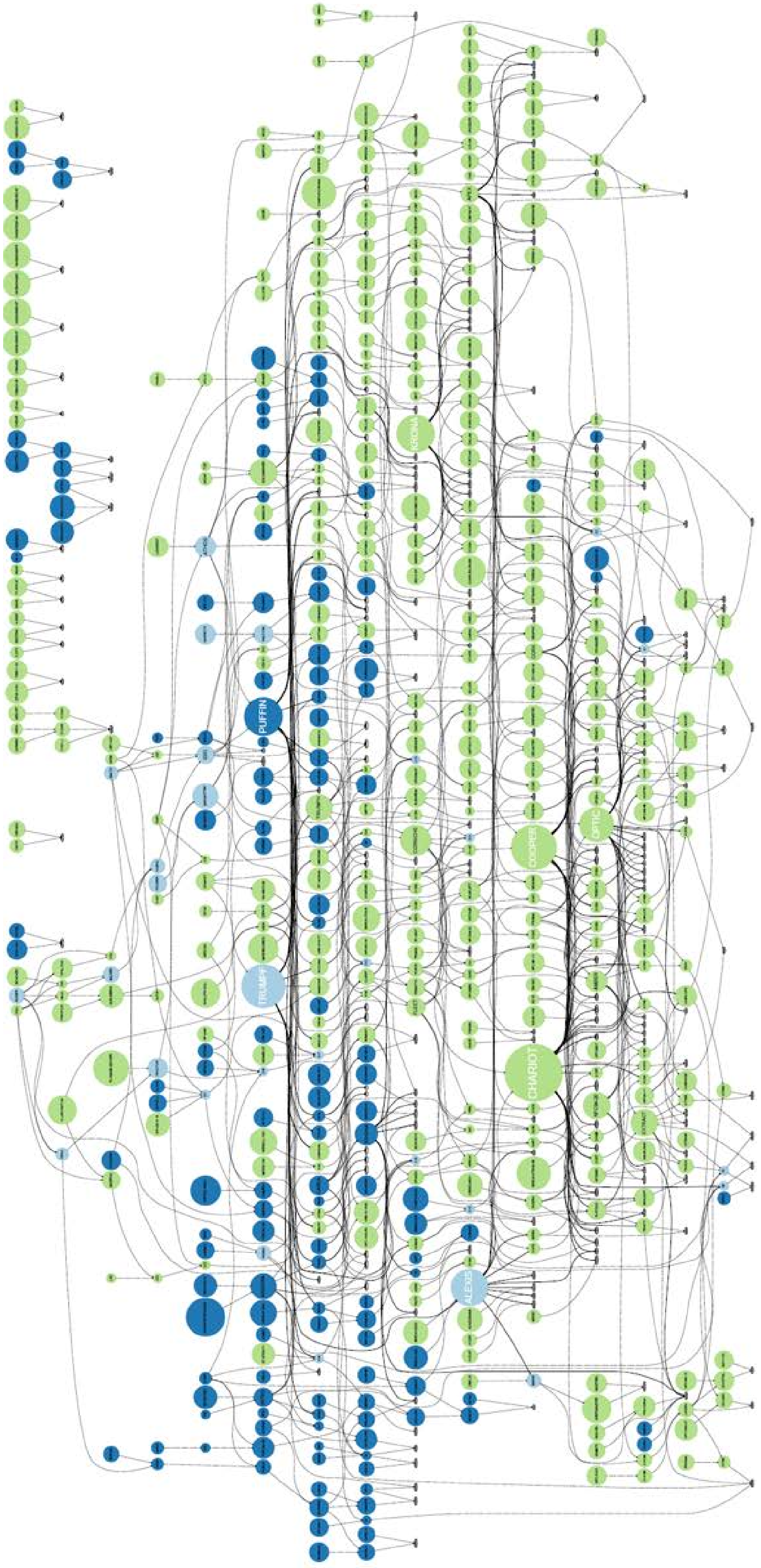


Figure 4-11 Barley pedigree created with Orb showing winter (blue) and spring (green) ecotypes. Plant lines that exist in both pedigrees are light blue in colour.

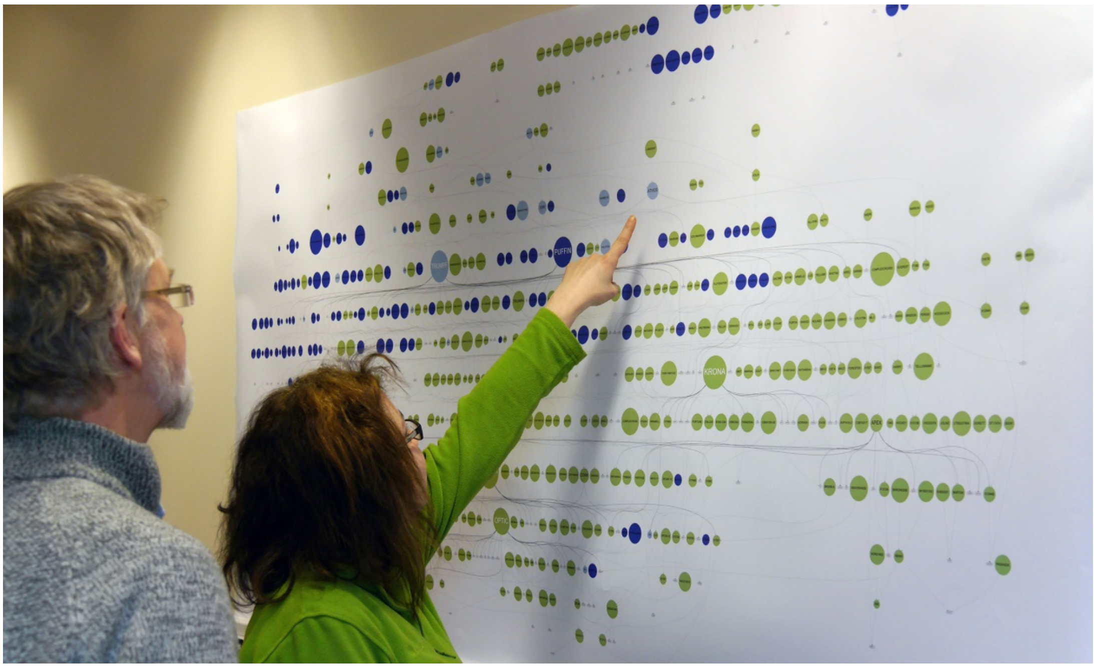


Figure 4-12 Users interacting with large pedigree visualization

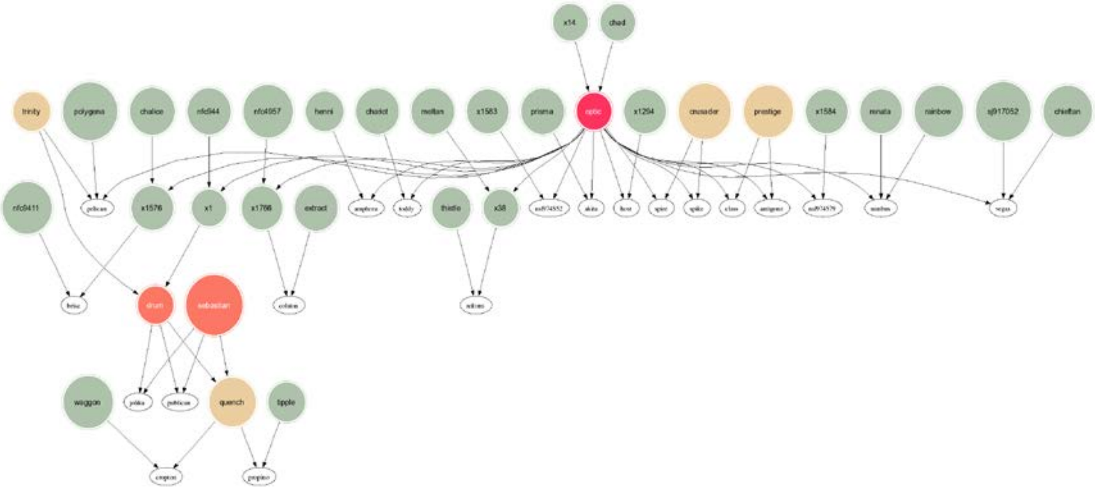


Figure 4-13 Successors for the plant line Optic created from Germinate

4.7.2 *Feedback on paper-based prototype*

Through observation and interviewing twelve geneticists and plant breeders while they interacted with the wall-mounted visualization it was clear that there were a number of issues associated with this implementation. Firstly, it was almost impossible to trace edges between nodes when the data was dense (even at a large output size) so we found ourselves falling back on examining text based records at a PC to confirm lineage.

Secondly, it is incredibly challenging to quickly locate specific plant lines with this density of data. With upwards of 500 nodes on the visualization it was almost impossible to find specific plant lines. While commonly used plant lines are immediately identifiable due to the use of size to represent the number of uses in breeding crosses, these are not always what users are most interested in. Users used these larger nodes as reference points, almost as if they were notable points on a map (Dieberger and Frank 1998; Muller et al. 2005) and attempts at using slightly different layouts or orientations were not well received. Figure 4-14 and Figure 4-15show the same data as in Figure 4-5 but in an inverted and left to right orientation layout. Figure 4-14 puts more recent plant lines at the top of the visualization and at the right of the image in Figure 4-15.

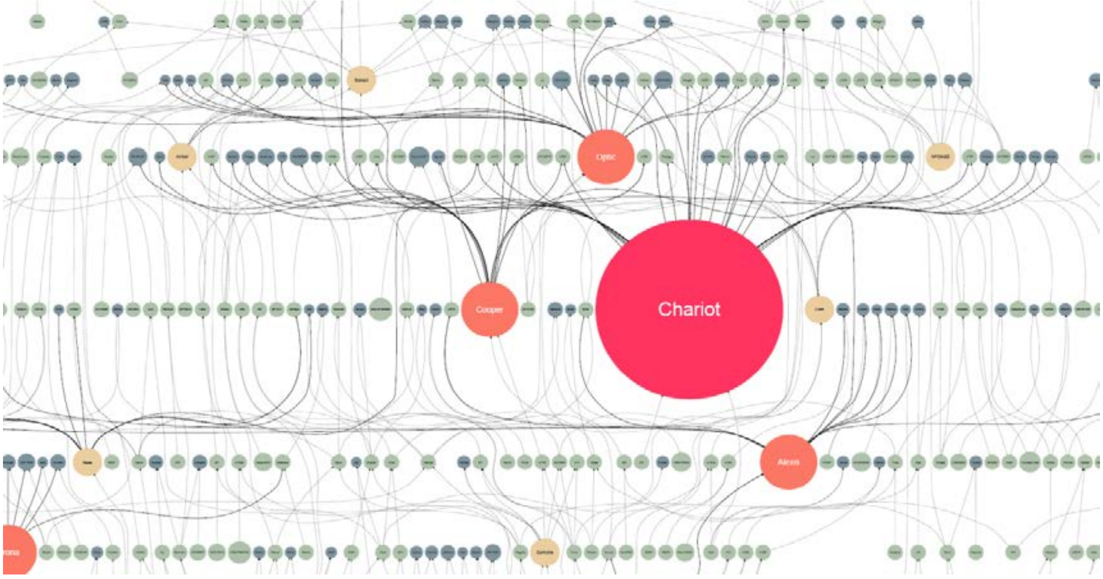


Figure 4-14 Top down (inverted) pedigree layout

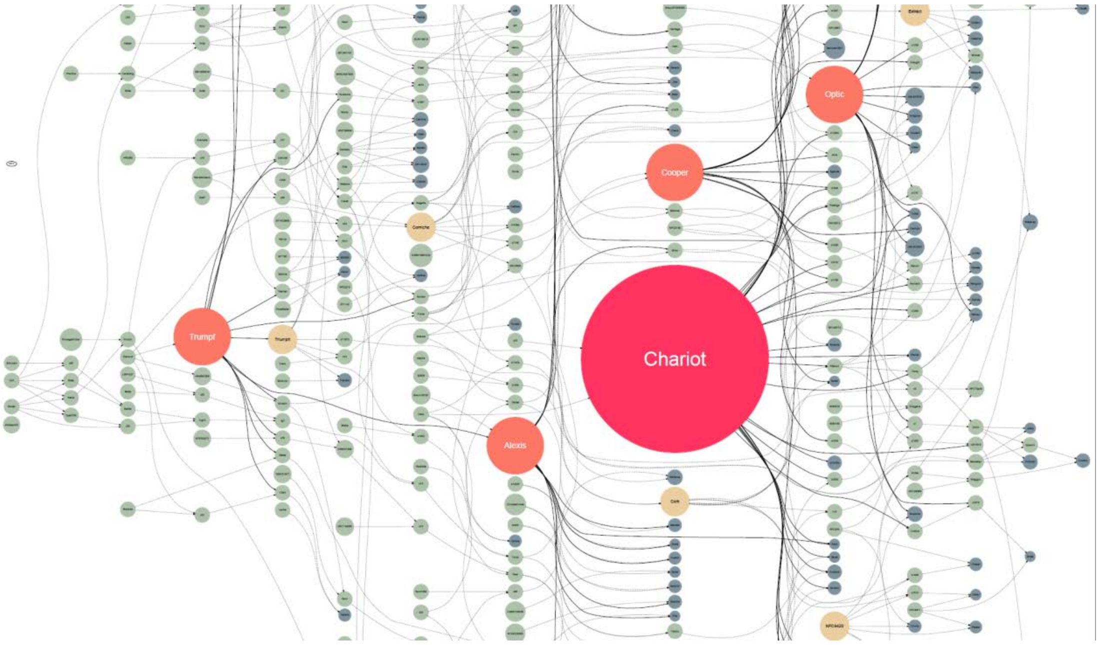


Figure 4-15 Left to right pedigree layout

Additionally, it was also clear that users were beginning to quickly spot pedigree problems. These problems related to the parentage of plant lines and in some cases the assignation of ecotype. These types of errors would be extremely difficult for a user without extensive experience to pick up on. This has not only shown that it is an effective technique for visualization but also an effective way of identifying errors with underlying datasets. When these errors were picked up by users the underlying data was changed as appropriate to reflect and adapt to the problems identified. At first while it is somewhat disheartening to see users finding so many problems with the data it highlighted, as clearly as possible, the problems with the underlying data and the errors it contained. It should be stated that this data has been being used in plant breeding and genetics experiments and in some cases is fundamentally flawed. This was a good opportunity to begin to fix some of these legacy issues.

4.8 Discussion

This prototype has shown that plant pedigrees often form what can be described as a ‘pedigree net’, whereby there is structure to the graph but it’s not as simple as traditional top-down pedigree representation that is seen in humans and to a lesser extent in farmed animals (Figure 4-16).

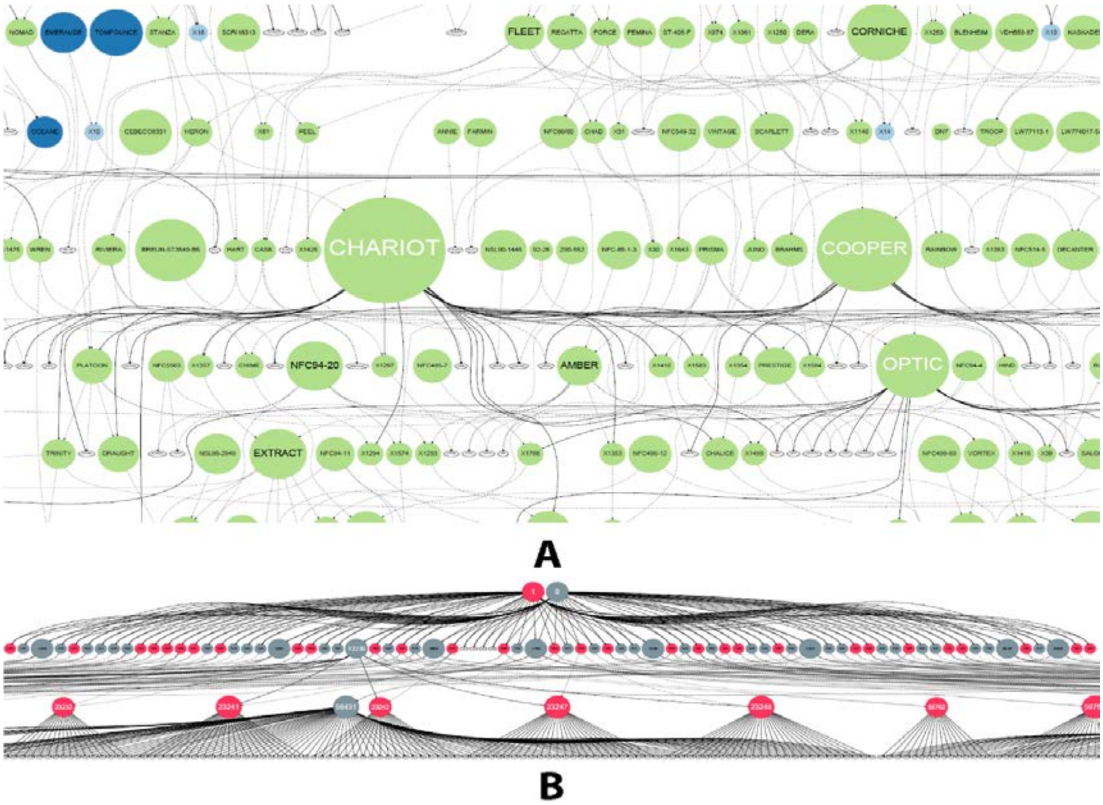


Figure 4-16 Plant (A) and animal (B) pedigree structure

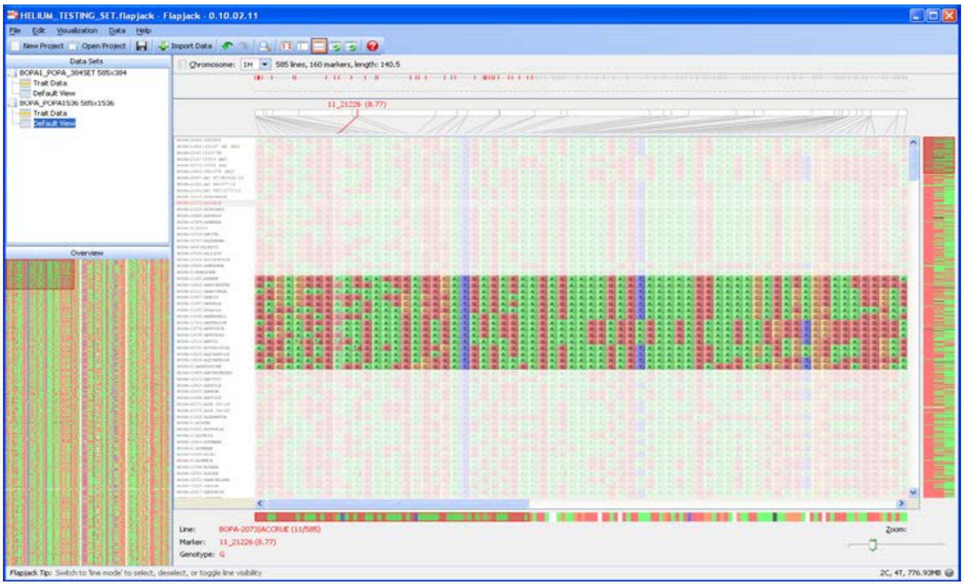
Feedback from users was that they liked this representation of large pedigrees. Not only is it visually attractive, but geneticists were using it to identify problems with the underlying pedigree and phenotypic data in a way that is more interactive, social, and tactile compared to the examination of text based records.

When presented with these results, plant 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’ plant lines in the UK Elite Barley germplasm that would have been difficult to see from textual records in the format seen in Figure 2-9. Missing data was also easily spotted thus allowing updating of the underlying datasets. Problems do however exist, especially in the inability to search for particular plant varieties and tracing of edges to establish lineage. In order to try and address these, it was quickly realised that a more interactive software tool would be required to address the problems that people had with the early static prototype.

One of the more hard-hitting measures of success of the first paper-based prototype came from the presentation of data to a meeting of UK plant breeders. While the pedigree data that was demonstrated was available to all in the room as written records the representation that was presented had a major impact through the provision of new insights as to how germplasm was very closely related. When written as a text string it is difficult to construct the bigger picture, but when displayed in this initial prototype, the relationships between competing breeders plant lines was much more striking. While this was privately known to the individual breeders, having it presented to them when they were all in the same room was very enlightening. This not only highlights the value of visualization but that a visualization tool with real-world impact has been implemented.

One of the considerations in the construction of the paper-based prototype was in the contributions of both geneticists and plant breeders in order to get a balanced viewpoint from both user groups. There was an inherent bias towards plant geneticists in the undertaking of this part of this research due to availability of staff. However, having enthusiastic end users available for feedback sessions and the continual input from them in the development of this static prototype was a considerable plus and a significant resource that not all projects have access to, or indeed utilise to the same extent.

Finally, as a proof of concept, tools were developed to link the Orb application to the Flapjack (Milne et al. 2010) graphical genotyping application to allow Flapjack users to select germplasm plant lines then visualise these in a pedigree context alongside the graphical genotypes for those plant lines. This allowed Flapjack to call a web-service which returned the static pedigree visualization which Flapjack can display within its graphical user interface. Figure 4-17 shows the selection mechanism for plant lines within the Flapjack application and Figure 4-18 the resulting image returned from the pedigree visualization tool.



585 lines, 160 markers, length: 140.5

Line: BOPA-2073|ACCRUE (11/585)
Marker: 11_21226 (8.77)
Genotype: G

Flapjack Tip: Switch to 'line mode' to select, deselect, or toggle line visibility

2C, 4T, 776.93MB

Figure 4-17 Plant line selection for pedigree visualization in Flapjack. The selected plant lines from within the Flapjack application are highlighted in an Orb representation (Figure 4-18)

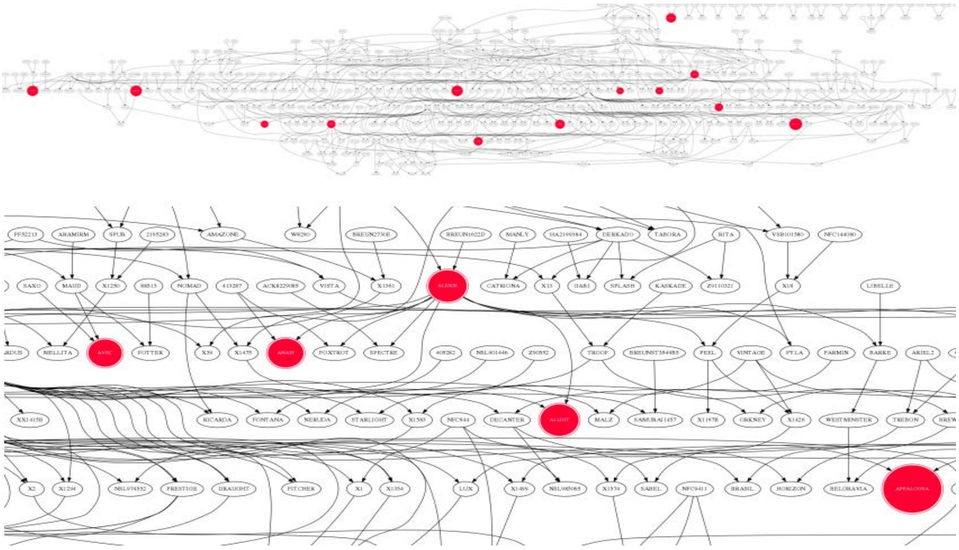


Figure 4-18 Pedigree visualization of selected plant lines highlighted in red.

The feedback from this initial static paper-based prototype showed that there was use for such a pedigree visualization tool and so the next stage was to begin to develop the visualizations to introduce interactive features.

Chapter 5 The Helium Prototype

5.1 Overview

The initial paper-based tool was effective in engaging users with pedigrees and has proven to be a valuable tool in the identification of errors in the underlying datasets. There were things that people liked about this prototype and things that people disliked but the main limiting factor was that it was static and users needed to be able to interactively explore their data.

Taking the feedback obtained from the initial informal user testing, an interactive detail and overview (Andy Cockburn, Karlson, and Bederson 2008) prototype pedigree visualization system was implemented.

The prototype was named 'Helium' after the balloon like appearance of the nodes and edges from the previous static prototype (Figure 4-5). It maintained the same visual metaphors described in Section 4.5 to describe pedigree structure as the static paper-based prototype, but added additional features to allow users to search and explore the data and link in plant passport, phenotype and background data from the Germinate database as well as facilitating the ability to quickly search for plant lines of interest within a complex pedigree structure.

5.2 Application Design and Development

In order to develop a prototype there were a number of decisions that needed to be made in relation to how a usable and testable application is developed. There were a number of components to the system that had to be developed that both interact with each other and provide a usable framework on to which more complex analysis and visualization tools can be built.

The basic functionality was that users should be able to visualize pedigree data loaded from external files or from a suitable Germinate database; this being the preferred option.

The application design was split into a number of key areas which were then designed and implemented to give basic core methods on to which more detailed and specific functionality would be added after user feedback and testing sessions.

A user centred iterative design process was used in the development of the Helium prototype. Due to being employed within a plant genetics department, and heavily involved in additional projects involving European plant breeding companies and academic partners, there was daily contact with plant geneticists, and regular contact with commercial plant breeding companies both through email and during organised project meetings at which this work was routinely presented. Prototypes were also shown at conferences both within the biological and visualization domains. Feedback was also gained from one-to-one sessions which were held with breeders to show them in greater detail the work that was being undertaken. Additionally, sample datasets were obtained from breeders to test the visualization but these were commercially sensitive and not represented in this work. What this allowed was the testing of Helium on data out with the barley datasets detailed in Section 2.6.

The daily interaction, constant dialog and iterative design process ensured that features added were directly beneficial to end users and development was in line with the precise feedback from domain experts.

This constant feedback meant that an Agile software development methodology could be used which facilitated the evolutionary development of the Helium application. It was important that working prototypes were used for demonstration and feedback throughout the development process to ensure that development was both heading in the right direction and was meeting the research requirements of end users. It was also common to hear users have additional ideas when presented with the prototypes, this software development approach allowed the incorporation of these ideas where appropriate.

5.2.1 *Design*

The overview provides a high level overview of all the data in the pedigree being examined. In the test dataset this would mean a layout representation of all the barley plant lines. The overview can be colour coded for a single parameter such as the winter spring genetic divide or DUS characters loaded from Germinate and node sizing can be enabled to draw emphasis to particular plant lines that are commonly used in crosses. In addition, the use of emphasis such as in the changing of thickness of line around a node to highlight a plant line of particular interest will allow the user to determine where it sits in relation to the other plant lines in the pedigree as well as offering a reference point for the other data display panels in the application. A selected node (plant line) becomes the focus for all other displays within the visualization insomuch as the information contained in additional displays would relate to the selected node.

The overview shows nodes and edges to show genetic structure and selectively highlight edges based on the selected nodes to show related plant lines. The amount of detail at this level is severely restricted and is intended only to give a broad overview of the dataset and its size. Details such as plant line names were omitted from this display as they are too small to read and therefore serve no useful purpose. DUS characters are selected from a drop-down list which can be changed on the fly by users.

5.2.2 *Detail level 1*

This level shows a more detailed layout of the pedigree based on a single plant line selected on the overview display. Moving the overview windows will update this display to show the plant lines that are under the highlighting box on the overview visualization. Once a plant line is selected this display will update to show all ancestors or descendants based on options shown in the overview dialogue. These plant lines will be colour coded based on the overview colouring scheme but there will be an option to subsequently colour this display based on other parameters relating to the dataset such as phenotypic or genotypic data. All plant lines at this zoom level will be visible and names clearly displayed within nodes. Edges will also be clearly displayed. This can be looked at as a detail and overview stage.

In detail level 1 there are a number of data types that can be represented using the node.

1. Varietal name or other naming convention.
2. High level overview of some phenotypic or genotypic characteristic.
    a. This could be something like resistance to a particular pathogen based on experimental work or;
    b. The existence of a particular allele or haplotype at a given genetic locus identified by genotyping or sequencing data.

The attributes of the graph node that are available are as follows; node position, node shape, node size and node colour.

For clarity focus was only on node size and node colour having already established that users were happier with round node aesthetics. Node position was determined by the yFiles Sugiyama type graph layout algorithm.

5.2.3 *Level 2 detailed data view*

Detail level 1 showed a very general overview of the data which is held in the dataset and subsequent zoom levels may, or may not; include additional information which would not be possible with the high level overview level. This data level shows background information about a plant line and displays the data in a data panel, thus forming a classical details on demand design pattern described by Shneiderman in his visual information seeking mantra (Shneiderman 1996) and the term originally defined by Kreitzberg (Kreitzberg 1991). Selecting a node or edge of interest on the graph representation of the pedigree will trigger the retrieval of additional information from the database backend. Due to the potential volume of additional data this data panel will most likely be a tabbed pane with panes facilitating the logical split of additional data types.

Examples of data that are displayed include background information on the plant line, background information on genetic markers relating to the plant line or calculated phenotypic data such as field trial results. The details will very much depend on the context in which the pedigree is being examined but the following data types currently exist.

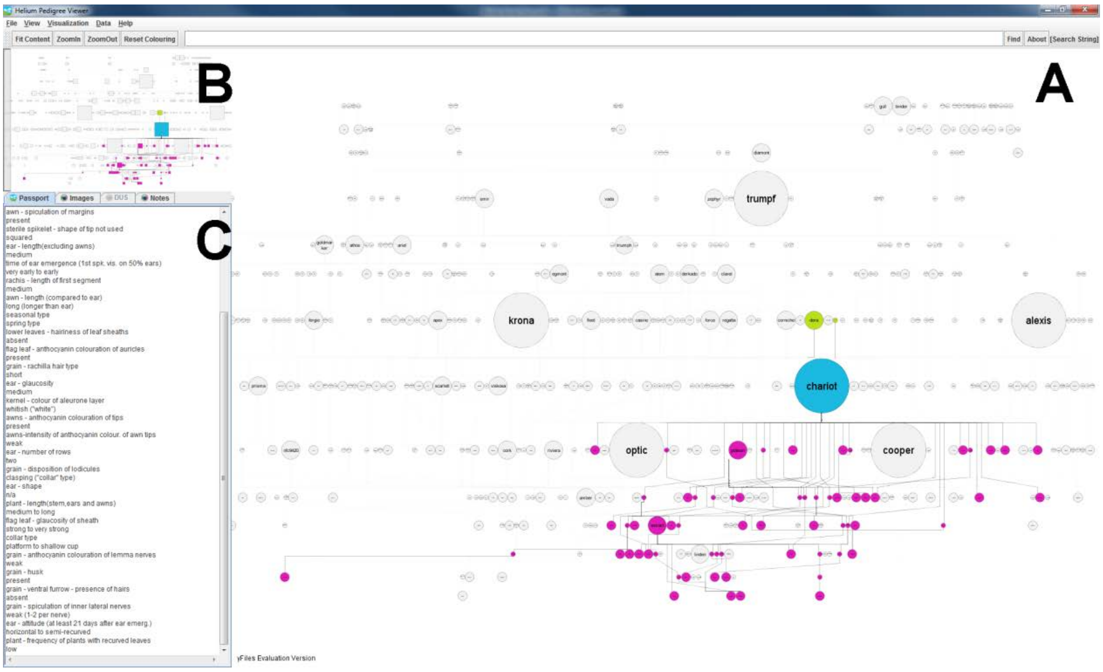
It has been already shown that there is a large volume and diversity of additional data that is held on each plant line although the coverage is somewhat sporadic (Section 2.6). While most of this data should be displayed as additional information in its own visualization space there are categories such as phenotype data where this information can be overlaid on the original pedigree representation. This also extends to include things such as breeder information which can be overlaid to give a representation of which breeders are responsible for which plant lines.

To summarise; there are a large number of additional data types that are important in the comprehension of the pedigree data. These data types form background information that is important for breeders to know in order effectively analyse the data for important information, trends or patterns. These data types may be represented by overlaying the pedigree view with the information in the case of simple data types or by the creation of custom views showing additional visualization methods. The data that is displayed should represent the plant line of interest on the main pedigree view and change, or be clear, if nothing is selected. This is to try and help avoid confusion with the visualization.

5.3 Implementation of first interactive prototype

The first implementation of Helium was developed and the layout of the application interface seen in Figure 5-1. While the paper prototype included a single static image it was clear that when users were viewing the visualization on computer monitors there would be a limitation on the number of nodes that could be displayed while still retaining legibility of plant line names. To address this, the main visualization panel (Figure 5-1A) can be zoomed and panned to allow users to explore data. An overview panel was added (Figure 5-1B) which would allow users to track where they were in the main visualization window and give a high-level overview of the pedigree structure. The overview would act as a common reference point for users that would not change as the main visualization window was manipulated. Feedback from the paper implementation also showed that users would want to get as much background information as possible on plant lines and so a detail panel was added (Figure 5-1C) which displays passport and general background information. Data from Germinate is displayed in the detail panel and is pulled on demand based on a user’s selection in the main visualization window.

The prototype application took the form of a Windows application written in Java and utilising the YFiles graph manipulation libraries.



awn - spiculation of margins
present
sterile spikelet - shape of tip not used
squared
ear - length(excluding awns)
medium
time of ear emergence (1st spk. vis. on 50% ears)
very early to early
rachis - length of first segment
medium
awn - length (compared to ear)
long (longer than ear)
seasonal type
spring type
lower leaves - hairiness of leaf sheaths
absent
flag leaf - anthocyanin colouration of auricles
present
grain - rachilla hair type
short
ear - glaucosity
medium
kernel - colour of aleurone layer
whitish ("white")
awns - anthocyanin colouration of tips
present
awns-intensity of anthocyanin colour. of awn tips
weak
ear - number of rows
two
grain - disposition of lodicules
clasping ("collar" type)
ear - shape
n/a
plant - length(stem,ears and awns)
medium to long
flag leaf - glaucosity of sheath
strong to very strong
collar type
platform to shallow cup
grain - anthocyanin colouration of lemma nerves
weak
grain - husk
present
grain - ventral furrow - presence of hairs
absent
grain - spiculation of inner lateral nerves
weak (1-2 per nerve)
ear - attitude (at least 21 days after ear emerg.)
horizontal to semi-recurved
plant - frequency of plants with recurved leaves
low

Figure 5-1 Initial interactive prototype

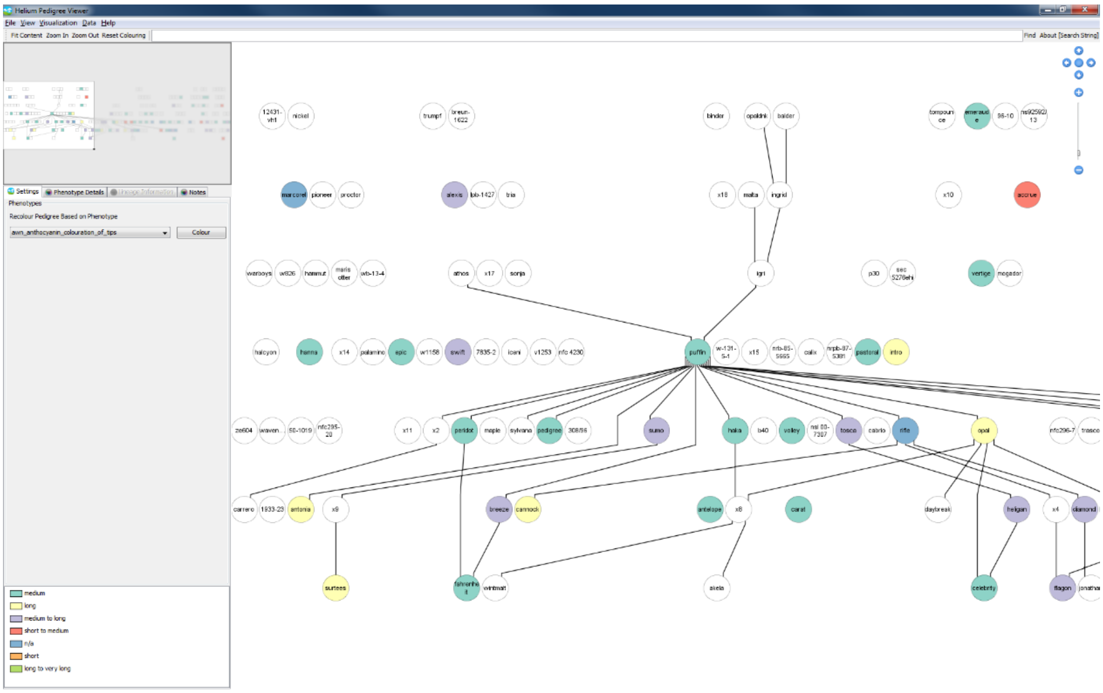
The inclusion of the option to resize nodes was included but was limited to node sizing based on the usage of specific plant lines in the pedigree. Both node sizing and the highlighting of node edges to show both predecessors (green nodes) and successors (purple nodes) are also shown. Orthogonal edge routing was used in the Helium prototype which can be clearly seen here also.

The Sugiyama based algorithm floats nodes with no incoming or outgoing edges towards the perimeter of the visualization/layout. This has a number of important advantages. Firstly it means that these nodes are easily seen as ones where limited information is held and acts as a pointer to look in more detail at the underlying data to fill in information and secondly it prevents the main visualization from becoming cluttered with nodes which may not add to the overall perception of the visualization.

The development of dominance was achieved by removing colour hue from unselected nodes, thus de-emphasising them.

Figure 5-2 and Figure 5-3 show the difference between ordinal and nominal colour coding pallets used in the prototype which were colour-coded in Helium using ColorBrewer2 palettes (Brewer, Hatchard, and Harrower 2003; Harrower and Brewer

2003). Hue was used to differentiate nominal data and saturation to distinguish ordinal data classes for phenotypes and genetic similarity metrics within the visualization (Ardi and Tan 2002).



Phenotypes

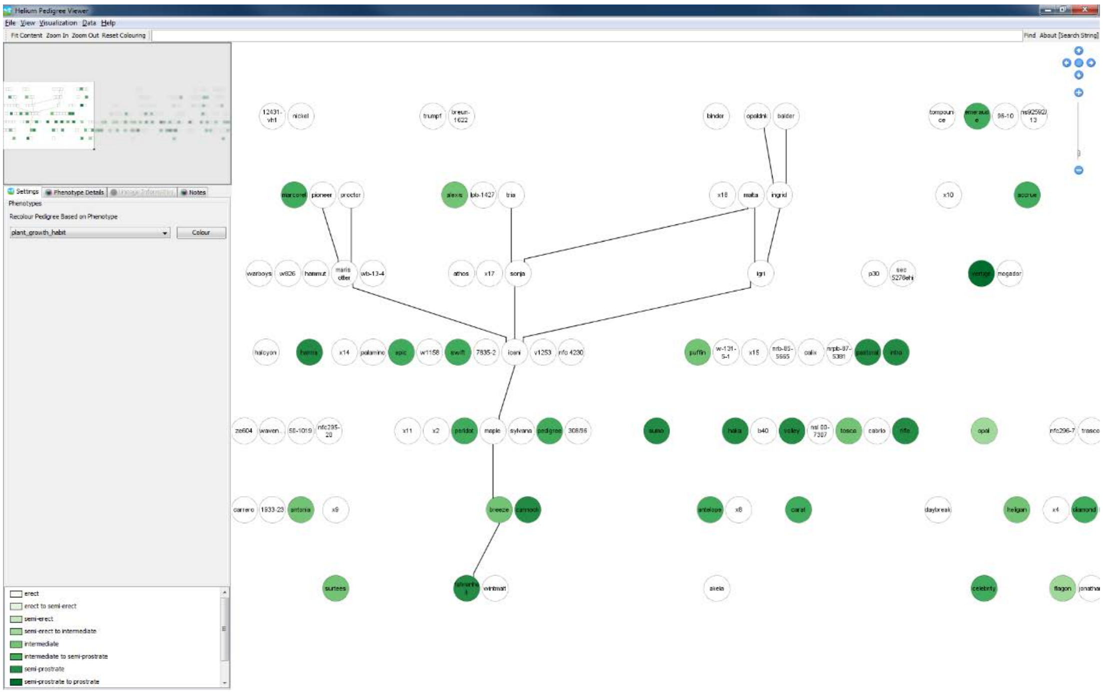
Recolour Pedigree Based on Phenotype

awn_anthocyanin_colouration_of_tips

Figure 5-2 Updated prototype with new edge layout and nominal DUS character coding

While originally it had been intended to show each phenotype as a different section on a node it was decided, through speaking to users during the initial evaluation that they would be interested in finding exact combinations and so it was decided to go with the single node colour to reduce clutter and keep the visualization clearer. There are however problems as the number of colours that have to be used can be around 20. Such a high number has been shown to be ineffectual at differentiating between classes (C. Ware 2004; Ardi and Tan 2002).

While users requested as much information as possible in the interface care was taken to only include necessary information and not turn Helium into a tool that presents so much unnecessary information to users it in itself becomes unusable or difficult to comprehend; situations where users are overloaded with information need to be avoided.



Recolour Pedigree Based on Phenotype: plant_growth_habit

Figure 5-3 Updated prototype with new edge layout and ordinal DUS character coding

5.4 Prototype feedback

Users when interviewed said that the overlaying of data onto the pedigree structure has in some ways more impact than showing the division of data in a bar chart or as a table. Having areas of colour in your face brings insight both into the location of clusters of similar data and visual impact of nodes changing from one colour to another.

While initial informal feedback from users was positive on this prototype there was one problem which users had and that was with the use of the orthogonal edge routing that was being used. While this helped to reduce edge clutter and overlapping within the visualization users thought that it would be more intuitive to be able to see all outgoing and incoming edges to a node. In this way they believed it would be a better representation of the complexity that exists in pedigrees and with the ability to selectively highlight edges only plant lines (nodes) and edges of interest would be visible. To this end the edge layout was changed (Figure 5-2 and Figure 5-3) to remove the orthogonal edge routing seen in Figure 5-1.

While edge bundling is often a useful technique in removing visual untidiness in complex graphs (Holten 2006; Newbery 1989), there are some inherent problems.

These include a reduction in specifics and/or detail about the connections (edges or in this case mating events) between nodes which was the exact problem that was seen in the initial informal feedback from users'. Additionally, work (McGee and Dingliana 2012) has shown that edge bundling does not aid user comprehension of complex networks.

During discussions with users it was also apparent that the ability to export plant line names would be a useful feature to allow scientists to make up lists for sending samples off for genotyping based on phenotypic or genotypic characteristics so the ability to allow users to export lists has been implemented. Users can select nodes then add them to an export list which can be saved to a text file.

The interactive prototype which was developed was now tested with domain experts.

5.5 Discussion

A prototype pedigree visualization system written in Java was developed in order to add additional features that feedback from the paper based prototype (Section 4.7.2). These features included the ability to interactively explore pedigree networks as well as search for information and overlay specific phenotypic data. The prototype system used linked views in a detail and overview based visualization in order to show higher level structure as well as a detailed representation of the pedigree.

The prototype maintained the layout of the paper based prototype using graph nodes to represent plant lines/varieties and graph edges to represent mating events. In this way a graph structure was constructed which was familiar to users of the paper based prototype..

Color Brewer colour palettes (Brewer and Harrower 2014) were used to encode ordinal and nominal phenotypic data classifications and tools added to allow users to select phenotypes and update the visualization to represent the classification of represented plant lines for the selected phenotypes.

User feedback for this first interactive prototype was positive, with exception of the use of orthogonal edge routing which was changed in subsequent prototypes.

Chapter 6 Initial user testing of the Helium prototype

6.1 Overview

This chapter details the process behind initial user testing on the Helium prototype that was detailed in Chapter 5. It details the methodology used for the user testing including the development of the questions that would be used in the testing questionnaire and then presents the results that were obtained from a detailed user testing carried out on 16 domain experts in order to identify potential problems with the prototype things that participants liked and things that participants did not like about the interface. Finally, it highlights some of the problems that test participants had with the Helium interface and identifies areas which could be focussed on to improve Helium going forward.

User testing is an important aspect of the development lifecycle of visualization systems (Sedlmair, Meyer, and Munzner 2012; Munzner 2009; Lam et al. 2011). Both Munzner and Lam lay out the requirements for testing, specifically relating to visualization studies in both contemplation and reflection of user studies.

6.2 User testing aims

The aim of this user testing was to establish if the abstract representations that had been implemented in the initial Helium interactive prototype were measurably useful to end users and were sufficiently intuitive to use that users, with minimal exposure to the visualization could perform complex pedigree operations with low error rates.

6.3 User testing methodology

A subjective evaluation was performed to establish user perception/acceptance and understanding of the visualization methods within Helium. This was to establish empirically if users were happy with representing data as graphs, moving away from the traditional family-tree type methods, and whether the use of graphs fits in with a user's perception of pedigree structure and function. Could the users perform basic pedigree operations such as accurately tracking back through generations and find information they require using the visualization? This was also done to ensure that users were able to interact well with the methods, which allow much greater data density and increased plant line density.

Test subjects were selected from a pool of plant geneticists, predominantly from The James Hutton Institute Cell and Molecular Sciences group. These users were not only the target audience for the visualization tool but also both extremely experienced and with international reputation.

The test datasets used in the evaluation were the main barley pedigree data along with DUS character data described in Section 2.6.3.2.

Ethical consent forms were obtained for testing which were read and signed by participants and the procedures and processes required to fulfil the ethical requirements of both Edinburgh Napier University and The James Hutton Institute (Appendix 3).

A pre-screening questionnaire, user tasks, and a follow up questionnaire centred on predefined tasks that users would be asked to perform was developed. The initial questions were to gain an overall impression of the length of experience the user has had in this field, and to classify their job title. There were two distinct groups of potential users: bioinformaticians/computational biologists and plant geneticists (experimental)/breeders (applied). User tasks were developed using the initial application requirements and were designed to force the users conducting the test to explore the experimental test datasets. The follow up questionnaire was clearly split into two sections; the first taking the form of attitude-scale questions on the user's opinion on the software and visualization in terms of both their use of it (assuming comparison to their current method of viewing these data types), and follow up subjective open-ended questions to get additional information that could be used to drive development of this software tool.

The testing data was obtained through a series of task-based objective questions, questionnaire and comment-based (open ended) feedback based on how intuitive users found the main features of the prototype to be (Table 6-1). Users were asked if this tool could be improved relating to general usage or new features. This is important as while initial user-requirements were gathered, when users actually started using this software it was fully expected of them to come up with new ideas on features or utility that would benefit their research.

The questions assume that a comparison is being made to other methods that test subjects are, or have been using to obtain the same information, and can be used to signify if the visualization and user interface brings significant improvements in visual representation and understanding of pedigree structure. Throughout the study, notes were taken and screen and audio capture was used to further examine a user's interaction with the interface and to aid in recount of the tests.

All testing was carried out on a Dell E6430 series laptop with an Intel i5-3360M processor and 8GB of RAM connected to a 1900 x 1200 (16:10 aspect ratio) Dell Ultrasharp monitor. Helium was running from a solid state drive (SSD) for increased performance and Germinate was from a local MySQL installation to reduce potential problems with network issues. User input was from a standard Microsoft keyboard and mouse.

Each test was scheduled to take around 45 minutes;

Questionnaire Section	Duration	Data Gathered
Pre-questionnaire	5 minutes	General profiling data
Familiarisation	5 minutes	NA
Test	25 minutes	Correctness and completion rate
Post-test questionnaire	10 minutes	Attitude scores and objective feedback

Table 6-1 User testing structure and data gathering

After completion of the main interaction study, users completed an attitude scale where they indicated their preference on a 5 point scale between ‘Very Difficult’ (1) to ‘Very Easy’ (5) relating to a number of statements about their use of this software.

The questionnaire asked users to detail features or concepts that they found to be confusing, those they found to be clear, and features that they feel would add value to their research. Finally users were asked to provide general comments about their use of the software; this would be used to allow tweaks and fine-tuning of the Helium interface to aid users with their research.

6.4 Test results

The following sections describe the results from the first stage Helium subjective evaluation.

6.4.1 *General background profiling*

The 16 expert users that undertook this study break down as follows; 5 bioinformaticians, 10 plant geneticists and breeders and 1 statistician. Out of the users, 94% were educated to PhD/MSc level and the average length of time working in their areas was 17 years. The minimum experience was 1 year, maximum 36 years giving a median length of experience of 13.5 years.

While all users were familiar with pedigree data, 69% used it on a day-to-day basis as part of their research and 38% regularly used alternative tools such as Microsoft Excel.

It should be noted that through verbal feedback it was established that the researchers who were using pedigree data were using paper records and spreadsheets to curate and maintain pedigree data used in their work and not a specific pedigree tool.

6.4.2 *Main user interaction study*

There were eight questions that users were asked to answer in using the pedigree interface. The questions were assigned an overall category and can be seen in Table 6-2 and Figure 6-1 and detailed here.

Unexplained Concepts (Category 1): the user would be required to speculate on what a specific type of node represented on the visualization. This had not been explained to them and was a representation where a parent was actually a cross between two plant lines but not assigned a varietal name.

Simple Grandparent Tracking (Category 2): the user was required to locate a plant line within the visualization using the search features then track back up the pedigree to locate the parents, and subsequently grandparents, of the plant line ‘Ayr’. This would involve them tracing back to identify the 2 parents and 4 grandparents by following lineage (graph edges).

Identifying Children (Category 3): the user was required to identify the children for a specific plant line by following lineage. The plant line was chosen specifically because while it had 3 children they were located on different areas of the pedigree representation which would force users to scroll around and follow edges. The plant line used was ‘Sebastian’.

Complex Grandparent Tracking (Category 4): this question asked the user to find a plant line using the search functions of the interface then go to that plant line’s children (of which there was only one). Then from the child track back up the pedigree to find the grandparents. This question was designed to make the user think about the question that was being asked and to force them to move both down through successors and up the pedigree through ancestors.

Phenotype Classes (Category 5): this question was asked to make the test users look at the colour scheme which had been implemented to encode DUS characters for the testing. The colour palette was the ColorBrewer 9 class quantitative BuGn palette.

Great Grandparent Tracking (Category 6): while similar to question classifications 2,3 and 4 in that it required the user undertaking the testing to track through the pedigree this added an additional generation level and therefore a large increase in the complexity of tracing the lineage and keeping track of what they had just done. An additional complexity was that not all the great grandparents were known with only three grandparents having their parents available.

Finding Additional Information (Category 7): the question used here was to ask the user to look for some additional background information on a specific plant line (‘Hart’) and provide the plant lines breeders code and AFP (Application for Protection) number. This information was held in the connected Germinate database so users needed to select the desired plant line by searching then use the Germinate data panel within the interactive interface to locate the desired information.

Colour Coding Perception (Category 8): this question was to try and asses if the test users were able to effectively use the ColorBrewer BuGN palette by asking them to find a plant line (in this case it was 2 plant lines ‘Scarlett’ and ‘Vegas’) in the interface, select a phenotype (‘Time of Ear Emergence’) then report the phenotypic values for the plant line and phenotype combination.

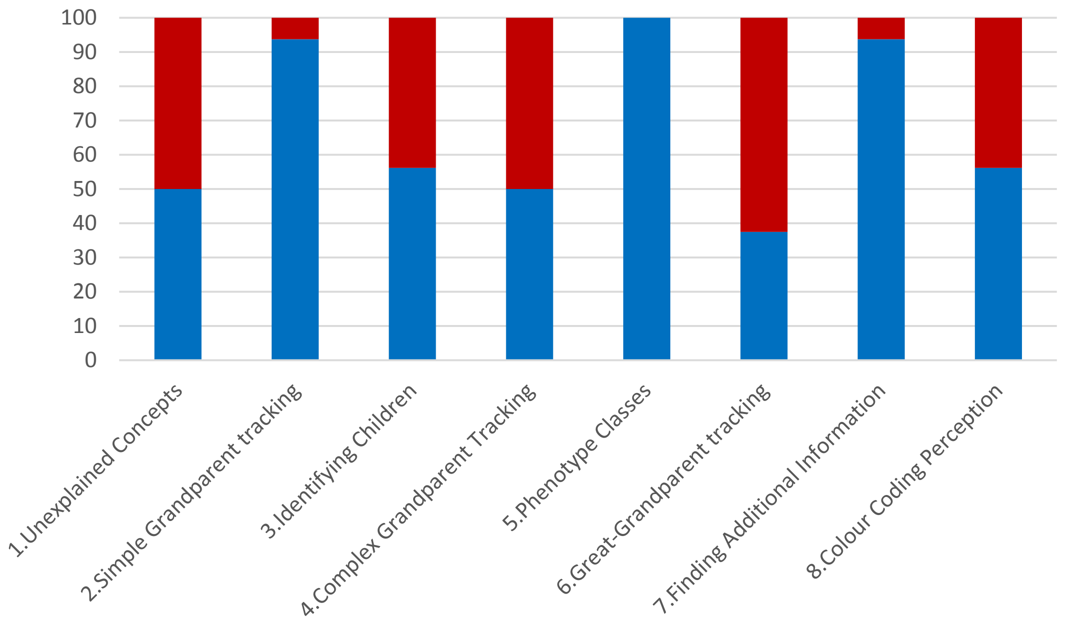


Figure 6-1 Interaction study correct and incorrect responses

Question Classification	Correct (%)	Incorrect (%)
1.Unexplained Concepts	50	50
2.Simple Grandparent tracking	93.75	6.25
3.Identifying Children	56.25	43.75
4.Complex Grandparent Tracking	50	50
5.Phenotype Classes	100	0
6.Great-Grandparent tracking	37.5	62.5
7.Finding Additional Information	93.75	6.25
8.Colour Coding Perception	56.25	43.75

Table 6-2 Interaction study correct answers

6.4.2.1 Post-study questionnaires (attitudinal and open ended)

After carrying out the main interaction study users were asked to fill in a series of questions that asked them to compare Helium to pedigree tools, or methods of handling pedigree data that they are familiar with using, and to get feedback on what they found easy and difficult to understand or perform with Helium. These results were Likert scale (1-5 with 5 being very easy) and are presented in Table 6-3.

Question / Test User	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean
1. Finding Parents	5	4	4	4	4	4	5	5	5	5	3	4	4	5	5	5	4.44
2. Phenotype Classes	3	3	5	4	4	3	5	5	4	4	2	4	3	3	4	4	3.75
3. Tracing Lineage	5	4	4	4	3	4	5	4	5	4	3	4	4	4	4	4	4.06
4. Understanding Data	5	4	4	4	4	3	5	4	4	4	4	5	4	4	4	4	4.13
5. Colour Coding	3	3	2	3	3	2	4	2	4	4	2	3	3	3	4	4	3.06
6. Children	5	4	5	4	4	4	5	4	5	5	4	5	5	4	5	5	4.56
7. Background Information	5	3	3	3	4	3	5	4	5	5	5	5	4	4	5	5	4.25
8. Clarity of Relationships	5	4	3	4	3	3	4	5	5	5	3	4	4	4	5	5	4.13
9. Finding Lines	5	5	5	5	4	4	5	5	5	5	5	5	5	3	5	5	4.75
10. Maintaining Bearings	5	4	2	4	4	2	5	4	4	4	4	3	4	4	4	4	3.81
11. Navigation	5	5	4	4	4	3	5	3	5	5	5	5	4	4	5	5	4.44
12. Ease of use	5	4	4	4	4	3	5	4	5	4	4	4	4	4	4	4	4.13

Table 6-3 Main user interaction study Likert responses colour coded (1-5 very difficult to very easy)

6.4.3 *General testing feedback*

Users were asked to comment on anything they found to be particularly confusing or easy to understand while using the interface and suggest improvements or additional functionality that would make the visualization tool more useful for their work.

6.4.3.1 Features users liked or found easy to understand

Features users liked included the layout which they thought was easy to understand and made scientific sense to them. They commented that it was similar to that of a more traditional family tree so they could understand quickly the concepts it was trying to convey. They also stated that it was easy to follow the edges and trace lineage and that searching for plant lines was simple. Finally users commented that it was incredibly useful bringing together multiple data sources into a single tool.

6.4.3.2 Features users disliked or found confusing

One of the main problems that users found was their ability to differentiate between phenotype classifications when, with ordinal data types there were more than a few data categories. A number of users thought that the long edges, which sometimes exist between plant lines, were disorienting. Users wanted to be able to jump quickly to a plant line when performing a search and automatically select that plant line so linked views would all be updated. Finally, users wanted clearer explanations of the phenotypes and associated classes available to them from within the visualization interface.

6.5 Test results discussion

An interesting outcome of the development of Helium is trying to quantify if this tool actually make a user's decision making better and does the software influence users into making more informed decisions about their data. One of the outcomes from the testing was to seek assurance that the decisions that had been made around the design of the tool were actually good foundations that users can build knowledge on and to that end Helium has made an impact. While standard approaches to the visualization tool were used, they have been developed and applied directly to a specific domain in which they have not been applied before, and the application tailored appropriately.

The testing of real experimental data and the integration of both pedigree and phenotypic data is something that users have had to previously conduct using alternative tools (mainly Excel) where pedigrees were represented as text strings. What Helium has done is allow the test participants to start to look critically for errors and patters in the test datasets which would have been much more difficult before where there was not a structure to the data. They have not been able to have a pedigree structure and quickly overlay different phenotypes in real-time with an interactive tool.

Test users liked the speed at which they could find data, the ease of tracing lineage through complex graphs (although the testing has shown that there were issues with this) and the intuitive layout of the visualization and supporting application. Tracing lineage when using the usual text files and Microsoft Excel approach is incredibly difficult and time consuming. What the testing did do was to highlight some issues, mainly around the use of colour gradients used in ordinal lists, which are ineffective and difficult for users' to distinguish when there are more than eight phenotype classes.

The user testing uncovered some interesting problems with this visualization. For example, the category 'Identifying Children' from

Table 6-2 asked participants to identify the progeny of a specific barley variety. In 44% of completed questionnaires this answer was incorrectly given. However, when examining 'Tracing Lineage' from Table 6-3 which related to this question, users thought that it was easy to trace lineage by following graph edges. Test users were continually missing the same progeny (one of three) of the plant line; the one whose complete edge was not immediately visible, and disappeared off the right-hand side of their computer display – position is clearly salient. When talking to a selection of users after the test had been carried out and asking them to perform the same question they did so without error (obviously suspicious to the reasons behind the request). Results from the Likert questions are detailed in Figure 6-2 and Table 6-4.

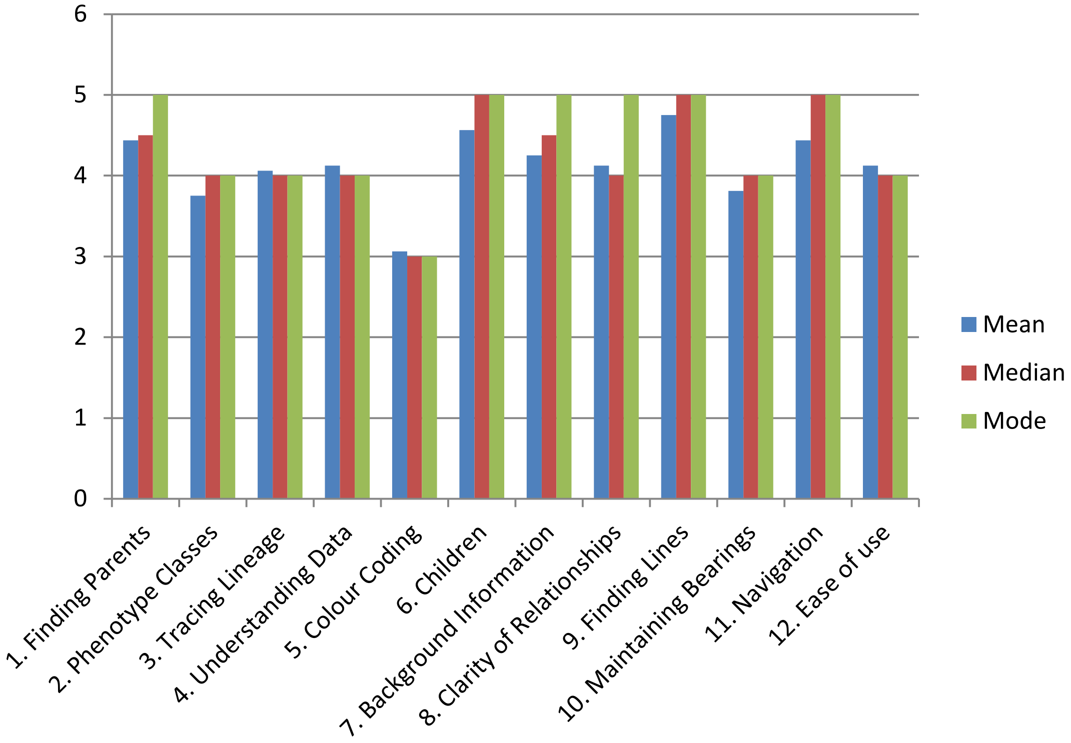


Figure 6-2 Main user interaction study results

Question	Mean	Median	Mode	SD
1. Finding Parents	4.43	4.5	5	0.61
2. Phenotype Classes	3.75	4	4	0.83
3. Tracing Lineage	4.06	4	4	0.56
4. Understanding Data	4.12	4	4	0.48
5. Colour Coding	3.06	3	3	0.75
6. Children	4.56	5	5	0.5
7. Background Information	4.25	4.5	5	0.83
8. Clarity of Relationships	4.12	4	5	0.78
9. Finding Lines	4.75	5	5	0.56
10. Maintaining Bearings	3.81	4	4	0.81
11. Navigation	4.43	5	5	0.7
12. Ease of use	4.12	4	4	0.48

Table 6-4 Main user interaction study results

The results from the user interaction study show that while there are questions that users had problems with, in particular colour coding the range of responses to the Likert questions shows that there was general problems as and not skewed by a number of particularly poor feedback. This is highlighted by the average standard deviation across all 12 questions of $\sigma = 0.81$. Figure 6-3 shows the median values of each of the 12 questions along with standard error assuming a 95% confidence limit. The median is used in this instance as the data is not a normal distribution with heavy skewing of data towards the more favourable responses with a low number of lower scores. The use of average/mean scores (Table 6-4) may give an inaccurate representation of data due to the score distribution. Again the data shows that while there was a lower mean response the error margins show that this was consistent across all responses and not skewed by outliers. The relatively low margins of error shown in Table 6-5. Results in Table 6-5 also indicates that overall, the responses from test users were consistent and showed minimal variation between favourable and poor Likert scores.

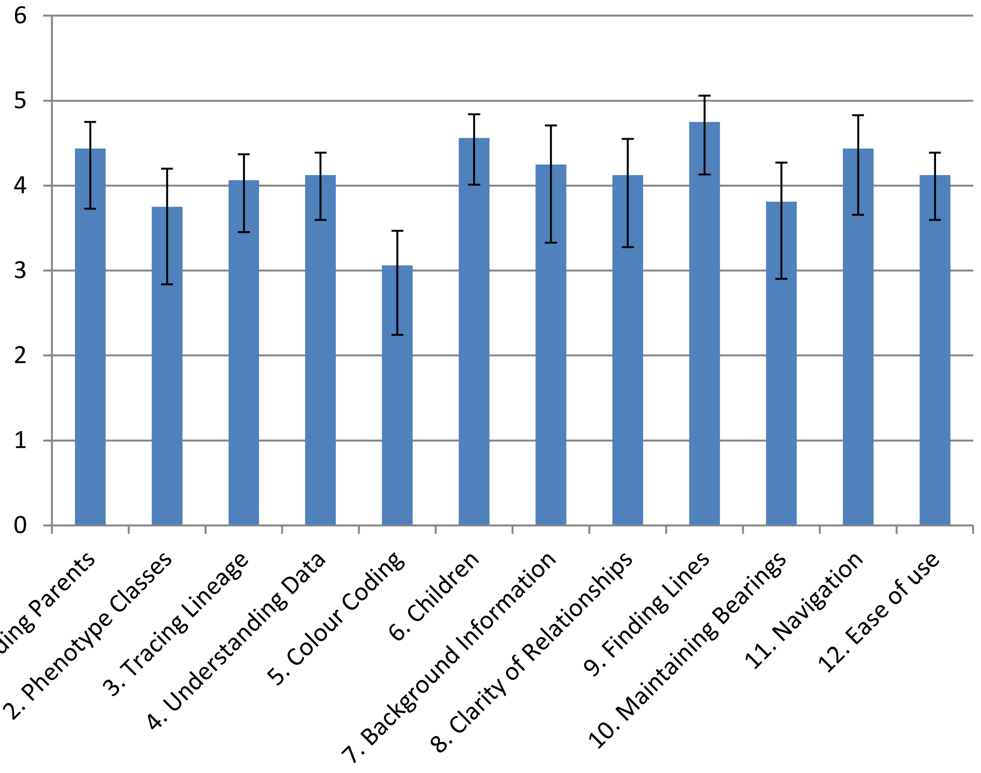


Figure 6-3 Main user interaction study medians including 95% confidence limit

Question	Upper Limit	Lower Limit	Margin of Error (0.95)
1. Finding Parents	4.75	4.04	0.35
2. Phenotype Classes	4.2	3.29	0.46
3. Tracing Lineage	4.37	3.76	0.31
4. Understanding Data	4.39	3.86	0.27
5. Colour Coding	3.47	2.65	0.41
6. Children	4.84	4.29	0.27
7. Background Information	4.71	3.79	0.46
8. Clarity of Relationships	4.55	3.7	0.43
9. Finding Lines	5.06	4.44	0.31
10. Maintaining Bearings	4.27	3.36	0.44
11. Navigation	4.83	4.05	0.39
12. Ease of use	4.39	3.86	0.27

Table 6-5 Main interaction questions confidence limits

The lower scores from the main interaction study (2. Phenotype Classes, 5. Colour Coding and 10. Maintaining Bearings) along with verbal feedback from users was used to direct and prioritise the iterative improvement of Helium's features in order to increase user understanding and comprehension of the system. This will be detailed in Chapter 7.

Chapter 7 Improvement to the Helium interface

7.1 Overview

The first user testing of Helium (Chapter 6) showed that there were a number of areas in which improvements could be made to increase both the functionality and usability of the Helium software. These were addressed then a follow up evaluation performed to verify the effectiveness of improvements made. This chapter discusses in detail improvements that were made to the Helium interface in order to try and help aid user comprehension in areas that were identified to be problematic and the process that was carried out to get Helium into a condition where a second round of user evaluation could be carried out in order to test the effectiveness of these described improvements.

7.2 Suggested improvements to the Helium interface based on initial user testing

The results in Chapter 6 highlighted areas in which improvements could be made to the Helium interface. These improvements were targeted using the results from the user testing which showed that users were having difficulty with some of the concepts that were used in the visualization. In order to try and improve understanding these problems were prioritised into three main categories; tracing lineage, unexplained concepts and colour coding. The results, feedback and subsequent analysis of the screen captures where appropriate were used to identify the areas in which improvements were made to the visualization interface to minimise problems.

7.2.1 *Tracing lineage problems*

Questions 2,3,4 and 6 in Part 2 of the user testing questionnaire (Appendix 4) asked users to find successors or predecessors of defined plant lines, each question had varying complexity ranging from immediate children and parents in questions 2 and 3 to more complex tracing of lineage in questions 4 and 6 to find grandparents and great-grandparents. The results in table 1 showed that there was clearly a problem with users obtaining the correct data from the prototype Helium visualization tool. Identifying children showed that only 56.35% of responses were correct, grandparents 50% correct and great grandparents 37.5% correct. These results show a clear reduction in correct responses as complexity increases.

The reasons for these problems are twofold. Firstly it was clear that there were cases where a user's stated ability did not match their understanding of the underlying biological concepts, for example having to explain what great-grandparents are but this accounted only for one user. The main problem was immediately obvious while watching users carry out this user testing. Question 3 asked to identify the children of plant line 'Sebastian'. This could be easily carried out with Helium by searching for the plant line Sebastian then counting the number of children it had (Figure 7-1).

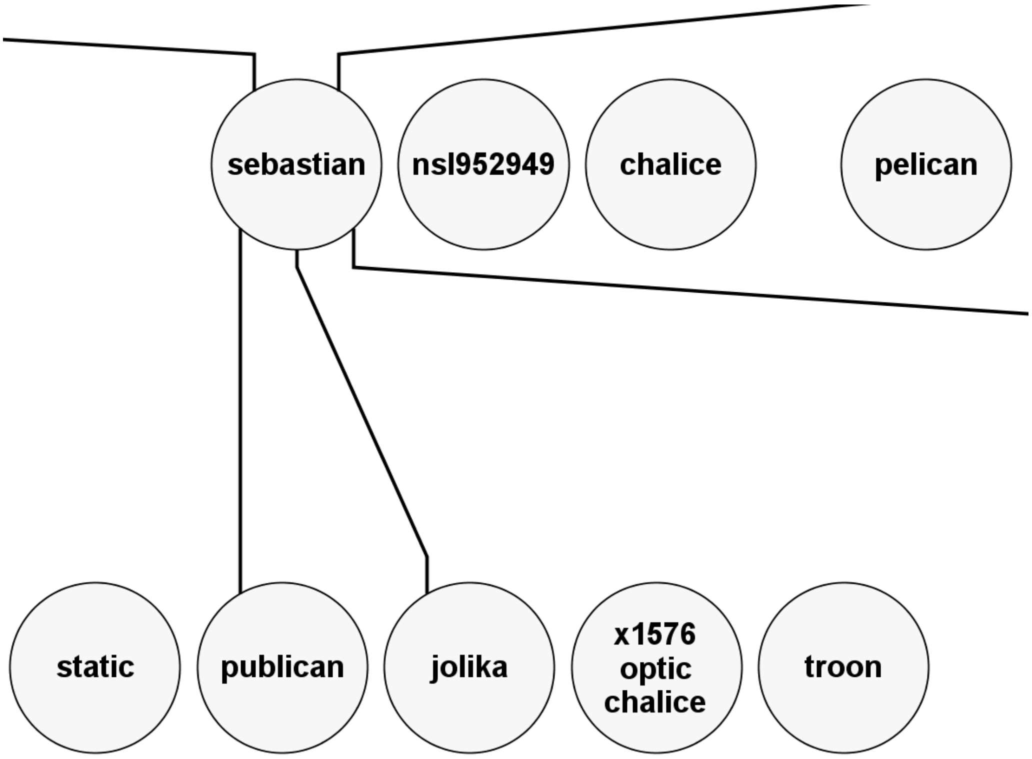


Figure 7-1 Barley variety Sebastian showing inwards and outgoing edges

It can be clearly seen that the children of Sebastian include Publican and Jolika plus one other (Quench) which cannot be seen in the above figure but represented by an outgoing edge from Sebastian. Panning the visualization would have allowed the user to identify this plant line (Figure 7-2). It is clear that there are salience issues with regards to the visual variables in this representation that were causing users issues.

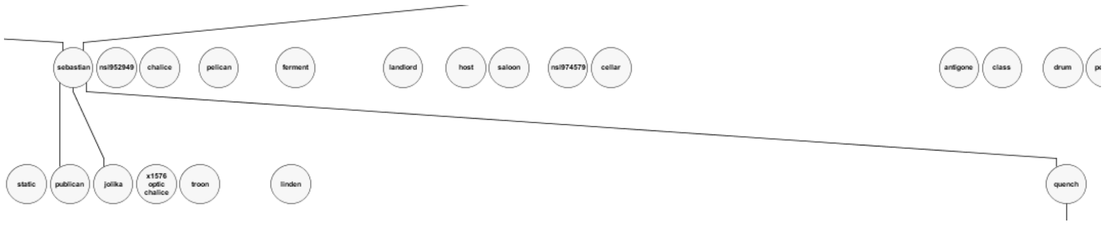


Figure 7-2 Sebastian showing 3 progeny (Publican, Jolika and Quench)

The position of the plant line Quench is a property of the positions in the layout of the parents and children of Quench. This clearly gave users problems with 43.75% of users incorrectly answering this question. Similar problems existed while finding grandparent and great-grandparents, 50% and 62.5% answering these questions incorrectly. Again some of this could be attributed to confusion on what a great-grandparent actually was but it was also clear that users were having problems tracing edges to nodes where the node was located off the current viewport which relates back to the salience issues on the positioning of nodes mentioned earlier. It is not the case that all children are positioned next to each other.

7.2.2 *Unexplained concepts problems*

The definition of unexplained concepts in this work is plant lines for which the parents are known and who have been used as parents but it is not known what they were called, and no record of them is available. This is common in the test datasets. In order to differentiate these plant lines from other known varieties they were shown as nodes with a name prefixed with ‘x’. An example which can be seen in Figure 7-1 would be ‘x1576 optic chalice’. Domain knowledge would allow the user to quickly identify that the plant lines Optic and Chalice were commercial varieties and experienced users quickly recognised the importance of these, this was not the case for inexperienced users.

It was clear however that this would need to be improved in future development to try and help users understanding.

7.2.3 *Colour coding problems*

The colour palettes used in the prototype version of Helium are based on quantitative and qualitative ColorBrewer palettes (Brewer, Hatchard, and Harrower 2003). When examining the ordinal data category ‘Time of Ear Emergence’ (question 8 in the testing questionnaire) there were 9 categories ranging from ‘very early’ to ‘very late’. These were coloured using the 9 class BuGn colour palette. When asked to distinguish between two colour classes which were consecutive only 56.25% of users gave the correct response. The main problem identified from feedback was that it was difficult to differentiate between consecutive colours in the ColorBrewer quantitative palette.

7.3 **Final iteration of the Helium pedigree visualization platform**

Feedback from the user evaluation allowed issues that users had with the initial prototype to be addressed. This would allow a more refined visualization application to be developed using the initial prototype as a foundation and tweaks and improvements made as required.

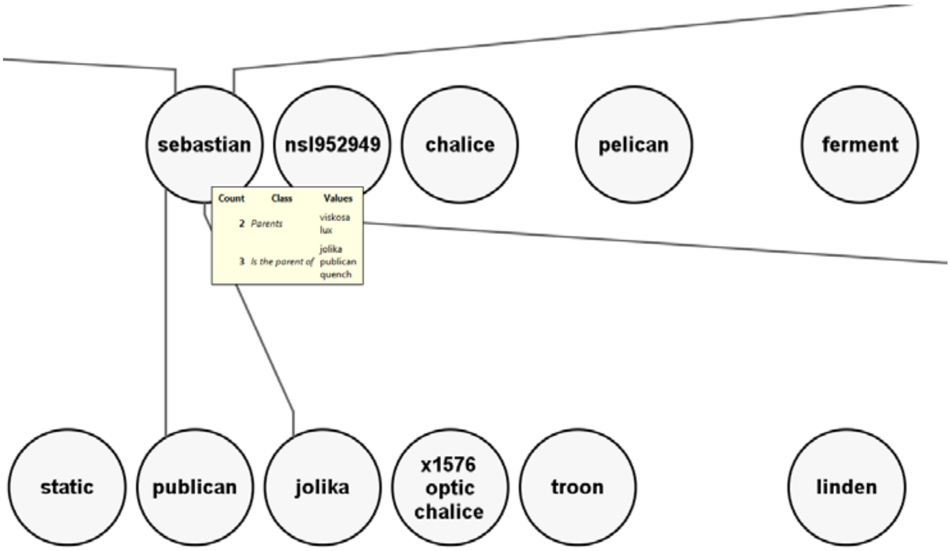
Problems relating to the three categories of problems identified in section 7.1 as well as user feedback obtained from the post study questionnaire could begin to be addressed to improve understanding of the pedigree visualizations. Any subsequent development would need to address these points if it was going to offer a usable and effective tool for users.

7.3.1 *Addressing user issues*

The main feedback gained from the initial prototype was that it was difficult to track lineage with overlapping edges. As previously stated users were having trouble identifying plant lines where children or parents spanned more than just the current viewport seen by the user. In order to address this problem, new features were added to the interface to increase awareness in situations such as these which are detailed below.

7.3.1.1 Addition of visual cues

The use of visual cues was implemented to aid comprehension. When a user hovered over a node a list of ancestors and successors was displayed showing both a numerical count, and the names of immediate parents and children (Figure 7-3). The main aim was to visually show the number of related nodes and thus try and reduce errors like those seen in user testing.



Count	Class	Values
2	*Parents*	viskosa lux
3	*Is the parent of*	jolika publican quench

Figure 7-3 Visual cues

7.3.2 *Modifications to Helium user interface subsequent to user evaluation*

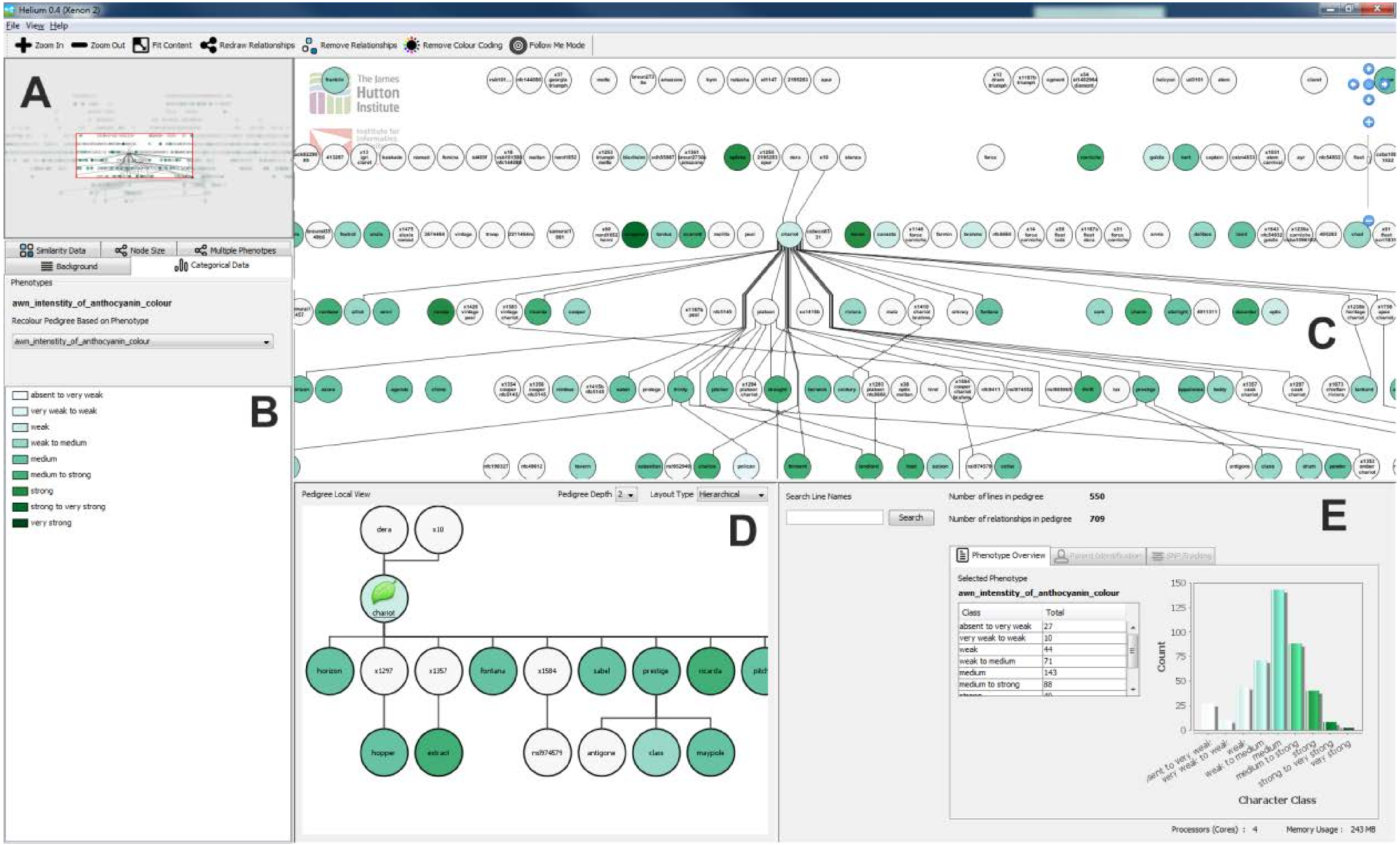
The Helium user interface was re-designed to show 4 main areas (Figure 7-4). These are described below.

7.3.2.1 Overview and data selection panel

This panel (Figure 7-4A and Figure 7-5A) also includes selection mechanisms for choosing ordinal and nominal categorical phenotypic classes as well as tools for visualizing genetic similarity data (Figure 7-4B and Figure 7-5A). Users can use the overview to navigate to a particular region within the main visualization window if required.

Interactive sliders allow users, in the case of similarity data, to set a percentage similarity value and in real-time highlight plant lines which match the search criteria (Figure 7-5A). In this way it is possible to see plant lines which should not be closely related appearing on the peripheries of the visualization as the slider is moved, which can indicate problems with pedigree definition or genotyping. Histograms have also been included, where appropriate, to show data distribution which can be an aid in the identification of problem markers. While the number of markers that have this problem is limited, it is nonetheless important to address.

Other features included in this panel are the ability to select more than one phenotype then recolour nodes based on the merged phenotype classes.



Phenotypes

awn_intensity_of_anthocyanin_colour

Recolour Pedigree Based on Phenotype

awn_intensity_of_anthocyanin_colour

□ absent to very weak
□ very weak to weak
□ weak
□ weak to medium
□ medium
□ medium to strong
■ strong
■ strong to very strong
■ very strong

Number of lines in pedigree 550

Number of relationships in pedigree 709

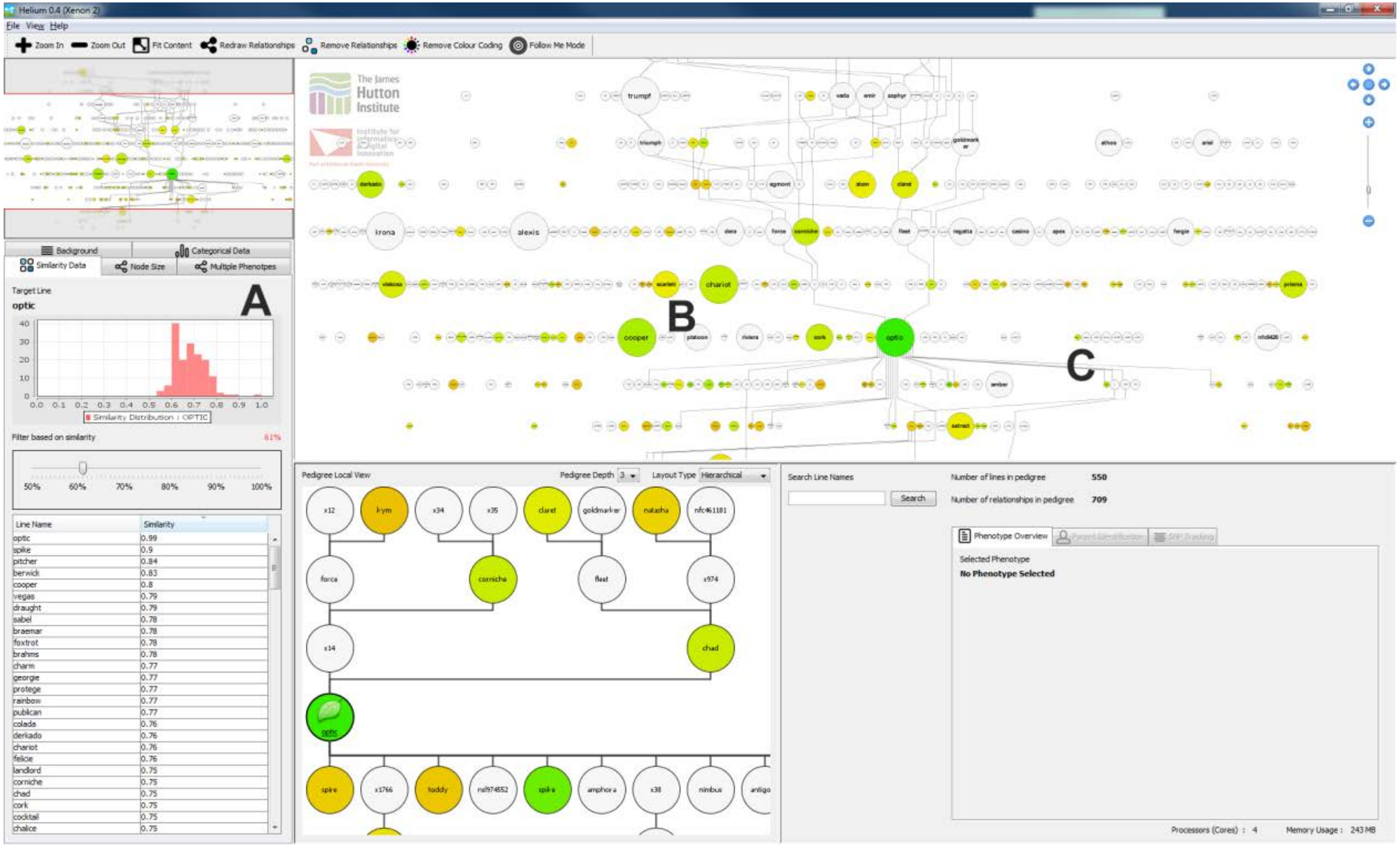
Selected Phenotype

awn_intensity_of_anthocyanin_colour

Class	Total
absent to very weak	27
very weak to weak	10
weak	44
weak to medium	71
medium	143
medium to strong	88
strong	40

Processors (Cores) : 4 Memory Usage : 243 MB

Figure 7-4 Helium modified interface after user testing



Target Line
optic

Filter based on similarity 61%

Line Name	Similarity
optic	0.99
spike	0.9
pitcher	0.84
berwick	0.83
cooper	0.8
vegas	0.79
draught	0.79
sabel	0.78
braemar	0.78
foxtrot	0.78
brahms	0.78
charm	0.77
georgie	0.77
protege	0.77
rainbow	0.77
publican	0.77
colada	0.76
derkado	0.76
chariot	0.76
felicie	0.76
landlord	0.75
corniche	0.75
chad	0.75
cork	0.75
cocktail	0.75
chalice	0.75

Pedigree Depth 3

Layout Type Hierarchical

Search Line Names

Number of lines in pedigree **550**

Number of relationships in pedigree **709**

Phenotype Overview | Parent Identification | SNP Tracking

Selected Phenotype

No Phenotype Selected

Processors (Cores) : 4 Memory Usage : 243 MB

Figure 7-5 Helium showing genetic similarity data

7.3.2.2 Main visualization panel

The main visualization window (Figure 7-4C and Figure 7-5B) was modified in a number of ways from the prototype. The move away from bundled orthogonal edge routing (Figure 7-5C) made the tracing of lineage easier. Slightly modified colour palettes were used to account for the situation where there are more than eight categorical classes. The new colour palette helps with the problem where adjacent classes were too similar in colour for users to accurately distinguish. In Table 6-2 the incorrect responses to 'Identifying Children' were high at 43.75%. In order to address this visual prompts when hovering over a node were added which display the number of ingoing and outgoing edges from a node and the names of the plant line's progeny (Figure 7-3). This makes the number of progeny immediately obvious, which will help prevent some of the problems seen in testing. When a user selects a node the edges connecting nodes of interest are made more prominent by both removing edges, which are not associated with the selected node, its ancestors, or successor, and by darkening the edges which are left.

Hovering over a graph edge will show the names of the two nodes that it connects, in this way with long edges, while using the main visualization window; it is easier to track their origin and destination.

7.3.2.3 Local view panel

Testing also showed that while users reported they found it easy to identify lineage there were some issues. What was termed a local-view was implemented which showed only plant lines that were directly related to a selected node. This is effectively a filtered view at the data level showing only related individuals. The reduction in complexity when unnecessary nodes were removed from the display had a number of benefits. These included the reduction in visual clutter and the reduction in space required to visualize the layout. This had the effect of bringing plant lines which were far apart in the main Helium visualization window close together meaning users would not have to 'chase edges' to find related nodes.

The local view would be shown when a user selects a node in the visualization. This view was implemented below the main visualization window. The local view can be panned and zoomed in the same way as the main visualization window. Within the local view the user has control of how many generations, forwards and backwards, they want to go. This addresses the problems highlighted in Table 6-2 where there were 50% and 62.5% of users incorrectly answering the 'Complex Grandparent Tracking' and 'Great-Grandparent Tracking' questions respectively. With appropriate selection of generation level, grandparents, or indeed any other generation, are now immediately obvious in the simplified pedigree. Additionally, the ability to layout the graph using a number of edge routing algorithms was added. Any changes made to the main pedigree visualization are propagated to the local view. While the local view includes another copy of a portion of the main visualization, it will increase the accuracy of tracing lineage when unnecessary plant lines are removed and edges between nodes shortened, thus addressing the problems highlighted in testing and reducing the need to 'chase edges'.

7.3.2.4 Detail panel

The details panel (Figure 7-4E) showed either information on the currently selected phenotype(s) or information from Germinate about the specific plant line. This example (Figure 7-4E) shows the distribution of the DUS character 'Anthocyanin Colour'. The histogram has been coloured in the same way as the phenotype classes in the main visualization window and was clickable to allow users to select a phenotype class and highlight these plant lines in the main visualization by means of a thick border.

The details panel also houses a search functionality which allows searching for plant lines with usual search features such as wild-card matching and an option which has been coined the 'follow me' mode which jumps to a search hit, selects it and subsequently updates the detail panel and main visualization window.

Finally, a user history panel has been included which records the plant lines and phenotypes that have been selected over a session so that if required, users can go back and see what they had been doing previously. This is important as with large quantities of data it is easy for users to forget what they have been doing over time.

7.4 **Discussion**

The Sugiyama style layouts offered by the two graph layout tools that were used (Dot and yFiles) were in the most part effective at laying out the pedigree structures that were available but there was some criticism about the positioning of edges and nodes on diagrams. Some of this criticism was users not comprehending how complex these pedigrees actually are and how it would be impossible to always layout large pedigrees with nodes and edges close together with few instances of edge crossing. This problem was partially addressed by the inclusion of the local view into the Helium tool.

It is also clear that while Dot and yFiles are useful tools when laying out up to a few thousand plant lines they do begin to become quite slow when dealing with any more than this. The problems with Dot are not quite as bad as it produces a static layout but in the case of yFiles with upwards of 3,000 plant lines things become noticeably less responsive. This is going to be a problem with whatever graph layout tools are used and some of the larger wheat breeding programmes who indicated they wanted to be able to handle upwards of 100,000 entries will need to work on how their pedigree structures can be broken down into logical units, or live with a less responsive system. This is not a problem which is unique to this work as it’s common for people to indicate they require to be able to visualize all their data when in fact basic pre-processing or dividing of data into more logical divisions is more appropriate and makes more biological and logical sense.

Other problems that Helium has is being able to use long plant line identifiers used by some breeding programmes. While in Europe most breeding programmes assign a short identification code to identify germplasm then have additional data that can be referenced back to this in the case of North American wheat and maize breeding for example pedigree data is often encoded into the plant line names which leads to identifiers with well in excess of 50 characters. Discussions with some users has shown that they want to be able to see this information within Helium nodes, something which is possible but not desirable due to the amount of space the visualization would occupy. However, additional work needs to be done with these sorts of data to try and form a set of guidelines for naming germplasm and more concrete guidance on what is a name and what is associated data that can go along with it. Were these names shortened they can be easily displayed in Helium and then give users’ access to the additional data types.

Chapter 8 User evaluation of updated Helium interface

8.1 Overview

This chapter describes the second round of user testing which was used to verify if the problems that were identified in the first round of user testing (Chapter 6) and addressed in Chapter 7 were effective. It describes the process behind the user testing which was similar to that carried out in Chapter 6 but also included the use of two industry standard usability questionnaires. The testing was carried out with 28 users. It concludes with a discussion about the results of statistical analysis on the test results and suggests features that could be improved upon in future Helium development.

8.2 Additional usability testing procedures

There was considerable enthusiasm, especially in this second round of user testing from staff members who had seen the Helium prototype and wanted to undergo the user testing. This was something that was never expected to happen where people were volunteering without being asked. This is one of the positives that was taken from the entire testing process.

The test users were asked to perform identical tasks to the first round of user testing (Chapter 6) so that a comparison to the previous testing with this round of testing could be made. This ensured that the changes made to the user interface subsequent to the first round of testing could be evaluated. Additionally standard questionaires used to evaluate software acceptability and usability were presented during the testing.

There were additional requirements in relation to carrying out tests involving human subjects which required application to the Human Ethics Committee at the James Hutton Institute before this round of testing could be carried out (Appendix 5). The test document can be seen in Appendix 6. These requirements came into affect after the first round of user testing.

User testing allows the quantification of results by presenting users with a standard series of questions which are selected to direct them into using the software in more detail than they would if they were presented with the software and asked to give unstructured feedback based on their opinions. Additionally, in order to make generalised comments on the usefulness of software, the use of a group of target users performing the same tasks gives greater weight to any argument on the efficiency or effectiveness of software over an uncontrolled harvesting of feedback from a random group of users.

The use of standard tests provides a framework for the logical reporting and in some cases comparison of results from user test experiments. They allow a value to be assigned to the software which may be used in comparisons with other applications which have been through the same testing procedures. The use of standard testing procedures also allows for the reproducibility of testing carried out and have been shown to be more reliable than ad-hoc usability studies (Hornbæk 2006). Finally, the use of standardised questionnaires may allow the comparison of results from different studies as a benchmark to which software may be compared. In essence, standard testing provides researchers with reliability, validity and quality across testing regimes.

Examples of commonly used standard frameworks for usability testing include the System Usability Scale or SUS (Brooke et al. 1996) which gives an indication or users perceived efficiency and learnability (although the learnability aspect was only recently described (Jeff Sauro and Lewis 2012)) of software (Adrion, Branstad, and Cherniavsky 1982; Brooke et al. 1996) and the Post Study Usability Questionnaire and PSSUQ (Gould and Lewis 1985; J. R. Lewis 1992; J. R. Lewis 1995; J. Lewis 2002) which indicates perceived satisfaction based on averaging subscales comprising 1. Information Quality (InfoQual), 2. System Quality (SysQual) and 3. Interface Quality (IntQual). The average is classified as ‘Overall Satisfaction’. Both questionnaire systems give commonly used and highly structured results from software testing. While other standard questionnaires do exist such as the Software Usability Measurement Inventory or SUMI (Kirakowski and Corbett 1993), these are often prohibitively expensive for small scale projects and research work.

It should also be noted that systems such as the PSSUQ are commonly known to be susceptible to acquiesce bias (Sica 2006) categorised by research test subjects being more willing to agree than disagree with a statement. Factors like these may also be amplified in research studies such as this whereby users were, by the most part, known before testing.

8.3 User testing

Each test was scheduled to take an hour and comprised of 5 minutes carrying out the pre-questionaire, 10 minutes familiarisation using the tool, 20 minutes of doing the test questions, 5 minutes filling out the SUS (System Usability Scale) questionaire 10 minutes discussing problems the user felt they had and any features they liked and finally 10 minutes doing the PSSUQ (Post Study System Usability Questionnaire) questions and giving additional feedback (Figure 8-1).

The hardware used was identical to that in the first round of user testing (Section 6.2).

The first section of the questionaire was used as a user profiling tool and asked for information such as the length of experience the user had in their respective area of employment and their academic and/or professional qualifications. This is important as Helium was developed as an expert tool to be used by professionals working in the areas of plant breeding and genetics and thus its primary function was to meet the need of these specialist audiences and not the general public.

Throughout the testing process users were observed and the interaction with the Helium tool was recorded using Microsoft Expression Encoder ("Microsoft Expression Changes | Microsoft Expression" 2014). Audio was not recorded and the users undergoing testing were made aware of this before carrying out the testing. It was important that users talked freely and openly. The screen captures would be used for clarification if required during the analysis of the testing data.

8.3.1 *Pre-study questionaire*

Question 1 from this section asked the test participant if they had taken part in the first round of user testing. In total there were 28 test subjects who undertook this round of user testing (11 in Mexico at CIMMYT and 17 at The James Hutton Institute site in Dundee). Of these 28 participants 8 indicated they had taken part in the first round of user testing (Chapter 6).

Question 2 asked for the test participants main job function. The results were broken down as follows: 14 geneticists, 6 breeders, 4 bioinformaticians, 1 cytogeneticist, 1 genebank manager, 1 research manager and 1 statistician. Question 3 asked for an indication of education level, 26 were eductated to PhD level and the remaining 2 to MSc level. Question 4 asked the user to indicate their amount of experience working in their current field. The minimum experience in the field from the group was 2 years and the maximum experience was 34 years. The test group had an average experience of $\bar{x}$=15.04 years, $\tilde{x}$ = 12.5 years with standard deviation σ_X = 9.21. The distribution of experience in years can be seen in the histogram shown in Figure 8-2 below.

Figure 8-1 Helium user testing in CIMMYT Mexico

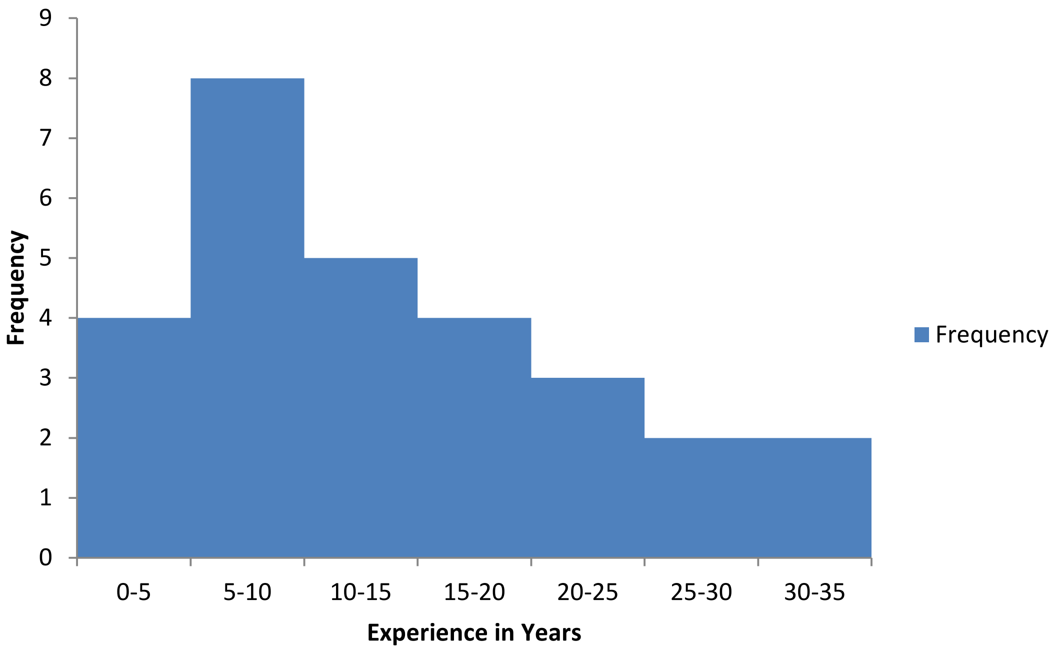


Figure 8-2 Test user experience

Questions 5 and 6 showed that 24 out of 28 (85.7%) of the respondents indicated that they use pedigree data as part of their work. And from this 20 used it on a monthly basis, 3 weekly and 1 on a daily basis.

Question 7 asked the user if they considered there was a problem with the current pedigree data that exists in terms of errors or problems in the way it was stored. The responses showed that 21 out of 28 (75%) thought there were issues with current recording, storage and handling of pedigree data while 6 (21.4%) thought there were no issues. One person did not answer this question.

The raw data for this first section can be found in Appendix 7.

8.3.2 *Pedigree and interface components – Quantitative task performance testing*

A series of 8 questions were asked in this section 6 of which were identical to questions asked in the first round of user testing. Of the questions that were different this was to get the users using the phenotype classifications which were problematic in the first round of testing. The aim was to get data that allowed a comparison between the first testing and this testing to be performed, with particular focus on where improvements to the interface had been made. The time between the two rounds of testing was 12 months so enough time that users would have forgotten what they had been asked, or at least not been fresh in their memories when performing the first round of testing.

The questions asked in this section would require users to interact with Helium to perform basic pedigree type operations such as the tracing of lineage through generations and overlaying additional data types on to the visualization to retrieve morphological characteristics about specific plant lines. It would also be used to test the efficiency of the colour coding that is used in Helium to differentiate data categories (Section 7.1.3).

The 8 questions were assessed to see if the user managed to answer the questions correctly and a strict marking schedule was imposed. Some questions had 2 parts but if one part was wrong the question was marked as wrong. This binary marking gave results represented as either 1 being correct or 0 being incorrect. This is one of the most fundamental usability questions that we can calculate: can a user correctly answer a question. Calculations could then be performed to assess effectiveness of Helium in being able to answer simple pedigree based questions that would be common in plant breeding. The results of these questions can be seen in Appendix 8 (raw data), Table 8-1 and Figure 8-3.

Question	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
Correct	24	25	21	21	26	28	27	23
Correct (%)	86	89	75	75	93	100	96	82

Table 8-1 User testing section 2 question responses. This table shows the responses for the 8 questions in section 2 of the user testing questionnaire. The correct results are a sum of the number of binary encoded 1's from the raw data which represents a correct response to the question. The total number of respondents was 28. The average row shows $\frac{correct\ responses}{total\ number\ of\ responses}$ or $\frac{n}{28}$ * 100 which gives us the percentage of correct responses to the question.

Table 8-1 shows the number of correct responses from a total of 28 test subjects. The average correct responses across all questions in this section was 0.87 (or 87% sample completion rate) which shows a high success rate. It is however clear from this data that some questions, such as Q3 'What are the great-great grandparents for the plant line Agenda?' and Q4 'What are the grandparents of the progeny of Oxbridge?' scored lower than the others. These questions can be classified in much the same way as questions 4 and 6 from Table 6-2. While the results here are low for these questions they are an improvement on the results in Table 6-2. Figure 8-3 shows the results from Table 8-1 but with calculated test statistics for margin of error and the upper and lower 95% confidence limits. These results show that the margins of error (and other test statistics) are reasonably low which would indicate that results were tightly clustered with a low degree of variation within the dataset.

The confidence limits give an indication of the most likely range of values from the data in an unknown group if they were asked to perform this task as it would not be the results that were obtained here (0.87). The range therefore gives an indication of the expected results from an unknown group of users. This is done by calculating the binomial confidence interval around the sample proportion (Jeff Sauro and Lewis 2012). In order to calculate this the Adjusted-Wald Interval (Agresti and Coull 1998) was a simplification of the methods proposed by Wilson (Wilson 1927).

The results in Figure 8-3 show the 95% confidence limit around the successful completion of the tasks. The p (maximum likelihood) and p (Wilson - binomial proportion confidence interval) give an indication of the probability of obtaining a value bettering, or at close to the observed test statistic. The margin of error is expressed in conjunction with the confidence interval (the degree of uncertainly) and expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter. These results are an indication of the successful completion rates that would be expected were the testing to be carried out on a larger population size than the 28 used here as using a different number of test participants would be unlikely to give exactly the same results as the testing carried out here. They give an indication of the likely range of successful completion rates that could be expected (defined by the 95% limits). The results therefore give an indication of the most likely, or plausible range for an unknown population size, the results would be defined as likely or plausible if they lie within the indicated by the 95% upper limit and 95% lower limit categories in the figure.

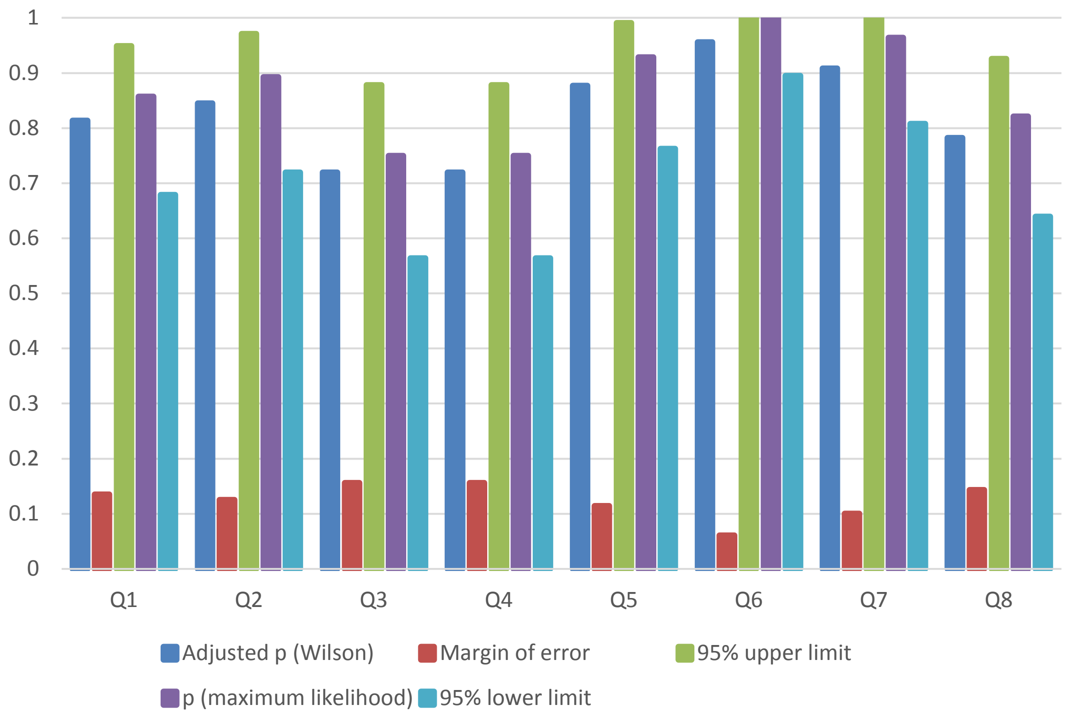


Figure 8-3 Confidence interval for completion rate user testing section 2 'User Interface Components' section. 95% confidence interval around the completion rate for the 28 users where users completed the task successfully.

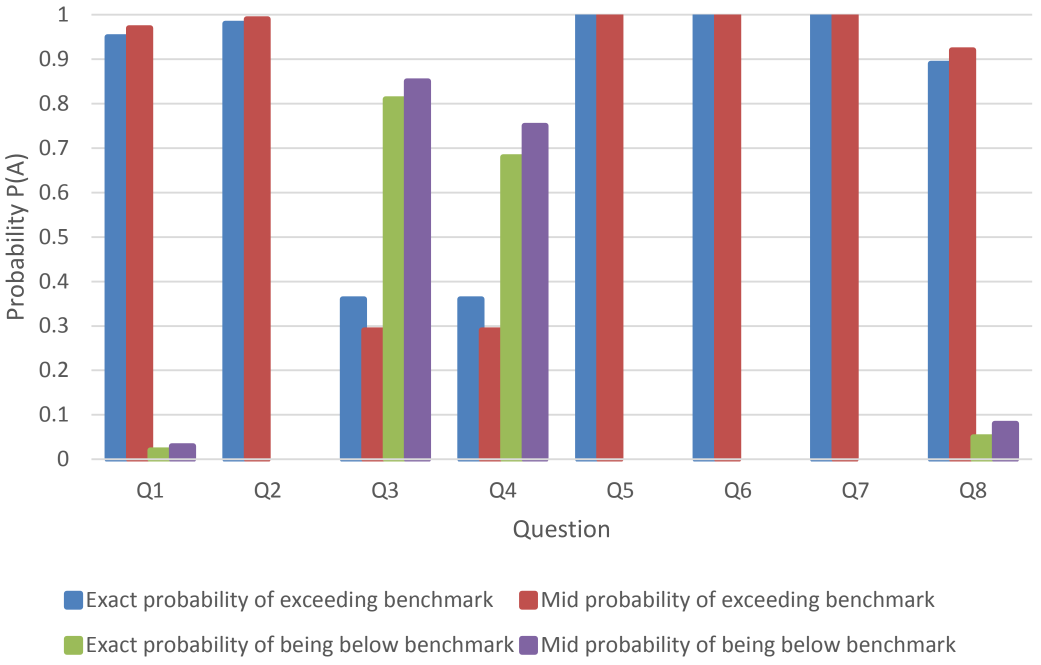


Figure 8-4 Comparison of success rate against 70% benchmark

While the confidence intervals (95% upper and lower limits) are an indication of the precision of our testing and an indication of the range of results that would be obtained from a sample group conducting this test, the ability to assess the probability of meeting a defined benchmark would be advantageous.

The results in Figure 8-4 show that there are immediate problems with questions 3 and 4 which related to great grandparent tracking and grandparent tracking. While this is an improvement on the first round of user testing only 75% of responses are correct and thus room for improvement. The results here show the probability of exceeding a sensible benchmark score of 70%. The score of 70% is chosen as a level above which it would appear that the application is performing its function well (Jeff Sauro and Lewis 2012). With the exception of questions 3, 4 and 8 the remaining questions all show a statistically significant probability ($P<0.05$) of exceeding a benchmark score thus indicating a high chance of answering the questions correctly across users. Questions 5, 6 and 7 are statistically highly significant with $P<0.1$.

8.3.2.1 Section A Feedback

This section was a series of 6 questions which asked users to rank a number of statements on a Likert scale from ‘Very Difficult’ to ‘Very Easy’. These questions were aimed at getting feedback on how easy or difficult the test users found with concepts within the pedigree visualization. Results were scored on a scale from 1-5 (Very Difficult to Very Easy) then averaged. These results are shown in Table 8-2 and Figure 8-5 below. Raw data can be found in Appendix 8. The boxplot shown in Figure 8-5 clearly shows the increased variability of Likert responses to question 3 which asked users to score how easy they found relating colour coding to phenotypic classes.

	Q1	Q2	Q3	Q4	Q5	Q6
Mode	5	4	3	4	4	5
Min	4	3	2	4	2	4
1st Quartile	4	4	3	4	4	5
Median	5	4	3	4	4	5
3rd Quartile	5	5	4	5	4	5
Max	5	5	5	5	5	5

Table 8-2 Likert response statistics

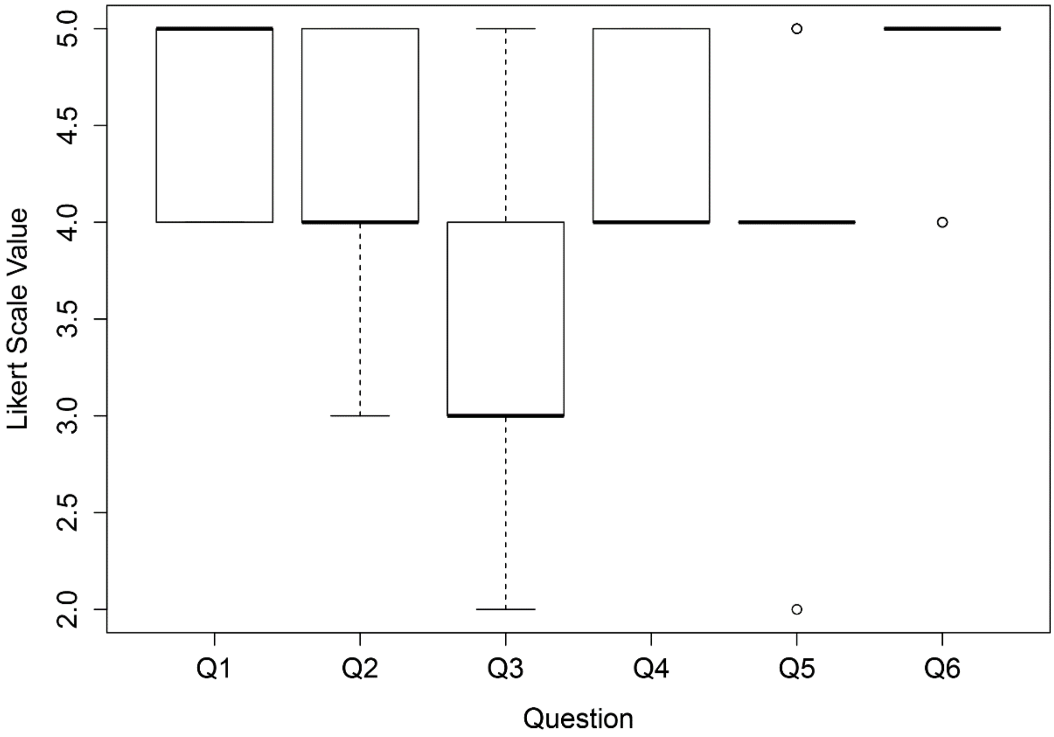


Figure 8-5 Likert Response Boxplot showing the variability of Likert response scores

It is also important to recognise that the Section A Feedback Likert response data is not normally distributed (Figure 8-6) with a clear skew towards the positive end of the response scale. Because of this, care is required when interpreting summary statistics. The potential reasons for this skewing of data are discussed in Section 8.4.

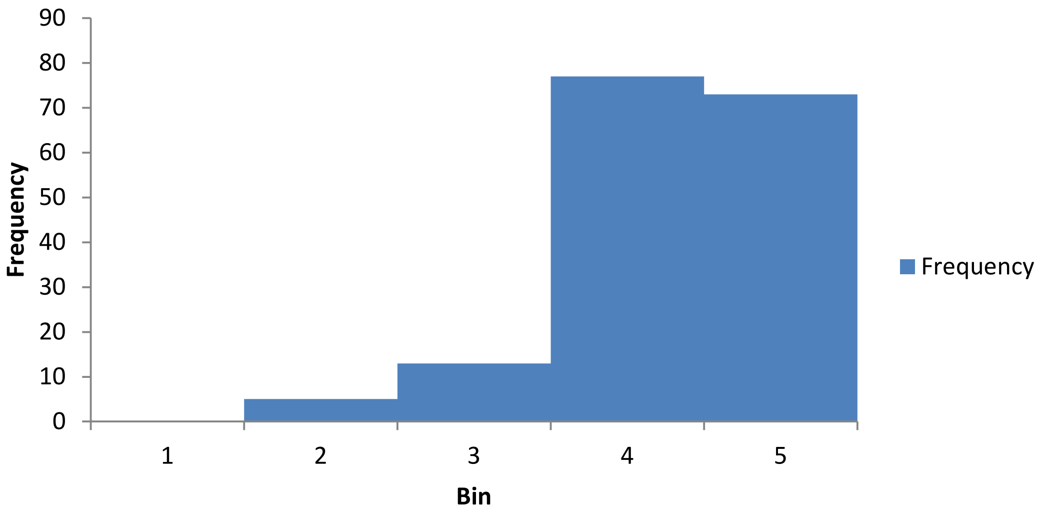


Figure 8-6 Section A Feedback Likert response value distribution

Table 8-2 shows test users and their standard deviation across the Likert type responses. The average standard deviation across all *users* was σ = 0.76. Standard deviations are shown in Figure 8-7 which shows data across individuals used in the testing. The results show that there is a low standard deviation across each of the questions for most users. This indicates that most users consistently scored questions (either high or low), there was no major shift and variation on the Likert 1-5 scale. However, for some test subjects (5, 11, 15 and 22) there was a high standard deviation which points towards a larger range of Likert responses for the specific user indicating they were more positive about some questions than others. It can be argued that the test subjects that gave higher standard deviations were potentially answering questions more carefully and not just ticking the same Likert score for each question.

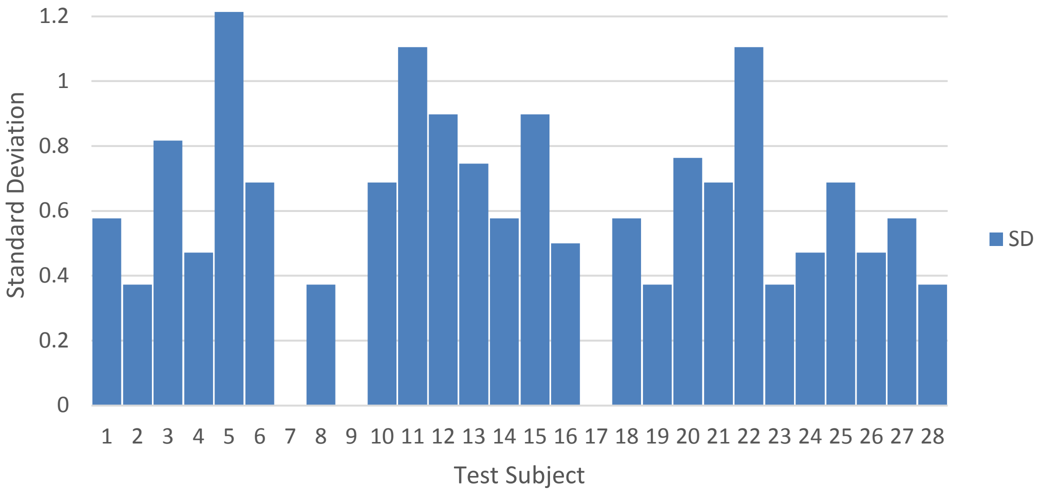


Figure 8-7 SD Likert responses / user Section A

Three users scored consistent scores across all 6 questions. Test subject 7 consistently scored 5 (Very Easy) for all questions and user 9 and 7 scored 4 (Easy) for all questions hence no data on this chart. While there is debate over the use of using the standard deviation and mean values with ordinal data not displaying a normal distribution, they are presented here in some circumstances but should be used with care.

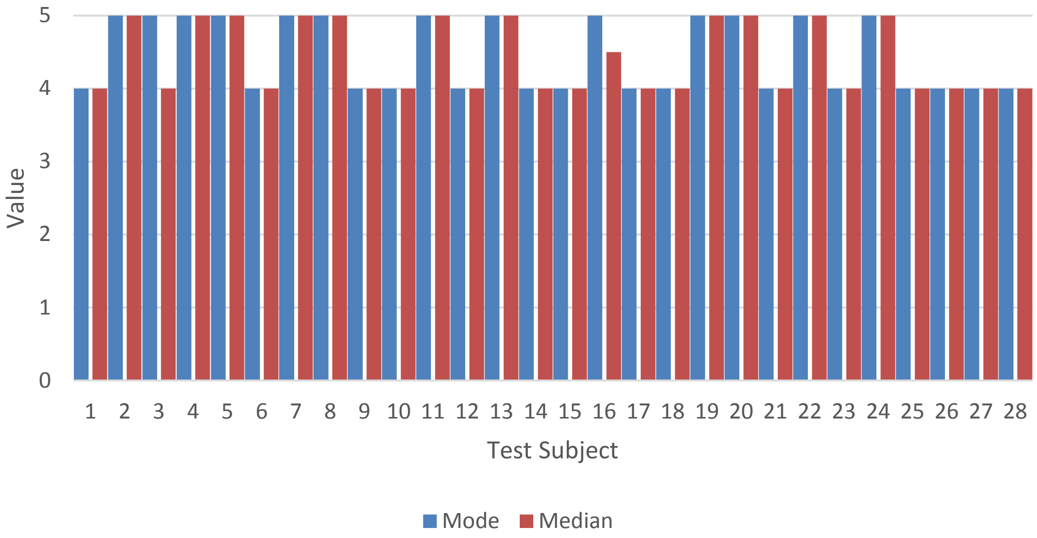


Figure 8-8 Mode and median values per user

Figure 8-8 shows mode and average values registered by each user in this part of the user testing. This chart is one indication of those users with a greater tendency to give erratic results and those more likely to pick consistent results across the questions. Examination of data such as this can be a good indication of users who just pick one response then apply it to all questions and those who genuinely try to answer to the best of their ability, although this may be somewhat controversial.

8.3.3 *Section B Advanced pedigree and interface questions*

Section B was comprised of three more advanced questions to get users using the interface and performing more difficult pedigree tasks involving varying data types and multiple areas within Helium. Questions 9, 10 and 11 gave average correct response rates of 79%, 89% and 89% (22/28 and 25/28 correct responses).

Figure 8-9 details margins of error and p values for these questions across the 28 test subjects. Raw data can be found in Appendix 10.

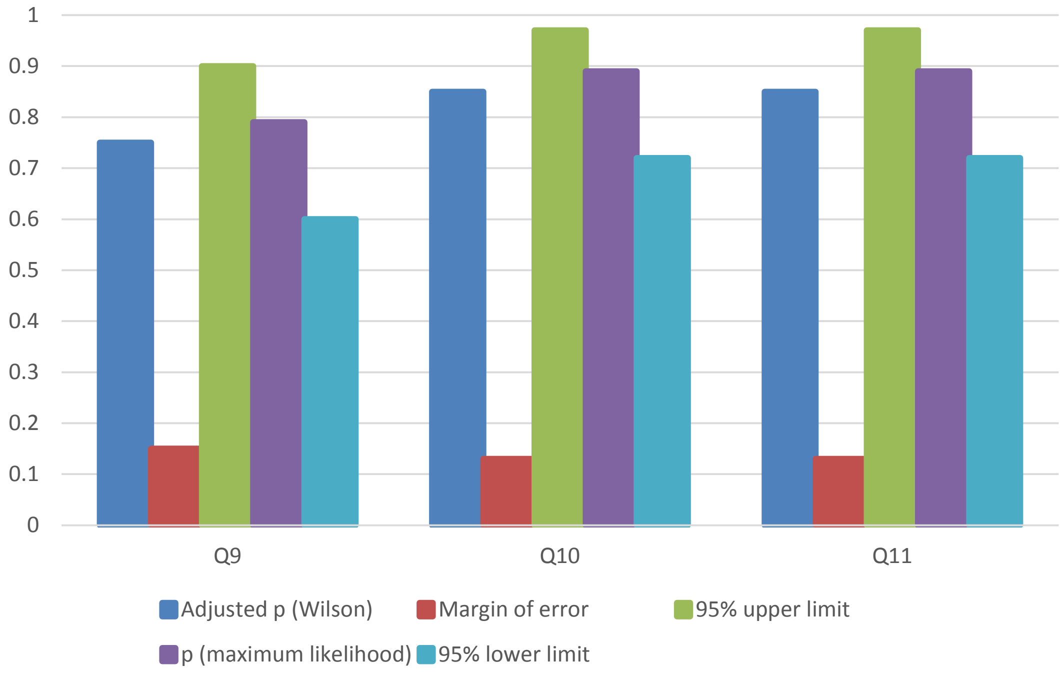


Figure 8-9 Part 2 responses

These results show that more or less there was a good response and accuracy when carrying out these tests. Question 9 has a slightly lower success rate (22/28) when compared to questions 10 and 11 but overall the likelihood of obtaining a results at least equal to these results (p0.79 – p0.89) is high. Figure 8-10 shows the average successful completion rate of each of the three questions in this section.

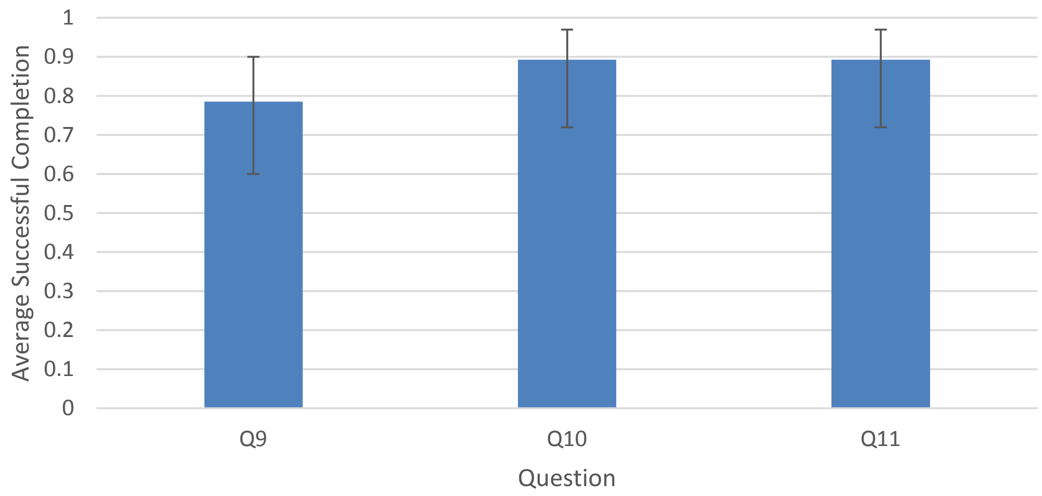


Figure 8-10 Average successful completion rate

Another way of looking at this data is to give an indication if after completion of the questions the Helium interface has met or exceeded defined goals. With small sample sizes a mid-probability binomial test is used (Figure 8-11).

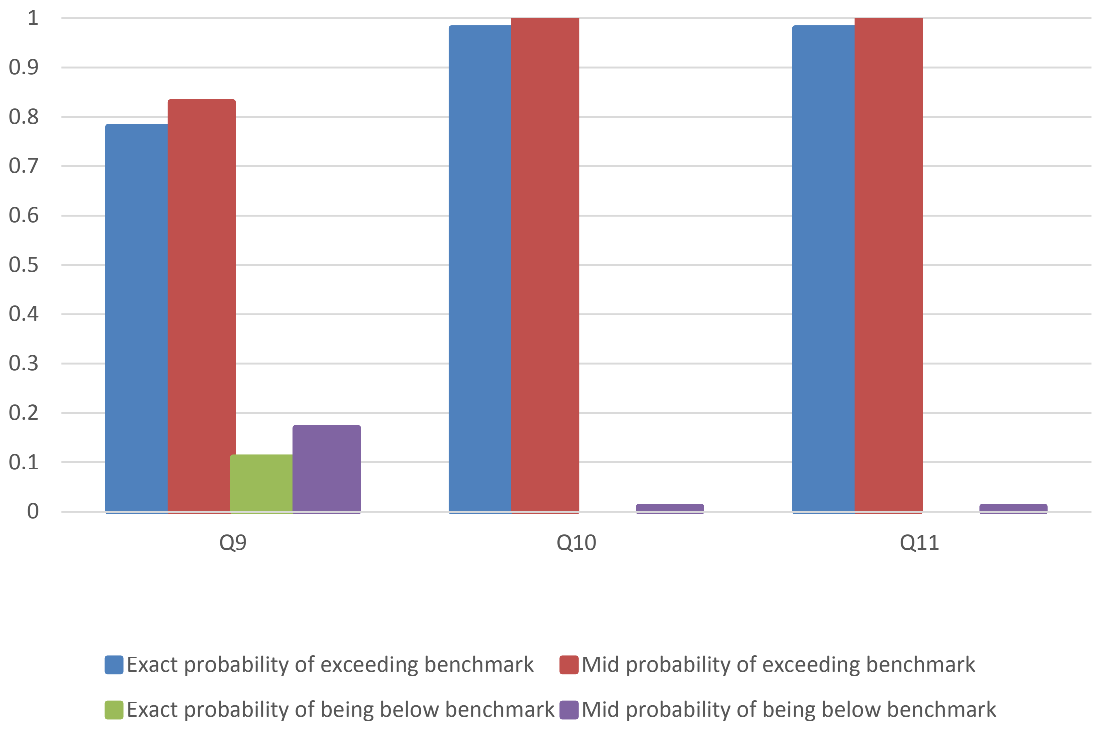


Figure 8-11 Success rate v 70% benchmark

What this tells us is that there is a high probability (Q9 $p < 0.17$, Q10 $p < 0.01$ and Q11 $p < 0.01$) of exceeding the 70% benchmark level. This equates to an 83%, 99% and 99% chance of exceeding the benchmark (with an unknown population/sample size) for these questions which is good.

For completion, Figure 8-12 and Figure 8-13 give the overall scores for all 11 questions in sections A and B combined into a single chart.

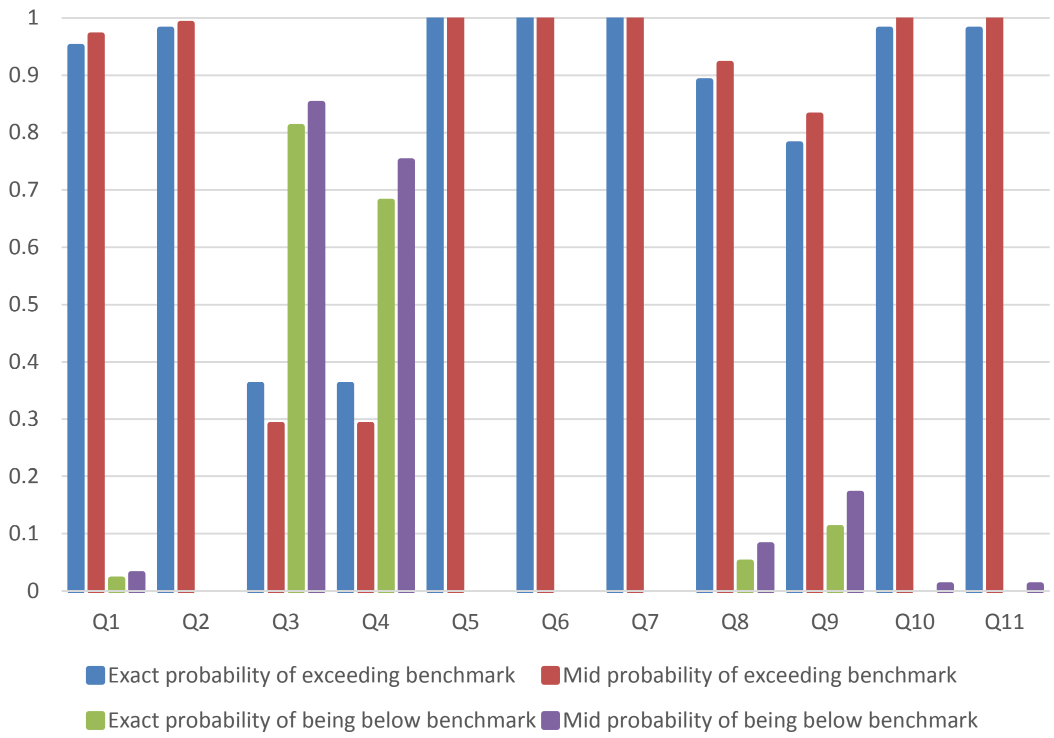


Figure 8-12 Comparison of success rate against 70% benchmark

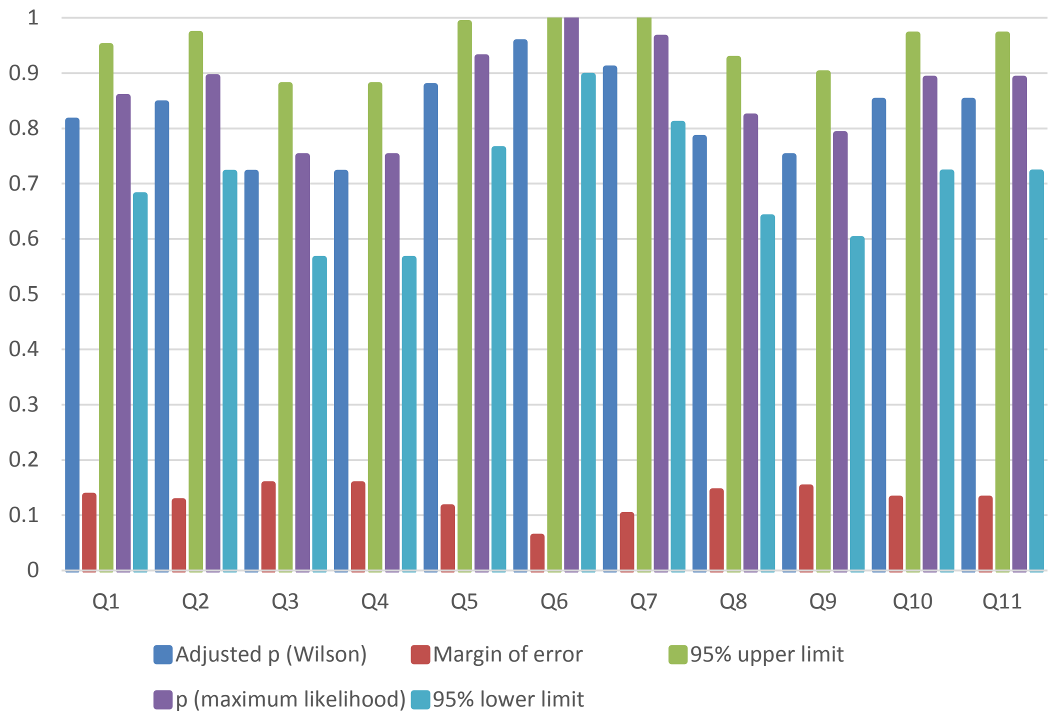


Figure 8-13 Confidence interval for completion rate user testing section 2 'User Interface Components' section. 95% confidence interval around the completion rate for the 28 users where users completed the task successfully.

8.3.4 *Section B SUS (System Usability Scale) results*

The results of the SUS analysis (Appendix 11, Appendix 12) show that the overall SUS mean score was 81.1 (75.2 -87 95% confidence interval) median and standard deviation of $\tilde{x}$ = 85 and σ_X = 14.99. The margin of error is 5.92. Items 4 and 10 represent a learnability dimension and the rest a usability dimension to the data so taking this into account the calculated SUS scores for these 2 subscales were; usable scale 81.2 (74.7-87.8) and learnable scale 80.4 (73.6 -87.1) Values in brackets denote 95% confidence intervals.

Sampling of SUS analysis data by Sauro (J Sauro 2011) puts the average SUS score across 68 usability studies at 68. This data can therefore shows that the SUS score of 81.1 is well above the industry average. The probability of exceeding the benchmark score of 68 is 0.99 using results from these test users which means that there is a 99% chance that Helium has an average SUS score in excess of the industry average.

The SUS scores were calculated by assigning the data on a 5-1 scale, strongly agree to strongly disagree then for each odd question number 1 was subtracted and each even question the value was subtracted from 5. This puts the data into a consistent scale ranging from 1-4 and reverses the scores for the negatively poised questions. The final SUS score is then calculated by multiplying the sum of each user by 2.5. These results for the summed (also called total) and derived or calculated SUS are shown below in Figure 8-14.

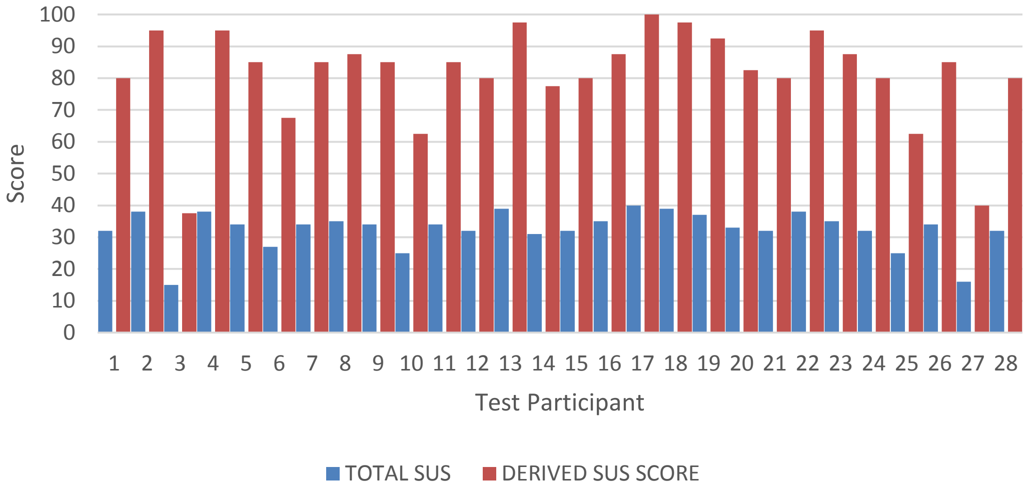


Figure 8-14 Total v derived SUS score

The distribution of scores is shown in the histogram in Figure 8-15. The higher the score the more positive the response to the questions in the SUS questionnaire. The histogram is a good representation of how skewed the data actually is in favour of the higher, and more positive SUS scores.

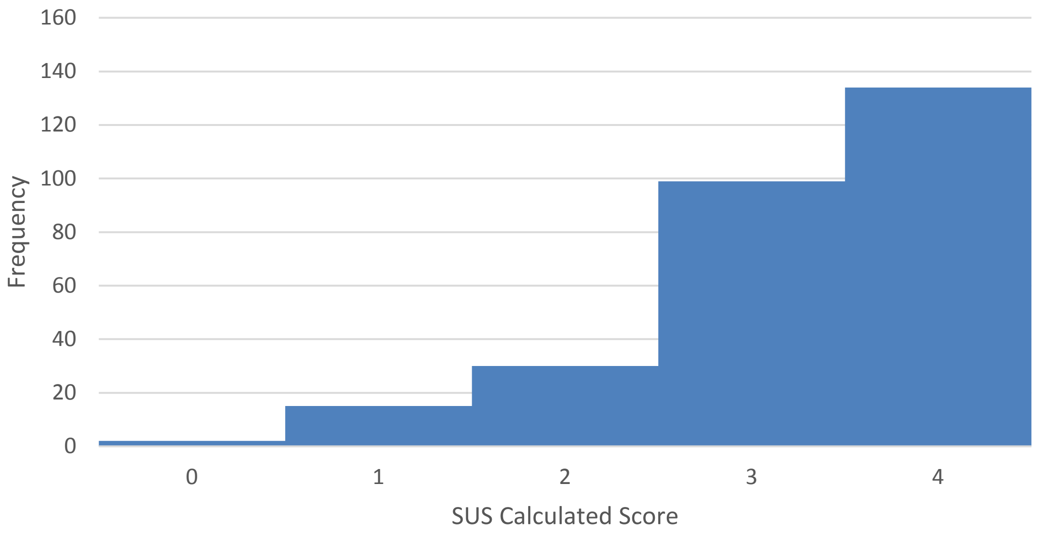


Figure 8-15 SUS score distribution (calculated SUS)

To highlight the nature of SUS questions; where there is a positive and negative switch between consecutive questions Figure 8-16 shows the raw average SUS scores across the 10 questions that make up the SUS questionnaire. The alternating values highlight why there is a requirement to convert the values into an appropriate merged scale before comparisons or calculations can be performed. The converted scale is shown in Figure 8-17.

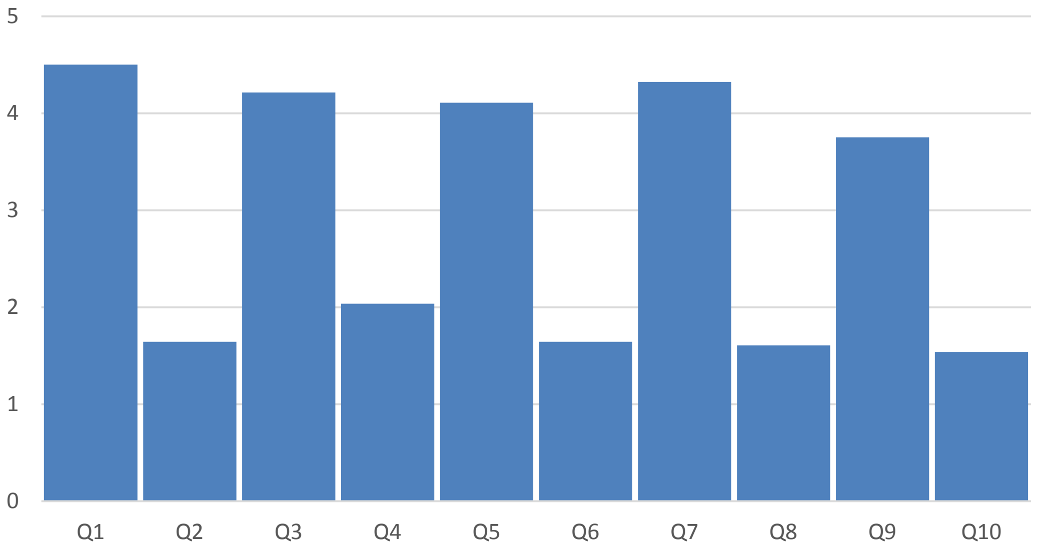


Figure 8-16 Raw SUS average responses

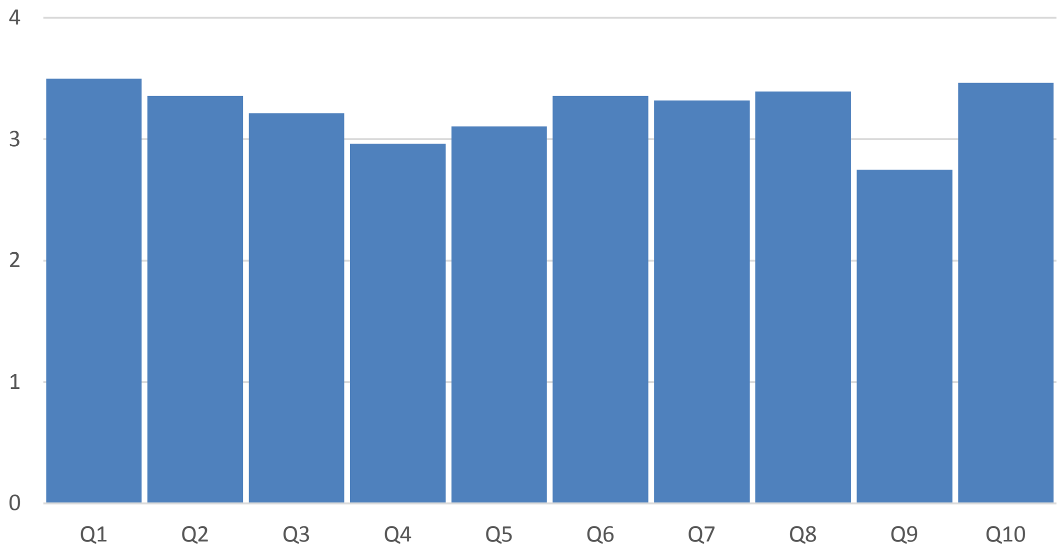


Figure 8-17 Calculated SUS merged scale

8.3.5 *Part 3 Post study questionnaire (Section C)*

Users were asked four questions in this section which was included as a way of getting more general feedback on the Helium visualization tool. The results from these questions are detailed here.

8.3.5.1 Question 1 Confusing elements in the Helium interface

Test users were asked if there was anything they found confusing about the Helium interface.

The main issue that users had was in the use of the colour scheme palettes used with the categorical datasets. Users found this to be confusing which was also highlighted in the first round of user testing (Section 6.3.3.2).

Users were generally very positive but found the genetic data similarity panel to be difficult to understand. This was a recurring issue with users. They also did not like the fact that they could not select plant lines from the local view panel and often tried to click on local view nodes and wondered why the main visualization window did not reflect this.

8.3.5.2 Question 2 Clear ideas and concepts in the Helium interface

The test users found that the layout of the application and the various panels within the Helium interface was very clear and uncluttered. The main positive feedback was that the way in which the pedigrees were represented (both nodes as plant lines and edges as mating events) was very clear and easy to understand. Users also liked the links between each of the information panels and how each updated based on selection from the main pedigree visualization window.

Users also commonly commented on how easy and intuitive the search functionality was within Helium and they liked how we had resized nodes based on the number of times they had been used as a parent. Users thought this feature was a really nice way of showing the plant lines which are used most often.

Finally users commented that the local view panel and the ability to remove nodes which were not related to a cross was excellent and increased their understanding of the structure of the pedigree.

8.3.5.3 Question 3 Questions you would like to be able to answer

This question focussed on things that users would not already be able to do with the Helium interface.

The main response to this question was that users wanted to be able to integrate their own data. What was striking about the testing was that although these were barley datasets, when we did testing on maize and wheat researchers they immediately were asking to be able to get their own data imported into the tool.

Users wanted to be able to import phenotypic and genotypic data and filter results based on these datasets. They also wanted to be able to look for genotyping or potential phenotyping errors based on their datasets and highlight these in the visualization window.

8.3.5.4 Question 4 General comments

The general comments was a section into which test users could note down any general comments or suggestions.

A recurring comment was that users wanted to be able to select alternate colour schemes. In the test version of Helium this was limited to a single ColorBrewer 2 BuGn colour palette for consistency.

While the example genotypic data focussed on genetic similarity; users wanted to be able to overlay specific genetic loci/gene specific data as well.

Another common comment was that it would be nice to be able to export both the data and pedigree images that could be used in publications and presentations.

In general the feedback from this testing was that Helium was a very useful program that is easy to use with an intuitive interface. Users commented on how responsive the interface was with our test datasets but some users (in the wheat community) had doubts over the ability to handle upwards of 100,000 plant lines. In reality no user is going to need to look at a pedigree of that size in its entirety.

One of the outcomes that came out of the user testing was that in more than half of the tests the test subject took more than the allotted time, not to complete the actual testing but in staying at the end to talk about the software and play with it. This is a good indication that Helium is, or at least has the possibility to be, a useful tool for plant breeders and geneticists.

8.3.6 *PSSUQ (Post Study System Usability) (Section D)*

The final section of the user testing asked users to complete a standard PSSUQ questionnaire which was used to assess users’ perceived satisfaction with the Helium interface once they had performed the defined tasks. The questions can be broken down into 4 main divisions; an overall score (Overall), a system quality score (SysQual), an information quality score (InfoQual) and finally an interface quality score (IntQual)(Figure 8-18)(Appendix 13, Appendix 14).

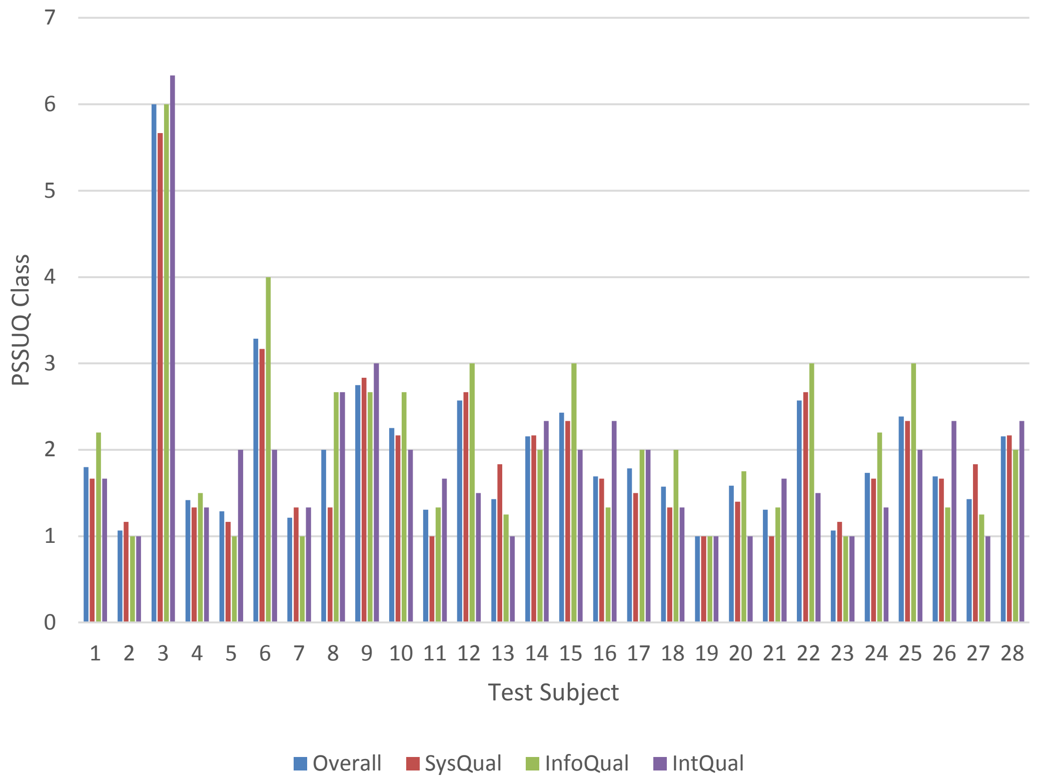


Figure 8-18 PSSUQ results

This figure has each user on the x-axis then the PSSUQ score on the y-axis. What this shows is that user 3 gave an unusually high set of scores in the questionnaire. The problem here is that PSSUQ scores were the reverse of the SUS scale thus the lower the better in this case. The test for user 3 was carried out near the start of the testing and the reason for this scoring was that the user, who was not a native English speaker, was moving quickly through the questions and it was clear that they were not carefully reading the statements. This can also be clearly seen in Figure 8-19 where average PSSUQ classifications are plotted against users. Even were a user having problems with an area of the user interface you would expect some of the questions to have a positive reaction, this was clearly not the case. Another potential for errors included the problem with the SUS being alternating positive negative on a scale from disagree to agree and the PSSUQ questions being short questions with no negative bias, and running on a scale from agree to disagree. If a user was not careful, and this was mentioned by a number of them, then it would be easy to get caught out if answering without paying attention.

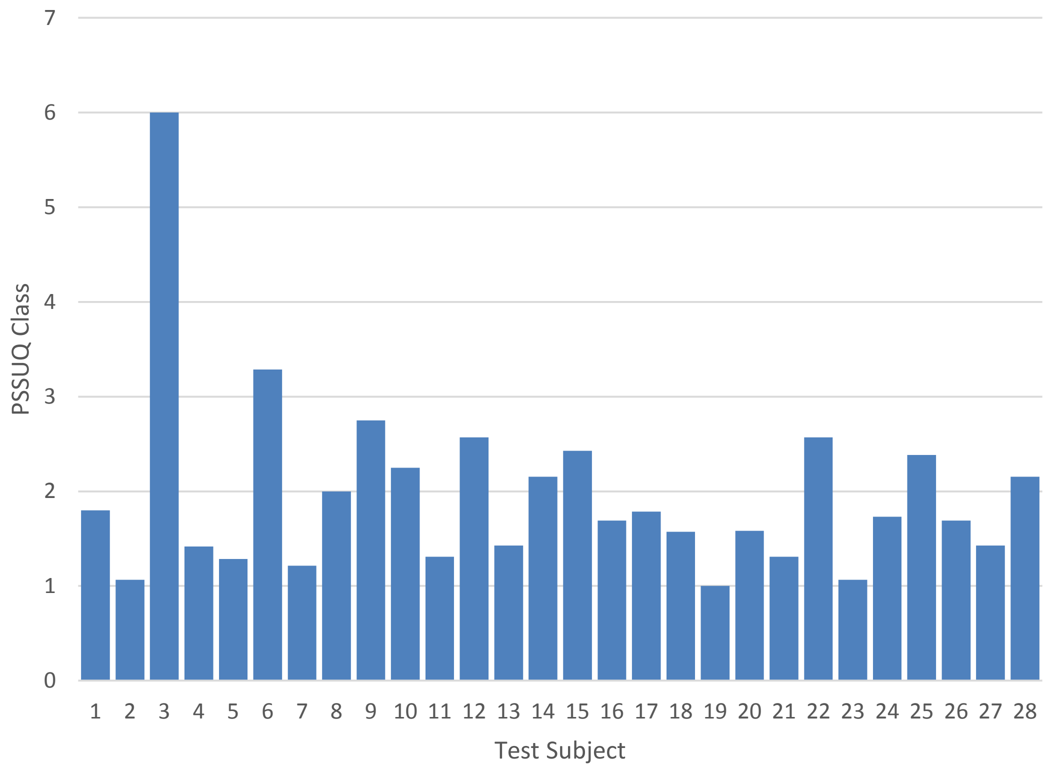


Figure 8-19 PSSUQ scores / test users

When plotting question number against average score an entirely different picture becomes apparent, and one that is referenced in the literature (J. Sauro 2004; J Sauro 2011). Question 7 can be seen in Figure 8-20 to be much higher than any of the other questions. The actual question asks the user to rate the effectiveness of error messages in the system, however, unless error messages are actually displayed, and they may not be for a number of valid reasons, then this usually results in the user crossing the NA option or marking it as poor. The raw data confirms this with only 2 scores (a 2 and 6 giving an average of 4) and 26 NA markings.

The low scoring of the remaining questions shows that users were happy with the questions and how they relate to Helium and have scored them highly.

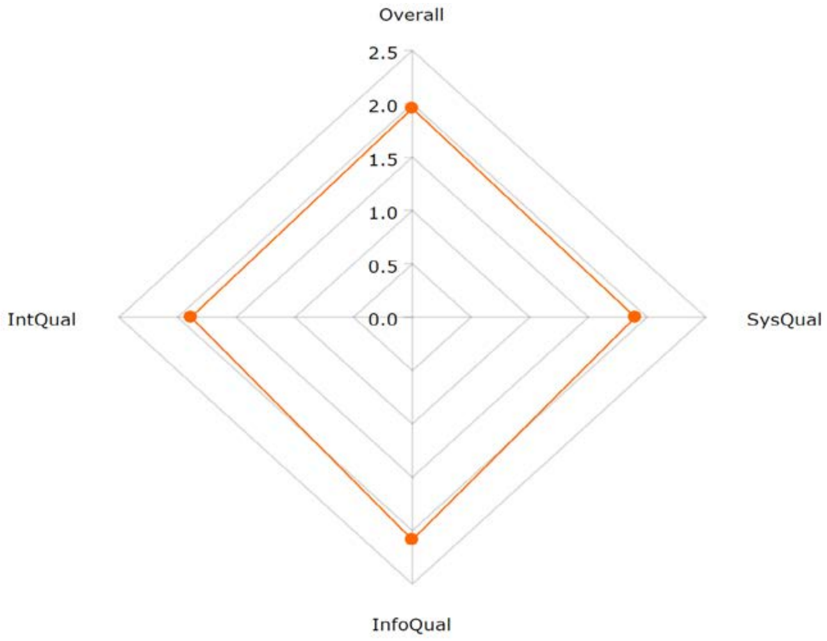


Figure 8-20 Radar diagram of 4 PSSUQ subscales

Figure 8-21 shows the 4 PSSUQ subscales (Overall (1.96), IntQual (1.88), SysQual (1.90) and InfoQual (2.09)) plotted as a radar diagram. This highlights again that InfoQual is affected by the high score for question 7.

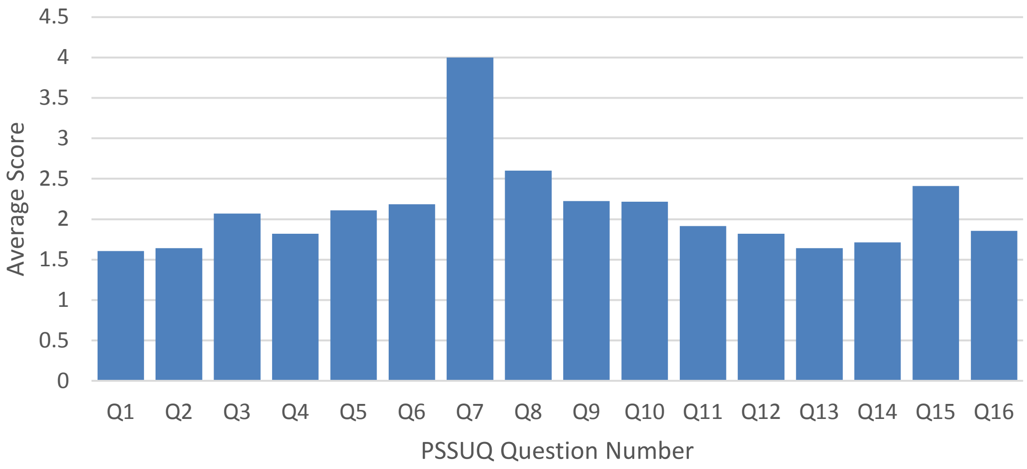


Figure 8-21 PSSUQ test users v score

Finally the distribution of PSSUQ scores shows a slightly skewed distribution towards the lower PSSUQ scores (higher acceptance) (Figure 8-22).

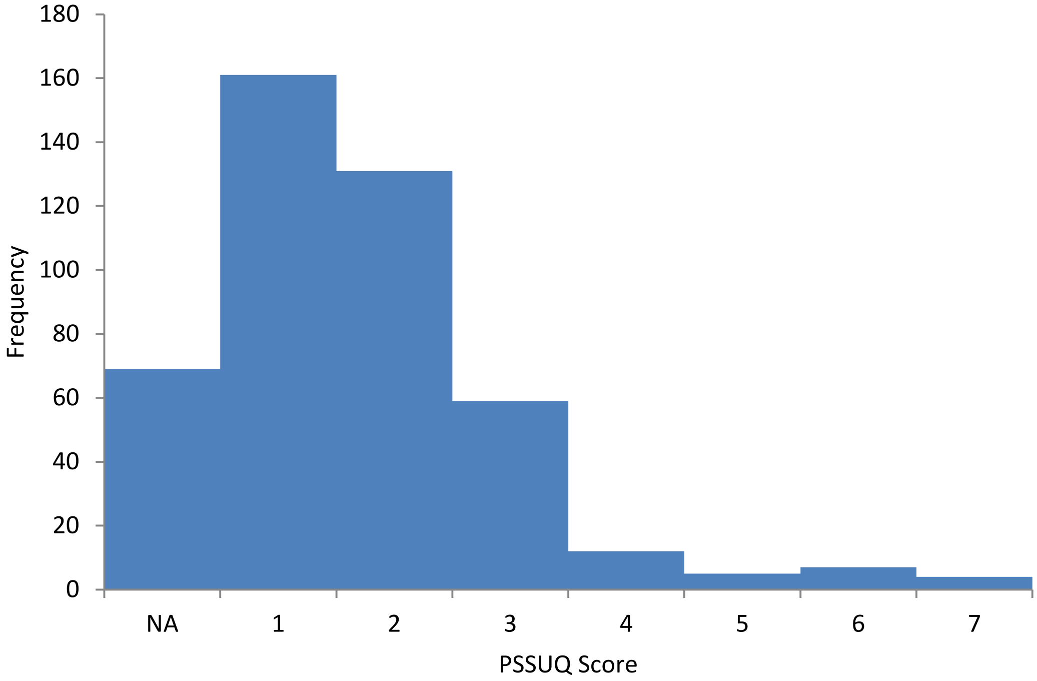


Figure 8-22 PSSUQ score distribution

8.4 Discussion

This second round of testing on a larger number of individuals was carried out to test the effectiveness of the tweaks that were made to the Helium interface which are described in Chapter 7. The testing was also carried out on a more diverse range of test subjects which was composed mainly of plant geneticists (14 test subjects) and 6 plant breeders (6 test subjects). While the median experience dropped slightly from 13.5 to 12.5 years between the two rounds of user testing there was an increased number of test subjects who carried out the testing. One of the interesting outcomes of this testing was that the test subjects were experts with a large amount of experience working with the sorts of data used in this work. With the exception of two PhD students every test subject was educated to PhD level and were active in the plant breeding and genetics community. The mixture of plant breeders and geneticists meant that both areas of work were well represented.

An interesting outcome of both sets of user testing comes about from the selection of the test subjects. The first round of testing was carried out predominantly with people who were known and who were, in the most part, happy to carry out the process. The second round of user testing however included two students who were asked to perform the test by their supervisors and it was clear that their enthusiasm was not that of their supervisors, one of whom ticked the same box for every question in the PSSUQ analysis which can be clearly seen in Figure 8-19 (Test subject 3). This poses the question as to whether the selection of the number of users (16 and 28 from the two rounds of testing) is sufficient to make any statistical assumptions on Helium as a small number of negative responses would adversely skew the data. Conversely, does the selection of specialist users in a small community lead to responses being more positive than expected due to familiarity with the person carrying out the testing, and not wanting to make negative comments on this work. This may be an issue in specialist areas where potential users are limited and you cannot perform the tests on the general public or users unknown to you. Sauro and Lewis (2012) state that there is an incorrect assumption that sample sizes must be above 30 to statistically analyse quantitative testing data and that a small test sample as one that has less than 15 positive and negative responses (Jeff Sauro and Lewis 2012). This does however bring into question the validity in user tests such as these where there is a low number of available test subjects where a low number of negative responses could potentially adversely skew data, it’s certainly something this should be acknowledged and taken into account when conducting future user testing in such specialised domains.

Another interesting outcome of this testing was when the test users were asked if they thought there was Question 7 where they were asked if they thought there were problems with pedigree data. 6 out of the 28 respondents indicated that they thought there were no general problems with pedigree data, these incidentally were the users with the least experience. There are well known problems with how lineage is described, something more appreciated by those with more experience in the area.

The results from this testing show that there was an increase in the amount of correct responses when users were asked to perform pedigree analysis tasks when. When looking at the comparable questions asked in the first round of user testing (Section 6.4.2 and Table 6-2), Simple Grandparent Tracking (Table 6-2 Category 2) which is the same as Question 1 in Table 8-1 showed an improvement from 86% correct responses to 93.75% correct responses (7.75% increase). Identifying Children, (Table 6-2 Category 3) which was the same as Question 2 in Table 8-1 showed an improvement from 56.25% correct responses to 89% correct responses (32.75% increase). Complex Grandparent Tracking, (Table 6-2 Category 4) which was same as Question 4 in Table 8-1 showed an improvement from 50% correct response to 75% correct responses (25% increase). Great Grandparent Tracking, (Table 6-2 Category 6) which was the same as Question 3 in Table 8-1 showed an improvement from 37.5% correct responses to 75% correct responses (37.5% increase). These results are a clear indication that the improvements that were made to the Helium interface as detailed in Chapter 7 have made a positive contribution to the accuracy of answering pedigree lineage type questions, something which was identified as what users wanted to be able to do with such a system.

It is not good practice to conduct statistical analysis on smaller groups of the sizes used in this work both 16 in the first round of user testing (Chapter 6) and 28 here. The ability of a few test participants to adversely skew data either positively or negatively is high. This is also a problem in a number of user evaluations that are carried out on software within the biological visualization domain with papers conducting detailed evaluation analysis on fewer numbers than carried out the evaluations in this work. These evaluations undertaken would come under heavy criticism from statisticians and it is perhaps an indication that more rigorous statistical reviewing is required. It is important to bear these problems in mind when examining the data presented in this work. While it is not deliberately misleading, caution must be taken when interpreting and drawing conclusions from the results presented here.

One potential issue that has been highlighted in this testing relates to the perception of colour used to differentiate ordinal data classes in the pedigree visualization. In Figure 6-1 and Table 6-2 for the first round of user testing there was a 56.25% correct response rate for using the colour coding within Helium to identify phenotype classes. In this second round of testing there was a 82% correct response rate (Question 8 Table 8-1) to the question in the testing that asked users to use colour coding then identify plant lines which were of a particular phenotype. While this was a definite improvement on the 56.25% rate from Chapter 7 there is clearly room for improvement. When conducting the user testing it was clear that users will still having issues distinguishing between shades of green on the ColorBrewer 9 class BuGn colour palette, even after including the features such as a clickable histogram (Section 7.3.2.4). This was highlighted by the high variability between Likert responses shown in Figure 8-5 for Question 3 which asked users to indicate if they found relating colour coding to phenotypic classes easy or difficult. There are a number of reasons why this was causing problems. Firstly, it was not always obvious to the test participant that they could click on the histogram to highlight plant lines which matched in the main visualization. Secondly, the number of colours on the gradient (9 in this case) meant that colours were very similar to each other and difficult to differentiate. Lastly, there is the potential problem with colour blindness although there was no indication that any of the test participants were affected by this.

Another outcome from the user testing was the enthusiasm in which most users talked about the data and the ideas that they had to improve Helium. Overall around 75% of user tests overrun their allotted time slots due to users wanting to play about with the system and talk about their own data. What was also clear was that it was highlighted that not only would Helium be a useful research tool; but it would also serve as an ideal platform for teaching crop genetics. This was not why the tool was developed so it would be exciting should it find a function in an educational setting.

The improvements that were identified in Section 7-2 have made a significant improvement to users’ ability to answer the types of pedigree problems that were identified as important. Tracing lineage of lines through previous and subsequent generations and the overlaying additional data types to identify plant lines which met specific criteria have been markedly improved. The improvements, in the case of tracing lineage the improvements to Helium by the inclusion of visual cues has improved an average success rate in the lineage problems in Table 6-2 (Questions 2, 3, 4 and 6) from 57.43% to 83.18% correct responses across all the questions, the largest increase being seen in the Great Grandparent Tracking (Table 6-2 Category 6) which showed a 37.5% increase in correct responses to the question.

Chapter 9 Conclusions and future work

This work has shown, through the development of the pedigree visualization tool Helium, that the visualization of real-life barley pedigree, genotypic and phenotypic data provides users with new insights into the genetics of crop breeding. The application of visualization techniques in the area of biological visualization, the development of a pedigree visualization tool and underlying database has allowed the development and verification of a useful plant pedigree visualisation tool, but is not without its limitations.

Helium's main research contributions are two-fold. Firstly it has applied visualization techniques used in information visualization (with regards to general layout principles) and applied them to the domain of plant pedigrees; this has addressed problems with handling large experimental plant pedigrees. The scale, complexity and diversity of data and the number of plant lines that Helium can handle exceed other currently available plant pedigree visualization tools. These techniques have been improved to deal with the differences that exist between human/mammalian pedigrees which take account of problems such as the complexity of crosses and routine inbreeding. The improvements have advanced both user understanding of pedigrees and allowed a much greater density and scale of data to be visualized. Secondly, the verification of the effectiveness of the visualizations has been demonstrated by performing two distinct rounds of user testing on a group of active domain experts. This testing has shown that the implementation and extensions to visualization techniques have improved user comprehension of plant pedigrees when asked to perform real-life tasks with barley datasets. When looking at the specific problem of tracing lineage the improvements that were made to Helium after the first round of user testing lead to an increase in correct responses to lineage tracing questions by 37.5% (up from 37.5% in the first testing to 75% correct responses in the second round of user testing). When overlaying phenotypic data and relating phenotypic categories to plant lines that expressed that character there was an increase from 56.25% to 82% correct responses after tweaks were made to the pedigree visualization. This user testing has also shown a high acceptance rate of the Helium software with the results of a SUS analysis giving a mean score of 81.1 (Section 8.3.4).

Helium has allowed the accurate representation of large complex plant pedigrees, something for which there were no current tools. Additionally, it has provided a means by which not only can these pedigrees be visualized, but also included the ability to overlay phenotypic and genotypic data in a user friendly interactive interface. Helium has addressed the requirements set out in the user requirements (Section 2.7.1) which included the ability to trace lineage of plant lines in complex pedigrees and overlay additional data types. It has also shown, through the identification and classification of plant lines within a pedigree (principal, flanking and terminal plant lines) that new patterns within these pedigrees can be seen. These classes are defined as; a) principal plant lines which are commonly used to generate new cultivars due to their possession of desirable characteristics b) flanking plant lines brought in to increase the genetic diversity of subsequent plant lines and less commonly used in crosses and finally c) terminal plant lines that are released, but have had little subsequent use. While this data had been available it had not been brought together in such a way that Germinate and Helium have facilitated.

Additionally Helium has highlighted and allowed definition of the term *pedigree net* which is a feature of not only the test barley datasets (Figure 4-11) but also those of wheat (Figure 4-7) and rice (Figure 4-8). This compares to what has been termed a *delta structure* which is often seen in animal breeding experiments.

The ability to identify errors in complex pedigrees such as these has been highlighted in the static paper prototype with users identifying lineage problems when presented with the pedigree visualization (these problems were fixed in the underlying datasets on identification by experts and therefore not present in the interactive prototype Helium).

This work has critically evaluated the currently available pedigree visualization tolls and methods and showed why current techniques and tools are not suitable for modern plant genetics and plant breeding applications. It has advanced the current thinking to identify the problems that exist with current visualization applications and show that pedigrees should be modelled as graphs instead of trees as this representation is often incorrect or confusing. Information Visualization techniques which improve user comprehension of such representations have been evaluated and verified using two rounds of user testing with domain experts.

Word of mouth from talking and presenting Helium at project meetings has led to contact from plant breeding companies wanting to look at their data with the application. This sort of enthusiasm from industry highlights that Helium is not just an academic tool that was developed with no applied use. There is considerable scope for working with companies and research groups to begin to introduce new features to aid in the plant breeding process and workflows. There has also been interest in developing Helium to aid in the teaching of genetics and plant breeding and a number of static posters have been created for breeding companies and educational purposes at cereals events within the United Kingdom.

Something which was obvious when setting up testing was that there was a clear bias towards people who actually wanted to do the user testing against people who would not have come forward. It is clear that this may have introduced bias in to results. While this may seem like it was selective in reality those who conducted the testing are the people who would ultimately use the software so it was perhaps no bad thing. The use of expert test subjects gives weight behind the testing results and from that end the test subjects that were available were not only experts, but had spent considerable time working in their areas of speciality. This is something which is not present in all user testing that is carried out with software and something which is relatively unique in academic software.

While Helium has been tailored to specific data types (genotypic/similarity, nominal and ordinal phenotypic data and pedigree definitions) it is intended to be a framework on to which, over time, additional data types can be added and work is currently ongoing with plant scientists and breeders to develop the Helium platform and Germinate to add additional functionality for plant breeders and geneticists worldwide.

The use specific nodes as landmarks in the visualization highlights that users are clearly using specific nodes to orientate themselves in a visualization which may have interest in the area of pre-attentive processing. It is not clear if this is due to the size of nodes, the colour or a combination of both factors. It is likely that the heightened salience of both large and colourful nodes are captured and processed quickly by users. This is probably true of the longer edges in the visualization which while have been shown to confuse users may have application in a user maintaining bearings, essentially acting like landmarks in the same way as the plant line nodes.

9.1 Limitations of this work

This work has dealt with large plant pedigrees of up to a few thousand individuals. In some plant breeding programmes this number is much greater and the ‘shape’ more closely resembles that of an animal pedigree. These pedigrees can have upwards of 20,000 individuals. Helium is not able to handle this sort of data which may preclude its acceptance and use in some areas. There is however an argument to be made that breeding programmes of such scale are not suited well to traditional pedigree visualization. The way the germplasm from these programmes is handled means that the ‘best’, whatever the definition of best may be, plant lines are selected and other plant lines discarded. In such situations where high volumes are involved there is scope for selection of plant lines by other techniques.

While the user testing was carried out with 16 then 28 users it would have been interesting, and perhaps added more weight to the results to have been able to conduct the testing across a larger section of the plant community, perhaps involving more commercial plant breeders and researchers working on additional plant species. This work was limited to barley, wheat and maize researchers.

One of the potential limitations of this work is that it dealt with a well curated barley dataset and therefore had a bias towards both barley data and well curated data. Helium has been used to look at pedigrees from commercial companies in a number of species. Additionally the static prototype has been used to create a number of wall posters in species including wheat, rice and citrus. The feedback from these, and from cereal researchers is that they would not envisage there would be significant differences in different datasets. Finally, the user testing involved barley, maize and wheat researchers.

9.2 Future work

There is clearly scope for additional work that could be carried out to improve knowledge in the area of plant pedigree visualization.

From a visualization perspective there is scope to look at the efficiency of current layout algorithms and whether these could be adapted to better suit the layouts that are expected in pedigree diagrams. This would be valuable in trying to help address the problems that people were having identifying parents and grandparents and the issues surrounding long edges which some users found disorienting. While the testing in Chapter 8 has shown that the improvements to the visualization detailed in Sections 7.3.1.1 and 7.3.2.3 (which describe the inclusion of visual cues and the local view implementation) there is potentially further scope for improvement to improve users ability to correctly answer lineage based questions to move it upwards of the 75% correct responses from the second round of user testing. Additional layout tweaks could further reduce this. There is scope to look at greater depth into graph aesthetics as a means to identify other variables which could be tweaked to increase user understanding of these pedigree representations.

There is also clearly scope for future work to be carried out in users' perception of colour coding when used with ordinal data such as 'continuous' phenotype scales and in colour coding for genetic similarity. There is however a debate to be had as to the efficacy of such an argument with regards to similarity data where there is a greater degree of tolerance between a scale showing similarity than that defining specific phenotypes. While there is room for misclassification based on a scale from 50-100% the difference between ordinal phenotype categories is potentially much more serious. While it seems reasonable, and good practice to use a single hue, multi-value based colour scales for genetic similarity data where small incremental changes in values may not be important, this testing has uncovered the need for better methods of representing ordinal data types which have a linear scale but differences between categories are critical. This is especially important in the context of this work with varietal testing and differentiation of DUS categories.

In the biological domain the most important aspect of future work is to get Helium out into the hands of more breeders. While this has been done with test users and a number of companies on an ad-hoc basis; the interest that has been received from companies working on crops ranging from barley, wheat, citrus all the way through to melon and medicinal cannabis has shown that there is a market and indeed need for such visualization tools. There have already been discussions started with plant breeding companies about how we can move forwards with Helium and suggestions on additional features which can be added, this is an area which will be explored.

Larger pieces of work which were discussed with breeders during the duration of this work which would be the logical next step to be included into Helium include the ability to aid in the targeted genotyping of plant lines to allow inference of genotype data in unknown pedigree regions. This would draw on the pedigree framework and high density genotypic data to aid in the selection of plant lines which, when more data becomes available, would bring greatest benefit to what is known about the genetics of the pedigree in which they sit. This would involve the selection of key plant lines in a pedigree which have not been genotyped but whose genotyping would open up other parts of the pedigree for imputation work thus offering the potential for both cost and time savings in experimental work.

While Helium has used defined barley datasets (although testing has been performed on breeders pedigree datasets from other species), there is clearly a requirement for more advanced data importing and integration tools. The capability to connect to a Germinate database that exists may not be appropriate to all users of such a system and therefore the ability to import data from defined text formats would be advantageous. This of course includes the problems associated with integrating data from disparate sources and so work would be required on a suitable data integration interface for Helium. Such tools would provide users a more flexible interface into using the tool and pooling both phenotypic and genotypic datasets from a variety of sources for use with their imported pedigree data.

Biology makes extensive use of ontologies to describe features and processed in biological systems. There is a clear path towards using standardised plant phenotype descriptors that exist in plants, however the DUS data used as part of this work have not yet been mapped onto these ontologies. Being able to use ontologies in Helium as a means to view, query and browse data would provide a common platform and framework for phenotypic analysis and bring added value to the visualization tool.

Another area of potential future work could involve the development of tools to allow, based on pedigree, the imputation of commercially important traits which are difficult to measure (both in terms of costs and time). This would use Helium to visualize haplotypes responsible for these characters and use the pedigree and knowledge of inheritance to predict these characters in subsequent generations and crosses. This would be of significant interest to plant breeding programmes. Additionally, the use of genotypic data would also allow the verification of pedigree definitions using molecular techniques as opposed to historical records, which have been found to be often unreliable and/or incomplete.

In a similar vein, the inclusion of Identity by Descent (IBD) data into the pedigree visualization would allow the examination of regions of IBD as they move through pedigrees. This would allow the categorisation and visualization of how IBD regions flow through pedigrees. This not only would provide researchers with an indication of the inheritance of specific regions of commercial or academic interest but also an indication of how these regions may change through generations. Tools such as these have both commercial and academic importance.

While the recognition of patterns has been demonstrated by this work there is a clear scope for researching the use of data mining as an automated means to facilitate pattern, and subsequently knowledge discovery in the underlying pedigree, genotypic and phenotypic datasets that Helium uses. This is especially pertinent as dataset sizes and pedigree complexity increases as well as the pressure on commercial breeding companies to develop new plant varieties to meet customer demands. Such tools would conceptually sit well within the current Helium visualization framework.

Finally, the user testing in Helium has been crucial in the continued development and refinement of the system, however, problems which have been highlighted (Section 8.4) relating to the suspicion that users may respond positively to questions because of familiarity with the individual carrying out the testing could be addressed. This could be carried out by conducting another series of user testing but using blind testing where the individual carrying out the testing has no involvement in the visualization tool, nor is known to the test subjects. This may lead to a more honest appraisal of the visualization tool. One problem however would be finding expert users who had no knowledge of the work that has been carried out in this area, something which could be addressed by going outside the barley community and carry out testing on another species user community.

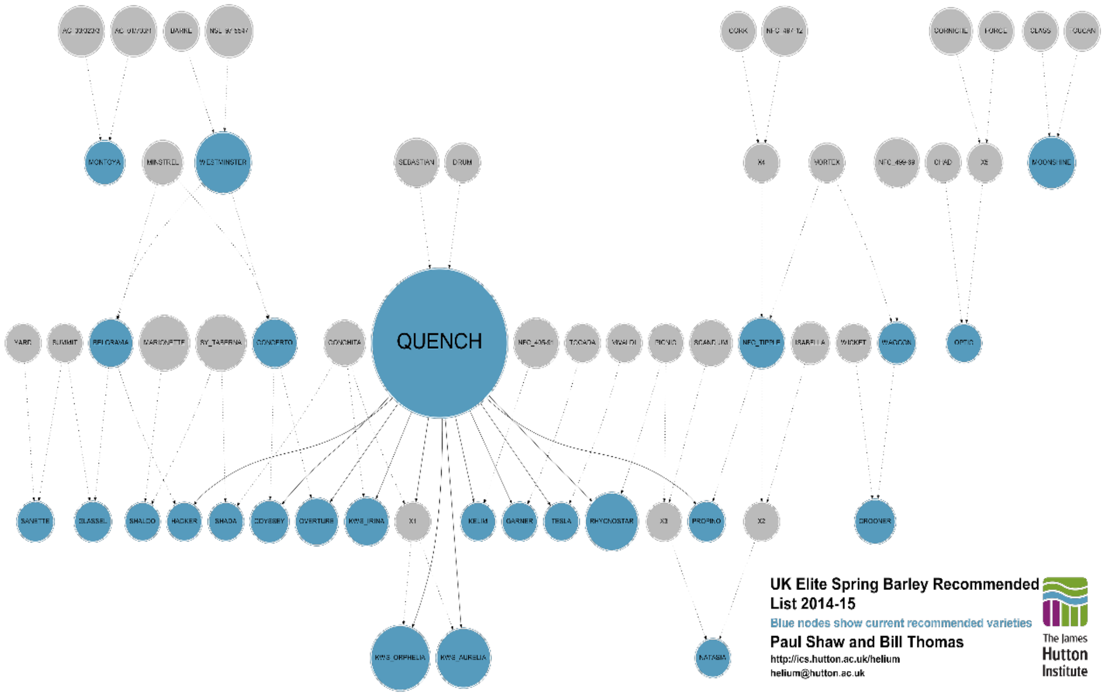


Figure 9-1 Current UK barley Recommended List showing the dominance of Quench

References

Adrion, W. Richards, Martha A. Branstad, and John C. Cherniavsky. 1982. “Validation, Verification, and Testing of Computer Software.” *ACM Computing Surveys*. doi:10.1145/356876.356879.

“AGOUEB -Association Genetics of UK Elite Barley.” 2014. Accessed December 3 2014. http://www.agoueb.org.

Agresti, Alan, and Brent A Coull. 1998. “Approximate Is Better than ‘Exact’ for Interval Estimation of Binomial Proportions.” *The American Statistician* 52 (2): 119–26.

“Agronomix Software - AGROBASE Generation II® Plant Breeding Software.” 2014. Accessed December 3 2014. http://www.agronomix.com/.

Ahlberg, C., and B. Shneiderman. 1994. “Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays.” In *Human Factors in Computing Systems, Chi ’94 Conference Proceedings - Celebrating Interdependence*, 313–17. Boston, MA, USA: ACM Press..

Akiyama, Kenji, Atsushi Kurotani, Kei Iida, Takashi Kuromori, Kazuo Shinozaki, and Tetsuya Sakurai. 2014. “RARGE II: An Integrated Phenotype Database of Arabidopsis Mutant Traits Using a Controlled Vocabulary.” *Plant & Cell Physiology* 55 (1): e4. doi:10.1093/pcp/pct165.

Anderson, Craig. 2014. “BBC News - English Barley Supplied to Scotch Whisky Distilleries.” *BBC*. http://www.bbc.co.uk/news/uk-scotland-highlands-islands-25863920.

Andrienko, Gennady, and Natalia Andrienko. 2007. “Coordinated Multiple Views: A Critical View.” In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, 72–74. CMV ’07. Washington, DC, USA: IEEE Computer Society. doi:10.1109/CMV.2007.4.

Ardi, Calvin, and Simon Tan. 2002. “Color Palette Generation for Nominal Encodings.” http://vis.berkeley.edu/courses/cs294-10-fa08/wiki/images/6/66/FP-STCAKY-paper.pdf.

Bartram, L., A. Ho, J. Dill, and F. Henigman. 1995. “The Continuous Zoom: A Constrained Fisheye Technique for Viewing and Navigating Large Information

Spaces.” In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*, 207–15. Pittsburgh, USA: ACM Press.

“Bayerische Landesanstalt Für Landwirtschaft.” 2014. Accessed November 12. http://www.lfl.bayern.de/.

Bennett, Chris, Jody Ryall, Leo Spalteholz, and Amy Gooch. 2007. “The Aesthetics of Graph Visualization.” *Computational Aesthetics in Graphics, Visualization and Imaging*, 1–8.

Bennett, Robert L, K A Steinhaus, S B Uhrich, C K O’Sullivan, R G Resta, D Lochner-Doyle, D S Markel, V Vincent, and J Hamanishi. 1995. “Recommendations for Standardized Human Pedigree Nomenclature. Pedigree Standardization Task Force of the National Society of Genetic Counselors.” *The American Journal of Human Genetics* 56 (3). Division of Medical Genetics, University of Washington Medical Center, Seattle 98195.: 745–52. doi:10.1007/BF01408073.

Bennett, Robin L, Kathryn Steinhaus French, Robert G Resta, and Debra Lochner Doyle. 1995. “Standardized Human Pedigree Nomenclature: Update and Assessment of the Recommendations of the National Society of Genetic Counselors.” *Journal of Genetic Counseling* 17 (3). Springer Netherlands: 241–60.

Bennett, Robin L, Kathryn Steinhaus French, Robert G Resta, and Debra Lochner Doyle.. 2008. “Standardized Human Pedigree Nomenclature: Update and Assessment of the Recommendations of the National Society of Genetic Counselors.” *Journal of Genetic Counseling* 17 (5): 424–33. doi:10.1007/s10897-008-9169-9.

Bertin, Jacques. 1983. *Semiology of Graphics*. University of Wisconsin Press.

Bezerianos, Anastasia, Pierre Dragicevic, Jean-Daniel Fekete, Juhee Bae, and Ben Watson. 2010. “GeneaQuilts: A System for Exploring Large Genealogies.” *IEEE Transactions on Visualization and Computer Graphics* 16 (6): 1073–81. doi:10.1109/TVCG.2010.159.

“Biodiversity International.” 2015. Accessed January 7. http://www.bioversityinternational.org/.

Brewer, Cynthia A., and Mark Harrower. 2014. “ColorBrewer 2 Color Advice for Cartography.” Accessed April 6 2015. www.colorbrewer2.org.

Brewer, Cynthia A., Geoffrey W. Hatchard, and Mark A. Harrower. 2003. “ColorBrewer in Print: A Catalog of Color Schemes for Maps.” *Cartography and Geographic Information Society*. doi:10.1559/152304003100010929.

Brooke, J., P.W. Jordan, B. Thomas, B.A. Weerdmeester, and I.L. McClelland. 1996. “SUS: A Quick and Dirty Usability Scale.” In , 189–94. London, UK: Taylor & Francis.

Brown, Terence A, Martin K Jones, Wayne Powell, and Robin G Allaby. 2009. “The Complex Origins of Domesticated Crops in the Fertile Crescent.” *Trends in Ecology & Evolution* 24 (2): 103–9. doi:10.1016/j.tree.2008.09.008.

Burney, Jennifer A, Steven J Davis, and David B Lobell. 2010. “Greenhouse Gas Mitigation by Agricultural Intensification.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (26): 12052–57.

Buschges, Rainer, Karin Hollricher, Ralph Panstruga, Guus Simons, Marietta Wolter, Adrie Frijters, Raymond Van Daelen, et al. 1997. “The Barley Mlo Gene: A Novel Control Element of Plant Pathogen Resistance.” *Cell* 88 (5): 695–705.

Card, S.K., G.G. Robertson, and J.D. Mackinlay. 1991. “The Information Visualizer, an Information Workspace.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology*, 181–88. New Orleans, Louisiana, USA: ACM Press. doi:10.1145/108844.108874.

Carollo, Victoria, David E Matthews, Gerard R Lazo, Thomas K Blake, David D Hummel, Nancy Lui, David L Hane, and Olin D Anderson. 2005. “GrainGenes 2.0. an Improved Resource for the Small-Grains Community.” *Plant Physiology* 139 (2): 643–51. doi:10.1104/pp.105.064485.

Carver, Tim, Nick Thomson, Alan Bleasby, Matthew Berriman, and Julian Parkhill. 2009. “DNAPlotter: Circular and Linear Interactive Genome Visualization.” *Bioinformatics* 25 (1): 119–20. doi:10.1093/bioinformatics/btn578.

Chi, E.H. 2000. “A Taxonomy of Visualization Techniques Using the Data State Reference Model.” In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, 69–75. Salt Lake City, Utah, USA: IEEE Computer Society Press. doi:10.1109/INFVIS.2000.885092.

Cockburn, A., and J. Savage. 2003. “Comparing Speed-Dependent Automatic Zooming with Traditional Scroll, Pan, and Zoom Methods.” In *BCS HCI*, 87–102. Bath, UK. doi:http://doi.acm.org/10.1145/1456650.1456652.

Cockburn, Andy, A M Y Karlson, and Benjamin B Bederson. 2008. “A Review of Overview + Detail , Zooming , and Focus + Context Interfaces.” *ACM Computing Surveys* 41 (1): 1–42. doi:10.1145/1456650.1456652.

Cockram, James, Jon White, Diana L Zuluaga, David Smith, Jordi Comadran, Malcolm Macaulay, Zewei Luo, et al. 2010. “Genome-Wide Association Mapping to Candidate Polymorphism Resolution in the Unsequenced Barley Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21611–16. doi:10.1073/pnas.1010179107.

Cole, John B. 2007. “PyPedal: A Computer Program for Pedigree Analysis.” *Computers and Electronics in Agriculture* 57 (1): 107–13. doi:10.1016/j.compag.2007.02.002.

Colmer, T. D., R. Munns, and T. J. Flowers. 2005. “Improving Salt Tolerance of Wheat and Barley: Future Prospects.” In *Australian Journal of Experimental Agriculture,* 45:1425–43. doi:10.1071/EA04162.

Comadran, Jordi, Benjamin Kilian, Joanne Russell, Luke Ramsay, Nils Stein, Martin Ganal, Paul Shaw, et al. 2012. “Natural Variation in a Homolog of Antirrhinum CENTRORADIALIS Contributed to Spring Growth Habit and Environmental Adaptation in Cultivated Barley.” *Nature Genetics*, no. 44 (November): 1388–92. doi:10.1038/ng.2447.

Craig, Paul, Alan Cannon, Robert Kukla, and Jessie Kennedy. 2013. “MaTSE: The Gene Expression Time-Series Explorer.” *BMC Bioinformatics* 14 Suppl 1 (Suppl 19): S1. doi:10.1186/1471-2105-14-S19-S1.

“DEFRA Farming and Food Brief.” 2014. *Department for Enviroment & Rural Affairs*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/300303/foodfarmbrief-02apr14.pdf.

Demange, Delphine, Vincent Laporte, Lei Zhao, Suresh Jagannathan, David Pichardie, and Jan Vitek. 2013. “Plan B: A Buffered Memory Model for Java.” *ACM SIGPLAN Notices* 48 (1). ACM: 329–42. doi:10.1145/2480359.2429110.

Dieberger, A., and A.U. Frank. 1998. “A City Metaphor to Support Navigation in Complex Information Spaces.” *Journal of Visual Languages and Computing* 9 (5): 597–622.

“DNA with Ethidium (1978).” 1978. *DNA with Ethidium*. https://www.youtube.com/watch?v=TD0-2lkvfgU.

Doan, Son, Ko-Wei Lin, Mike Conway, Lucila Ohno-Machado, Alex Hsieh, Stephanie Feudjio Feupe, Asher Garland, et al. “PhenDisco: Phenotype Discovery System for the Database of Genotypes and Phenotypes.” *Journal of the American Medical Informatics Association : JAMIA* 21 (1): 31–36. doi:10.1136/amiajnl-2013-001882.

Draper, Geoffrey M. 2008. “Interactive Fan Charts: A Space-Saving Technique for Genealogical Graph Exploration.” *Proceedings of the 8th Annual Workshop on Technology for Family History and Genealogical Research (FHTW 2008).*

“EURISCO.” 2014. http://eurisco.ecpgr.org/.

Faberova, I. 2010. “Standard Descriptors and EURISCO Development.” *Czech J. Genet. Plant Breed* 46: 106–9.

FAO. 2012. “FAO - News Article: Globally Almost 870 Million Chronically Undernourished - New Hunger Report.” http://www.fao.org/news/story/en/item/161819/icode/.

"FAO/Bioversity Multi-Crop Passport Descriptors V.2 [MCPD V.2]." 2015. Accessed January 7 2015. http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v2-mcpd-v2/.

Field, Christopher B. 2014. "IPCC WGII AR5 Summary for Policymakers." *Intergovernmental Panel on Climate Change*. https://ipcc-wg2.gov/AR5/images/uploads/IPCC_WG2AR5_SPM_Approved.pdf.

Fishbeck, G. 2003. "Chapter 3 Diversification through Breeding." In *Diversity in Barley - Hordeum Vulgare*, 29–52. Elsevier.

Foley, Jonathan A, Navin Ramankutty, Kate A Brauman, Emily S Cassidy, James S Gerber, Matt Johnston, Nathaniel D Mueller, et al. 2011. "Solutions for a Cultivated Planet." *Nature* 478 (7369): 337–42. doi:10.1038/nature10452.

Friendly, Michael. 1995. "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization." *13th International Conference on Database and Expert Systems Applications 9* (August): 59–66. doi:10.1016/S1360-1385(01)02193-8.

Fry, Ben. 2007. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. 1st ed. O'Reilly Media.

Fuller, Dorian Q. 2007. "Contrasting Patterns in Crop Domestication and Domestication Rates: Recent Archaeobotanical Insights from the Old World." *Annals of Botany* 100 (5): 903–24. doi:10.1093/aob/mcm048.

Gehlenborg, Nils, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, et al. 2010. "Visualization of Omics Data for Systems Biology." *Nature Methods* 7 (3 Suppl): S56–68. doi:10.1038/nmeth.1436.

"Genesys PGR." 2013. https://www.genesys-pgr.org/welcome.

Gentle, J.E., and Wolfgang Hardle. 2012. *Handbook of Computational Statistics*. 2nd ed. Springer.

"Germinate 3." 2014. *Germinate 3*. Accessed October 28. http://ics.hutton.ac.uk/germinate/.

Gershon, Nahum, Stephen G. Eick, and Stuart Card. 1998. "Information Visualization." *Design Interactions* 5 (2): 9–15. doi:10.1145/274430.274432.

Ghoniem, M., J.-D. Fekete, and P. Castagliola. 2004. "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations," October. Austin, Texas, USA: IEEE Computer Society Press, 17–24.

Gkoutos, Georgios V., Eain C.J. Green, Simon Greenaway, Andrew Blake, Ann-Marie Mallon, and John M. Hancock. 2005. "CRAVE: A Database, Middleware and Visualization System for Phenotype Ontologies." *Bioinformatics* 21 (7): 1257–62. doi:10.1093/bioinformatics/bti147.

Goldenberg, Suzanne. 2014. "Climate Change 'Already Affecting Food Supply' – UN | Environment | The Guardian." *The Guardian*. http://www.theguardian.com/environment/2014/mar/31/climate-change-food-supply-un.

Gould, J.D., and C. Lewis. 1985. "Designing for Usability: Key Principles and What Designers Think." *Communications of the ACM* 28 (3): 300–311.

Graham, Martin, Jessie Kennedy, Trevor Paterson, and Andy Law. 2011. "Visualising Errors in Animal Pedigree Genotype Data." *Computer Graphics Forum* 30 (3): 1011–20. doi:10.1111/j.1467-8659.2011.01950.x.

"Graph - Graph Data Structures and Algorithms - Metacpan.org." 2014. Accessed November 6. https://metacpan.org/pod/distribution/Graph/lib/Graph.pod.

Harlan, J R, and D Zohary. 1966. "Distribution of Wild Wheats and Barley." *Science (New York, N.Y.)* 153 (3740): 1074–80. doi:10.1126/science.153.3740.1074.

Harrower, Mark, and Cynthia A. Brewer. 2003. "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps." *The Cartographic Journal* 40 (1): 27–37. doi:10.1179/000870403235002042.

"HarvEST." 2014. http://harvest.ucr.edu/.

Hayden, Erika. 2009. "Genome Sequencing: The Third Generation." *Nature*, no. 457 (February): 768.

HGCA. 2014. "HGCA : HGCA Recommended Lists." http://www.hgca.com/varieties/hgca-recommended-lists.aspx.

Hoekstra, A. Y., and M. M. Mekonnen. 2012. "The Water Footprint of Humanity." *Proceedings of the National Academy of Sciences* 109 (9): 3232–37. doi:10.1073/pnas.1109936109.

Holten, D. 2006. "Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data." *IEEE Transactions on Visualization and Computer Graphics* 12 (5): 741–48. doi:10.1109/TVCG.2006.147.

Hornbæk, Kasper. 2006. "Current Practice in Measuring Usability: Challenges to Usability Studies and Research." *International Journal of Human Computer Studies* 64 (2): 79–102. doi:10.1016/j.ijhcs.2005.06.002.

"Illumina BeadArray Microarray Technology." 2015. Accessed March 11. http://www.illumina.com/technology/beadarray-technology.html.

Intergovernmental Panel on Climate Change. 2014. "Climate Change, Adaptation, and Vulnerability." *Organization & Environment* 24 (3): 1–44.

IPCC. 2013. “Summary for Policymakers.” In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 33. doi:10.1017/CBO9781107415324.

Jaccoud, D, K Peng, D Feinstein, and A Kilian. 2001. “Diversity Arrays: A Solid State Technology for Sequence Information Independent Genotyping.” *Nucleic Acids Research* 29 (4): E25. doi:10.1093/nar/29.4.e25.

Jakubowska, Joanna, Ela Hunt, Matthew Chalmers, Martin McBride, and Anna F Dominiczak. 2007. “VisGenome: Visualization of Single and Comparative Genome Representations.” *Bioinformatics (Oxford, England)* 23 (19): 2641–42. doi:10.1093/bioinformatics/btm394.

Jing, Runchun, Alexander Vershinin, Jacek Grzebyta, Paul Shaw, Petr Smýkal, David Marshall, Michael J Ambrose, T H Noel Ellis, and Andrew J Flavell. 2010. “The Genetic Diversity and Evolution of Field Pea (Pisum) Studied by High Throughput Retrotransposon Based Insertion Polymorphism (RBIP) Marker Analysis.” *BMC Evolutionary Biology* 10 (January): 44. doi:10.1186/1471-2148-10-44.

Johnson, Chris. 2004. “Top Scientific Visualization Research Problems.” *IEEE Comput. Graph. Appl.* 24 (4): 13–17. doi:10.1109/MCG.2004.20.

Jolly, W, J Froom, and M G Rosen. 1980. “The Genogram.” *The Journal of Family Practice* 10 (2): 251–55.

Jorgensen, I. Helms. 1992. “Discovery, Characterization and Exploitation of Mlo Powdery Mildew Resistance in Barley.” *Euphytica* 63 (1-2): 141–52. doi:10.1007/BF00023919.

Karl, Thomas R, Jerry M Melillo, and Thomas C Peterson. 2009. *Global Climate Change Impacts in the United States*. Edited by Thomas R Karl, Jerry M Melillo, and Thomas C Peterson. *Society*. Vol. 54. U.S. Global Change Research Program. Cambridge University Press.

Keller, Lukas F., and Donald M. Waller. 2002. “Inbreeding Effects in Wild Populations.” *Trends in Ecology and Evolution* 17 (5): 230–41. doi:10.1016/S0169-5347(02)02489-8.

Kestler, Hans A., Andre Muller, Thomas M. Gress, and Malte Buchholz. 2005. “Generalized Venn Diagrams: A New Method of Visualizing Complex Genetic Set Relations.” *Bioinformatics* 21 (8): 1592–95. doi:10.1093/bioinformatics/bti169.

Kilian, Benjamin, Hakan Özkan, Carlo Pozzi, and Francesco Salamini. 2009. “Domestication of the Triticeae in the Fertile Crescent.” In *Genetics and Genomics of the Triticeae*, edited by Gary J Muehlbauer and Catherine Feuillet, 7:81–119. Springer New York. doi:10.1007/978-0-387-77489-3_3.

Kingsbury, Noel. 2009. *Hybrid: The History and Science of Plant Breeding*. University of Chicago Press.

Kirakowski, J, and M Corbett. 1993. “SUMI: The Software Usability Measurement Inventory.” *British Journal of Educational Technology*, no. 3: 10–12. doi:10.1111/j.1467-8535.1993.tb00076.x.

Kopecky, David, Jan Bartos, Adam Lukaszewski, James Baird, Vladimir Cernoch, Roland Kolliker, Odd Arne Rognli, et al. 2009. “Development and Mapping of DArT Markers within the Festuca - Lolium Complex.” *BMC Genomics* 10 (1): 473. doi:10.1186/1471-2164-10-473.

Kreitzberg, Charles B. 1991. “Hypertext Models for Coping with Infomation Overload.” *Interfaces for Information Retrieval and Online Systems*, 169–76.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research* 19 (9): 1639–45. doi:10.1101/gr.092759.109.

Lam, Heidi, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2011. “Empirical Studies in Information Visualization: Seven Scenarios.” *IEEE Transactions on Visualization and Computer Graphics* 18 (9): 1520–36. doi:10.1109/TVCG.2011.279.

Lamacraft, R R, and K W Finlay. 1973. “A Method for Illustrating Pedigrees of Small Grain Varieties for Computer Processing.” *Euphytica* 22 (1). Springer Netherlands: 56–60. doi:10.1007/BF00021556.

Lathe, Warren C., Jennifer M. Williams, and Donna Karolchik. 2008. “Genomic Data Resources: Challenges and Promises.” *Scitable - Nature Education*. http://www.nature.com/scitable/topicpage/Genomic-Data-Resources-Challenges-and-Promises-743721.

Lee, Jennifer M., Guy F. Davenport, David Marshall, T.H. Noel Ellis, Michael J. Ambrose, Jo Dicks, Theo J.L. van Hintum, and Andrew J. Flavell. 2005. “GERMINATE. A Generic Database for Integrating Genotypic and Phenotypic Information for Plant Genetic Resource Collections.” *Plant Physiology* 139 (2): 619–31. doi:10.1104/pp.105.065201.

Leung, Y.K., and M.D. Apperley. 1994. “A Review and Taxonomy of Distortion-Oriented Presentation Techniques.” *ACM Transactions on Human-Computer Interaction* 1 (2): 126–60.

Lewis, James. 2002. “Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies.” *International Journal of Human-Computer Interaction* 14 (3): 463–88. doi:10.1207/S15327590IJHC143&4_11.

Lewis, James R. 1995. "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use." *International Journal of Human-Computer Interaction* 7 (1): 57–78. doi:10.1080/10447319509526110.

Lewis, James R. J.R. 1992. "Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ." In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 36:1259–63. Human Factors and Ergonomics Society.

Li, J-L, M-X Li, H-Y Deng, P E Duffy, and H-W Deng. 2005. "PhD: A Web Database Application for Phenotype Data Management." *Bioinformatics (Oxford, England)* 21 (16): 3443–44. doi:10.1093/bioinformatics/bti557.

Loh, Angeline M, Steven Wiltshire, Jon Emery, Kim W Carter, and Lyle J Palmer. 2008. "Celestial3D: A Novel Method for 3D Visualization of Familial Data." *Bioinformatics* 24 (9): 1210–11.

Mackinlay, Jock. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–41. doi:10.1145/22949.22950.

Mahadevappa, M, R A Descenzo, and R P Wise. 1994. "Recombination of Alleles Conferring Specific Resistance to Powdery Mildew at the Mla Locus in Barley." *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 37 (3): 460–68.

Mäkinen, Ville-Petteri, Maija Parkkonen, Maija Wessman, Per-Henrik Groop, Timo Kanninen, and Kimmo Kaski. 2005. "High-Throughput Pedigree Drawing." *European Journal of Human Genetics: EJHG* 13 (8): 987–89. doi:10.1038/sj.ejhg.5201430.

Mann, Bob, Roy Williams, Malcolm Atkinson, Ken Brodlie, Amos Storkey, and Chris Williams. 2002. *Scientific Data Mining, Integration and Visualization*. National e-Science Centre. https://www.nesc.ac.uk/talks/sdmiv/report.pdf.

Marchese, Francis T. 2011. "Exploring the Origins of Tables for Information Visualization." In *2011 15th International Conference on Information Visualisation*, 395–402. IEEE. doi:10.1109/IV.2011.36.

Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." *Nature* 498 (7453): 255–60. doi:10.1038/498255a.

Matthews, David E, Victoria L Carollo, Gerard R Lazo, and Olin D Anderson. 2003. "GrainGenes, the Genome Database for Small-Grain Crops." *Nucleic Acids Research* 31 (1): 183–86. doi:10.1093/nar/gkg058.

McCormick, B. H. 1988. "Visualization in Scientific Computing." *SIGBIO Newsl.* 10 (1): 15–21. doi:10.1145/43965.43966.

McCormick, B.H., T.A. Defanti, and M.D. Brown. 1987. “Visualization in Scientific Computing.” *IEEE Computer Graphics & Applications* 21 (6): 61–70.

Mertz, Ole, Kirsten Halsnaes, Jørgen E Olesen, and Kjeld Rasmussen. 2009. “Adaptation to Climate Change in Developing Countries.” *Environmental Management* 43 (5): 743–52. doi:10.1007/s00267-008-9259-3.

Meyer, Miriah, Tamara Munzner, and Hanspeter Pfister. 2009. “MizBee: A Multiscale Synteny Browser.” *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 897–904. doi:10.1109/TVCG.2009.167.

Mian, Afaq, Ronald J F J Oomen, Stanislav Isayenkov, Hervé Sentenac, Frans J M Maathuis, and Anne Aliénor Véry. 2011. “Over-Expression of an Na +- and K +-Permeable HKT Transporter in Barley Improves Salt Tolerance.” *Plant Journal* 68 (3): 468–79.

“Microsoft Expression Changes | Microsoft Expression.” 2014. Accessed November 12. http://www.microsoft.com/expression/eng/.

Miedaner, Thomas, and Viktor Korzun. 2012. “Marker-Assisted Selection for Disease Resistance in Wheat and Barley Breeding.” *Phytopathology* 102 (6): 560–66. doi:10.1094/PHYTO-05-11-0157.

Milne, Iain, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. 2009. “Tablet - Next Generation Sequence Assembly Visualization.” *Bioinformatics* 26 (3): 401–2. doi:10.1093/bioinformatics/btp666.

Milne, Iain, Paul Shaw, Gordon Stephen, Micha Bayer, Linda Cardle, William T B Thomas, Andrew J Flavell, and David Marshall. 2010. “Flapjack--Graphical Genotype Visualization.” *Bioinformatics (Oxford, England)* 26 (24): 3133–34. doi:10.1093/bioinformatics/btq580.

Milne, Iain, Gordon Stephen, Micha Bayer, Peter J a Cock, Leighton Pritchard, Linda Cardle, Paul D Shaw, and David Marshall. 2013. “Using Tablet for Visual Exploration of Second-Generation Sequencing Data.” *Briefings in Bioinformatics* 14 (2): 193–202. doi:10.1093/bib/bbs012.

Moore, Gordon. 1965. “Cramming More Components onto Integrated Circuits.” *Electronics Magazine*, April.

Muller, Michael J., Olga Kuchinskaya, Suzanne O. Minassian, John C. Tang, Catalina Danis, Chen Zhao, Beverly Harrison, and Thomas P. Moran. 2005. “Shared Landmarks in Complex Coordination Environments.” In *CHI EA ’05*, 1681–84. ACM. doi:10.1145/1056808.1056996.

Munzner, Tamara. 2009. “A Nested Model for Visualization Design and Validation.” *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 921–28. doi:10.1109/TVCG.2009.111.

“National Statistics.” 2014. *Farming Statistics Provisional Crop Areas, Yields and Livestock Populations At June 2014 - United Kingdom*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/364157/structure-jun2013prov-UK-16oct14.pdf.

Nesbitt, M. Samuel, D. 1995. “Hulled Wheat.” In *Proceedings of the First International Workshop on Hulled Wheats*, 40.

Newbery, F.J. 1989. “Edge Concentration: A Method for Clustering Directed Graphs.” In *SCM ’89 Proceedings of the 2nd International Workshop on Software Configuration Management*, 14:76–85. Princeton, New Jersey, USA: ACM Press. doi:10.1145/72910.73350.

Ni, Junjian, Anuradha Pujar, Ken Youens-Clark, Immanuel Yap, Pankaj Jaiswal, Isaak Tecle, Chih-Wei Tung, et al. 2009. “Gramene QTL Database: Development, Content and Applications.” *Database* 2009 (May): bap005. doi:10.1093/database/bap005.

O’Donoghue, Seán I, Anne-Claude Gavin, Nils Gehlenborg, David S Goodsell, Jean-Karim Hériché, Cydney B Nielsen, Chris North, et al. 2010. “Visualizing Biological Data-Now and in the Future.” *Nature Methods* 7 (3 Suppl): S2–4.

Oxford English Dictionary. 2002. *Paperback Oxford English Dictionary*. *Paperback Oxford English Dictionary*. OUP Oxford.

Paterson, Trevor, Martin Graham, Jessie Kennedy, and Andy Law. 2011. “Evaluating the VIPER Pedigree Visualisation: Detecting Inheritance Inconsistencies in Genotyped Pedigrees.” *2011 IEEE Symposium on Biological Data Visualization BioVis*. IEEE Computer Society Press, 119–26. doi:10.1109/BioVis.2011.6094056.

Paterson, Trevor, Martin Graham, Jessie Kennedy, and Andy Law. 2012. “VIPER: A Visualisation Tool for Exploring Inheritance Inconsistencies in Genotyped Pedigrees.” *BMC Bioinformatics* 13 Suppl 8 (Suppl 8): S5. doi:10.1186/1471-2105-13-S8-S5.

Pavlopoulos, Georgios, Sean O’Donoghue, Venkata Satagopam, Theodoros Soldatos, Evangelos Pafilis, and Reinhard Schneider. 2008. “Arena3D: Visualization of Biological Networks in 3D.” *BMC Systems Biology* 2 (1): 104. doi:10.1186/1752-0509-2-104.

Peterson, C.J. 2011. “Domestication, Breeding, and Genetic Modification of Crop Plants.” http://www.cof.orst.edu/cof/teach/agbio2011/Presentations/Hayes-Breeding_04.05.11.pdf.

Piffanelli, Pietro, Luke Ramsay, Robbie Waugh, Abdellah Benabdelmouna, Angelique D’Hont, Karin Hollricher, Jorgen Helms Jorgensen, Paul Schulze-Lefert, and Ralph Panstruga. 2004. “A Barley Cultivation-Associated Polymorphism Conveys Resistance to Powdery Mildew.” *Nature* 430 (7002): 887–91. doi:10.1038/nature02781.

Pimentel, David, Bonnie Berger, David Filiberto, Michelle Newton, Benjamin Wolfe. Elizabeth Karabinakis, Steven Clark, Elaine Poon, Elizabeth Abbett, Sudha Nandagopal. 2004. "Water Resources: Agricultural and Environmental Issues." *BioScience*. doi:10.1641/0006-3568(2004)054[0909:WRAAEI]2.0.CO;2.

Pritchard, Leighton, Jennifer A. White, Paul R.J. Birch, and Ian K. Toth. 2006. "GenomeDiagram: A Python Package for the Visualization of Large-Scale Genomic Data." *Bioinformatics* 22 (5): 616–17. doi:10.1093/bioinformatics/btk021.

Purchase, H C, R F Cohen, and M James. 1995. "Validating Graph Drawing Aesthetics." In *Graph Drawing*, edited by Franz Brandenburg, 1027:435–46. Lecture Notes in Computer Science. Springer Verlag.

Purchase, H.C. 2000. "Effective Information Visualisation: A Study of Graph Drawing Aesthetics and Algorithms." *Interacting with Computers* 13 (2): 147–62. doi:10.1016/S0953-5438(00)00032-1.

Purchase, H.C. 2002. "Metrics for Graph Drawing Aesthetics." *Journal of Visual Languages and Computing* 13 (5): 501–16. doi:10.1016/S1045-926X(02)90232-6.

Purchase, Helen, David Carrington, and Jo-anne Allder. 2002. "Empirical Evaluation of Aesthetics-Based Graph Layout." *Empirical Software Engineering* 7 (3): 233–55. doi:10.1023/A:1016344215610.

Purdy, Laurence H, W Q Loegering, C F Konzak, C J Peterson, and R E Allan. 1968. "A Proposed Standard Method for Illustrating Pedigrees of Small Grain Varieties." *Crop Science* 8 (4). Crop Science Society of America: 405. doi:10.2135/cropsci1968.0011183X000800040002x.

Pyne, Alice, Ruth Thompson, Carl Leung, Debdulal Roy, and Bart W Hoogenboom. 2014. "Single-Molecule Reconstruction of Oligonucleotide Secondary Structure by Atomic Force Microscopy." *Small (Weinheim an Der Bergstrasse, Germany)*, 1–5.

Rahmstorf, Stefan, Jason E. Box, Georg Feulner, Michael E. Mann, Alexander Robinson, Scott Rutherford, and Erik J. Schaffernicht. 2015. "Exceptional Twentieth-Century Slowdown in Atlantic Ocean Overturning Circulation." *Nature Climate Change* 5 (5): 475–80. doi:10.1038/nclimate2554.

Raun, William R., and Gordon V. Johnson. 1999. "Improving Nitrogen Use Efficiency for Cereal Production." *Agronomy Journal* 91 (3): 357. doi:10.2134/agronj1999.00021962009100030001x.

Rhyne, T.-M. 2003. "Does the Difference between Information and Scientific Visualization Really Matter." *IEEE Computer Graphics and Applications* 23 (3): 6–8. doi:10.1109/MCG.2003.1198256.

Rhyne, T.-M., M. Tory, T. Munzner, M.O. Ward, C. Johnson, and D.H. Laidlaw. 2003. "Information and Scientific Visualization: Separate but Equal or Happy Together at Last?" In , 611–13. Seattle, Washington, USA: IEEE Computer Society Press.

Rosario, G.E., E.A. Rundensteiner, D.C. Brown, and M.O. Ward. 2003. "Mapping Nominal Values to Numbers for Effective Visualization." April. http://davis.wpi.edu/~xmdv/docs/tr0311_nominal.pdf.

Ruths, D.A., E.S. Chen, and L. Ellis. 2000. "Arbor 3D: An Interactive Environment for Examining Phylogenetic and Taxonomic Trees in Multiple Dimensions." *Bioinformatics* 16 (11): 1003–9. doi:10.1093/bioinformatics/16.11.1003.

Sanderson, Michael J. 2006. "Paloverde: An OpenGL 3D Phylogeny Browser." *Bioinformatics* 22 (8): 1004–6. doi:10.1093/bioinformatics/btl044.

Sansom, Claire. 2007. "The DNA Deluge." *Scientific Computing World*. http://www.scientific-computing.com/features/feature.php?feature_id=168.

SASA. 2014. "The Scottish Barley Variety Database." *SASA*. Accessed October 28. http://barley.agricrops.org/menu.php.

Sauro, J. 2011. "Measuring Usability With The System Usability Scale (SUS)." *Measuring Usability*.

Sauro, J. 2004. "Premium Usability: Getting the Discount without Paying the Price." *Interactions* 11 (4): 30–37. doi:10.1145/1005261.1005276.

Sauro, Jeff, and James Lewis. 2012. *Quantifying the User Experience*. Edited by Steve Elliot and Dave Bevans. Waltham, MA: Morgan Kaufmann Publishers Inc.

Savage, David B, Maura Agostini, Inês Barroso, Mark Gurnell, Jian'an Luan, Aline Meirhaeghe, Anne-Helen Harding, et al. 2002. "Digenic Inheritance of Severe Insulin Resistance in a Human Pedigree." *Nature Genetics* 31 (4): 379–84. doi:10.1038/ng926.

"Scotch Whisky Association - Home." 2014. Accessed November 11. http://www.scotch-whisky.org.uk/.

"Scotch Whisky Association 2012 Statistical Report." 2014. Accessed June 3. http://www.scotch-whisky.org.uk/media/62024/2012_statistical_report.pdf.

Scottish Government. 2012. " Whisky Galore – an International Asset ." Scottish Government, St. Andrew's House, Regent Road, Edinburgh EH1 3DG. http://www.scotland.gov.uk/News/Releases/2012/12/whisky31122012.

Sedlmair, Michael, Miriah Meyer, and Tamara Munzner. 2012. "Design Study Methodology : Reflections from the Trenches and the Stacks." *IEEE*

Transactions on Visualization and Computer Graphics 18 (12): 2431–40. doi:10.1109/TVCG.2012.213.

Shneiderman, B. 1996a. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” *Proceedings 1996 IEEE Symposium on Visual Languages*. doi:10.1109/VL.1996.545307.

Shneiderman, B. 1996b. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” In , 336–43. Boulder, Colorado, USA: IEEE Computer Society Press.

Shneiderman, B. 2008. “Extreme Visualization: Squeezing a Billion Records into a Million Pixels.” In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 3–12. Vancouver, Canada: ACM Press. doi:10.1145/1376616.1376618.

Sica, Gregory T. 2006. “Bias in Research Studies.” *Radiology* 238 (3): 780–89. doi:10.1148/radiol.2383041109.

Sorger, Johannes, Katja Buhler, Florian Schulze, Tianxiao Liu, and Barry Dickson. 2013. “neuroMAP — Interactive Graph-Visualization of the Fruit Fly’s Neural Circuit.” In *2013 IEEE Symposium on Biological Data Visualization (BioVis)*, 73–80. IEEE. doi:10.1109/BioVis.2013.6664349.

Stasko, J., R. Catrambone, M. Guzdial, and K. McDonald. 2000. “An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures.” *International Journal of Human-Computer Studies* 53 (5): 663–94. doi:10.1006/ijhc.2000.0420.

Sugiyama, K., S. Tagawa, and M. Toda. 1981. “Methods for Visual Understanding of Hierarchical Systems.” *IEEE Transactions on Systems, Man and Cybernetics* 11 (2): 109–25. doi:10.1109/TSMC.1981.4308636.

“The Scottish Barley Variety Database.” 2014. Accessed November 12. http://www.agricrops.org/menu.php?

Thiele, Holger, and Peter Nürnberg. 2005. “HaploPainter: A Tool for Drawing Pedigrees with Complex Haplotypes.” *Bioinformatics (Oxford, England)* 21 (8): 1730–32. doi:10.1093/bioinformatics/bth488.

Thomas, W.T.B., E. Baird, J.D. Fuller, P. Lawrence, G.R. Young, J. Russell, L. Ramsay, R. Waugh, and W. Powell. 1998. “Identification of a QTL Decreasing Yield in Barley Linked to Mlo Powdery Mildew Resistance.” *Molecular Breeding* 4 (5): 381–93. doi:10.1023/A:1009646115967.

Thomas, J.T., Cook, K.A. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEE Press.

Thorisson, Gudmundur A., Juha Muilu, and Anthony J. Brookes. 2009. "Genotype-Phenotype Databases: Challenges and Solutions for the Post-Genomic Era." *Nat Rev Genet* 10 (1): 9–18. doi:10.1038/nrg2483.

Thormann, Imke, Hannes Gaisberger, Federico Mattei, Laura Snook, and Elizabeth Arnaud. 2012. "Digitization and Online Availability of Original Collecting Mission Data to Improve Data Quality and Enhance the Conservation and Use of Plant Genetic Resources." *Genetic Resources and Crop Evolution* 59 (5): 635–44.

Trager, Edward H, Ritu Khanna, Adrian Marrs, Lawrence Siden, Kari E H Branham, Anand Swaroop, and Julia E Richards. 2007. "Madeline 2.0 PDE: A New Program for Local and Web-Based Pedigree Drawing." *Bioinformatics* 23 (14): 1854–56. doi:10.1093/bioinformatics/btm242.

Treisman, A M, and G Gelade. 1980. "A Feature-Integration Theory of Attention." *Cognitive Psychology* 12 (1): 97–136.

"Trellis - The Chart With Everyone On It." 2014. Accessed November 11. http://progenygenealogy.com/products/family-tree-charts/trellis.aspx.

Tuttle, C, L G Nonato, and C Silva. 2010. *PedVis: A Structured, Space-Efficient Technique for Pedigree Visualization*. *IEEE Transactions on Visualization and Computer Graphics*. Vol. 16. doi:10.1109/TVCG.2010.185.

Tuttle, Claurissa, Luis Gustavo Nonato, and Cláudio T Silva. 2010. "PedVis: A Structured, Space-Efficient Technique for Pedigree Visualization." *IEEE Transactions on Visualization and Computer Graphics* 16 (6): 1063–72. doi:10.1109/TVCG.2010.185.

Unfccc. 2007. *Climate Change: Impacts, Vulnerabilities and Adaptation in Developing Countries*. *United Nations Framework Convention on Climate Change*. United Nations Framework Convention on Climate Change.

Van Berloo, R, and R C B Hutten. 2005. "Peditree: Pedigree Database Analysis and Visualization for Breeding and Science." *The Journal of Heredity* 96 (4): 465–68. doi:10.1093/jhered/esi059.

Van Berloo, Ralph. 2008. "GGT 2.0: Versatile Software for Visualization and Analysis of Genetic Data." *The Journal of Heredity* 99 (2): 232–36. doi:10.1093/jhered/esm109.

Van Hintum, Theo, Frank Menting, and Elisabeth van Strien. 2011. "Quality Indicators for Passport Data in Ex Situ Genebanks." *Plant Genetic Resources* 9 (03): 478–85. doi:10.1017/S1479262111000682.

Van Wijk, J.J., and W.A.A. Nuij. 2003. "Smooth and Efficient Zooming and Panning." In *IEEE Symposium on Information Visualization, INFO VIS (2003)*, 15–22. Seattle, Washington, USA: IEEE Computer Society Press. doi:10.1109/INFVIS.2003.1249004.

Voorrips, Roeland E, Marco C A M Bink, and W Eric van de Weg. 2012. “Pedimap: Software for the Visualization of Genetic and Phenotypic Data in Pedigrees.” *The Journal of Heredity* 103 (6): 903–7. doi:10.1093/jhered/ess060.

“Walrus - Graph Visualization Tool.” n.d. http://www.caida.org/tools/visualization/walrus/.

Wang, Minghui, Ning Jiang, Tianye Jia, Lindsey Leach, James Cockram, Robbie Waugh, Luke Ramsay, Bill Thomas, and Zewei Luo. 2012. “Genome-Wide Association Mapping of Agronomic and Morphologic Traits in Highly Structured Populations of Barley Cultivars.” *TAG. Theoretical and Applied Genetics* 124 (2): 233–46. doi:10.1007/s00122-011-1697-2.

Ware, Colin. 2004. *Information Visualization: Perception for Design*. *Information Visualization*. 1st ed. Waltham, MA: Morgan Kaufmann Publishers Inc. doi:10.1016/B978-0-12-381464-7.00018-1.

Ware, Doreen H, Pankaj Jaiswal, Junjian Ni, Immanuel V Yap, Xioakang Pan, Ken Y Clark, Leonid Teytelman, et al. 2002. “Gramene, a Tool for Grass Genomics.” *Plant Physiology* 130 (4): 1606–13. doi:10.1104/pp.015248.

Wei, Fusheng, Rod A Wing, and Roger P Wise. 2002. “Genome Dynamics and Evolution of the Mla (powdery Mildew) Resistance Locus in Barley.” *The Plant Cell* 14 (8): 1903–17. doi:10.1105/tpc.002238.

Weiss, Ehud, and Daniel Zohary. 2011. “The Neolithic Southwest Asian Founder Crops.” *Current Anthropology* 52 (S4): S237–54. doi:10.1086/658367.

Wernert, E.A., and J. Lakshmipathy. 2005. “PViN - A Scalable and Flexible System for Visualizing Pedigree Databases.” In *Proceedings of the 2005 ACM Symposium on Applied Computing*, 115–22. Santa Fe, New Mexico, USA. doi:10.1145/1066677.1066709.

Wigginton, J. E. 2005. “PEDSTATS: Descriptive Statistics, Graphics and Quality Assessment for Gene Mapping Data.” *Bioinformatics* 21 (16): 3445–47. doi:10.1093/bioinformatics/bti529.

Wilson, EB. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22 (158): 209–12. http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1927.10502953.

Winkler, Harald. 2005. “Climate Change and Developing Countries.” *South African Journal of Science* 101 (July/August): 355–64.

WIPO. 2015. “WIPO World Intellectual Property Organization.” http://www.wipo.int/portal/en/index.html.

Wolfe, J.M., and T.S. Horowitz. 2004. “What Attributes Guide the Deployment of Visual Attention and How Do They Do It?” *Nature Reviews: Neuroscience* 5 (6): 495–501. doi:10.1038/nrn1411.

Wong, Bang. 2012. “Points of View: Visualizing Biological Data.” *Nature Methods* 9 (12). 1131. doi:10.1038/nmeth.2258.

Wong, L. 2000. “Visualization and Manipulation of Pedigree Diagrams.” *Genome Informatics Workshop on Genome Informatics* 11: 63–72.

Wu, Dezhi, Shengguan Cai, Mingxian Chen, Lingzhen Ye, Zhonghua Chen, Haitao Zhang, Fei Dai, Feibo Wu, and Guoping Zhang. 2013. “Tissue Metabolic Responses to Salt Stress in Wild and Cultivated Barley.” *PLoS ONE* 8 (1). doi:10.1371/journal.pone.0055431.

Young, N. D., and S. D. Tanksley. 1989. “Restriction Fragment Length Polymorphism Maps and the Concept of Graphical Genotypes.” *Theoretical and Applied Genetics* 77 (1). Springer-Verlag: 95–101. doi:10.1007/BF00292322.

Zohary, Daniel, Maria Hopf, and Ehud Weiss. 2012. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford University Press.

Appendix 1 DUS data types

DUS ID	Character Name	Data Type	State Identifier	Value
1	Plant Growth Habit	Ordinal	1	erect
1	Plant Growth Habit	Ordinal	2	erect to semi-erect
1	Plant Growth Habit	Ordinal	3	semi-erect
1	Plant Growth Habit	Ordinal	4	semi-erect to intermediate
1	Plant Growth Habit	Ordinal	5	intermediate
1	Plant Growth Habit	Ordinal	6	intermediate to semi-prostrate
1	Plant Growth Habit	Ordinal	7	semi-prostrate
1	Plant Growth Habit	Ordinal	8	semi-prostrate to prostrate
1	Plant Growth Habit	Ordinal	9	prostrate
2	Lower Leaves Hairiness of Leaf Sheaths	Nominal	1	absent
2	Lower Leaves Hairiness of Leaf Sheaths	Nominal	9	present
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	1	absent or very low
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	2	very low to low
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	3	low
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	4	low to medium
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	5	medium
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	6	medium to high
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	7	high
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	8	high to very high
6	Plant Frequency of Plants with Recurned Leaves	Ordinal	9	very high
7	Flagleaf Anthocyanin Colouration of Auricles	Nominal	1	absent
7	Flagleaf Anthocyanin Colouration of Auricles	Nominal	9	present
8	Time of Ear Emergence	Ordinal	1	very early
8	Time of Ear Emergence	Ordinal	2	very early to early
8	Time of Ear Emergence	Ordinal	3	early
8	Time of Ear Emergence	Ordinal	4	early to medium
8	Time of Ear Emergence	Ordinal	5	medium
8	Time of Ear Emergence	Ordinal	6	medium to late
8	Time of Ear Emergence	Ordinal	7	late
8	Time of Ear Emergence	Ordinal	8	late to very late
8	Time of Ear Emergence	Ordinal	9	very late
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	1	absent to very weak
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	2	very weak to weak
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	3	weak
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	4	weak to medium
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	5	medium
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	6	medium to strong
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	7	strong
10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	8	strong to very strong

10	Flagleaf Intensity of Anth Colour of Auricles	Ordinal	9	very strong
14	Ear Glaucosity	Ordinal	1	absent or very weak
14	Ear Glaucosity	Ordinal	2	very weak to weak
14	Ear Glaucosity	Ordinal	3	weak
14	Ear Glaucosity	Ordinal	4	weak to medium
14	Ear Glaucosity	Ordinal	5	medium
14	Ear Glaucosity	Ordinal	6	medium to strong
14	Ear Glaucosity	Ordinal	7	strong
14	Ear Glaucosity	Ordinal	8	strong to very strong
14	Ear Glaucosity	Ordinal	9	very strong
15	Flagleaf Glaucosity of Sheath	Ordinal	1	absent or very weak
15	Flagleaf Glaucosity of Sheath	Ordinal	2	very weak to weak
15	Flagleaf Glaucosity of Sheath	Ordinal	3	weak
15	Flagleaf Glaucosity of Sheath	Ordinal	4	weak to medium
15	Flagleaf Glaucosity of Sheath	Ordinal	5	medium
15	Flagleaf Glaucosity of Sheath	Ordinal	6	medium to strong
15	Flagleaf Glaucosity of Sheath	Ordinal	7	strong
15	Flagleaf Glaucosity of Sheath	Ordinal	8	strong to very strong
15	Flagleaf Glaucosity of Sheath	Ordinal	9	very strong
16	Awn Anthocyanin Colouration of Tips	Nominal	1	absent
16	Awn Anthocyanin Colouration of Tips	Nominal	9	present
17	Ear Attitude	Ordinal	1	erect
17	Ear Attitude	Ordinal	2	erect to semi-erect
17	Ear Attitude	Ordinal	3	semi-erect
17	Ear Attitude	Ordinal	4	semi-erect to horizontal
17	Ear Attitude	Ordinal	5	horizontal
17	Ear Attitude	Ordinal	6	horizontal to semi-recurved
17	Ear Attitude	Ordinal	7	semi-recurved
17	Ear Attitude	Ordinal	8	semi-recurved to recurved
17	Ear Attitude	Ordinal	9	recurved
19	Awn Intensity of Anthocyanin Colour	Ordinal	1	absent to very weak
19	Awn Intensity of Anthocyanin Colour	Ordinal	2	very weak to weak
19	Awn Intensity of Anthocyanin Colour	Ordinal	3	weak
19	Awn Intensity of Anthocyanin Colour	Ordinal	4	weak to medium
19	Awn Intensity of Anthocyanin Colour	Ordinal	5	medium
19	Awn Intensity of Anthocyanin Colour	Ordinal	6	medium to strong
19	Awn Intensity of Anthocyanin Colour	Ordinal	7	strong
19	Awn Intensity of Anthocyanin Colour	Ordinal	8	strong to very strong
19	Awn Intensity of Anthocyanin Colour	Ordinal	9	very strong
20	Sterlie Spikeles Attitude Mid 1 3	Ordinal	1	parallel
20	Sterlie Spikeles Attitude Mid 1 4	Ordinal	2	parallel to weakly divergent
20	Sterlie Spikeles Attitude Mid 1 5	Ordinal	3	divergent

21	Sterile Spikelet Shape of Tip	Nominal	1	pointed
21	Sterile Spikelet Shape of Tip	Nominal	2	rounded
21	Sterile Spikelet Shape of Tip	Nominal	3	squared
22	Ear Development of Sterlie Spikelets	Nominal	1	none or rudimentary(deficiens)
22	Ear Development of Sterlie Spikelets	Nominal	2	full
25	Median Spikelet Length of Glume Awn CF Grain	Ordinal	1	shorter
25	Median Spikelet Length of Glume Awn CF Grain	Ordinal	2	equal
25	Median Spikelet Length of Glume Awn CF Grain	Ordinal	3	longer
29	Ear Length	Ordinal	1	very short
29	Ear Length	Ordinal	2	very short to short
29	Ear Length	Ordinal	3	short
29	Ear Length	Ordinal	4	short to medium
29	Ear Length	Ordinal	5	medium
29	Ear Length	Ordinal	6	medium to long
29	Ear Length	Ordinal	7	long
29	Ear Length	Ordinal	8	long to very long
29	Ear Length	Ordinal	9	very long
30	Ear Length	Ordinal	3	short (shorter than ear)
30	Ear Length	Ordinal	4	shorter to +/- equal
30	Ear Length	Ordinal	5	medium (+/- equal to ear)
30	Ear Length	Ordinal	6	+/- equal to longer
30	Ear Length	Ordinal	7	long (longer than ear)
33	Plant Length Stem Ear Awns	Ordinal	1	very short
33	Plant Length Stem Ear Awns	Ordinal	2	very short to short
33	Plant Length Stem Ear Awns	Ordinal	3	short
33	Plant Length Stem Ear Awns	Ordinal	4	short to medium
33	Plant Length Stem Ear Awns	Ordinal	5	medium
33	Plant Length Stem Ear Awns	Ordinal	6	medium to long
33	Plant Length Stem Ear Awns	Ordinal	7	long
33	Plant Length Stem Ear Awns	Ordinal	8	long to very long
33	Plant Length Stem Ear Awns	Ordinal	9	very long
38	Collar Type	Ordinal	1	decurrent
38	Collar Type	Ordinal	2	decurrent to platform
38	Collar Type	Ordinal	3	platform
38	Collar Type	Ordinal	4	platform to shallow cup
38	Collar Type	Ordinal	5	shallow cup
38	Collar Type	Ordinal	6	shallow cup to cup
38	Collar Type	Ordinal	7	cup
39	Ear Number of Rows	Nominal	1	two
39	Ear Number of Rows	Nominal	2	six
40	Ear Density	Ordinal	1	very lax
40	Ear Density	Ordinal	2	very lax to lax

40	Ear Density	Ordinal	3	lax
40	Ear Density	Ordinal	4	lax to medium
40	Ear Density	Ordinal	5	medium
40	Ear Density	Ordinal	6	medium to dense
40	Ear Density	Ordinal	7	dense
40	Ear Density	Ordinal	8	dense to very dense
40	Ear Density	Ordinal	9	very dense
41	Ear Shape	Ordinal	3	tapering
41	Ear Shape	Ordinal	4	tapering to parallel
41	Ear Shape	Ordinal	5	parallel
41	Ear Shape	Ordinal	7	fusiform
44	Rachis Length of First Segment	Ordinal	3	short
44	Rachis Length of First Segment	Ordinal	4	short to medium
44	Rachis Length of First Segment	Ordinal	5	medium
44	Rachis Length of First Segment	Ordinal	6	medium to long
44	Rachis Length of First Segment	Ordinal	7	long
46	Rachis Curviture of First Segment	Ordinal	1	absent
46	Rachis Curviture of First Segment	Ordinal	2	very weak
46	Rachis Curviture of First Segment	Ordinal	3	weak
46	Rachis Curviture of First Segment	Ordinal	4	weak to medium
46	Rachis Curviture of First Segment	Ordinal	5	medium
46	Rachis Curviture of First Segment	Ordinal	6	medium to strong
46	Rachis Curviture of First Segment	Ordinal	7	strong
46	Rachis Curviture of First Segment	Ordinal	8	strong to very strong
46	Rachis Curviture of First Segment	Ordinal	9	very strong (angular)
62	Kernel Colour of Aleuron Layer	Ordinal	1	whitish (white)
62	Kernel Colour of Aleuron Layer	Ordinal	2	weakly coloured
62	Kernel Colour of Aleuron Layer	Ordinal	3	strongly coloured (blue)
65	Grain Rachilla Hair Type	Nominal	1	short
65	Grain Rachilla Hair Type	Nominal	2	long
69	Awn Spiculation of Margins	Ordinal	1	absent
69	Awn Spiculation of Margins	Ordinal	5	reduced
69	Awn Spiculation of Margins	Ordinal	9	present
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	1	absent/v. weak (0-2 per nerve)
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	2	v. weak to weak
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	3	weak (1-2 per nerve)
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	4	weak to medium
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	5	medium (3-5 per nerve)
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	6	medium to strong
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	7	strong (5-10 per nerve)
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	8	strong to v. strong
73	Grain Spiculation of Inner Lateral Nerves	Ordinal	9	very strong (>10 per nerve)

75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	1	absent or very weak
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	2	very weak to weak
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	3	weak
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	4	weak to medium
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	5	medium
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	6	medium to strong
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	7	strong
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	8	strong to very strong
75	Grain Anthocyanin Colouration of Lemma Nerves	Ordinal	9	very strong
76	Grain Ventral Furrow Presence of Hairs	Nominal	1	absent
76	Grain Ventral Furrow Presence of Hairs	Nominal	5	VFH/sharkskin-;fence hairs+
76	Grain Ventral Furrow Presence of Hairs	Nominal	9	present
86	Grain Husk	Nominal	1	absent (grains thresh free)
86	Grain Husk	Nominal	9	present
88	Grain Disposition of Lodicules	Nominal	1	frontal (bib type)
88	Grain Disposition of Lodicules	Nominal	2	clasping (collar type)
90	Seasonal Type	Nominal	1	winter type
90	Seasonal Type	Nominal	2	alternative type
90	Seasonal Type	Nominal	3	spring type

Appendix 2 Initial requirements gathering questionnaire

Purpose

The purpose of this questionnaire is to try and determine possible tools that we can add to the Germinate 3 database and interface that will help to make your research easier! You have been asked to contribute to this as you are seen as someone who works in the genetic resources or quantitative genetics area and would be able to give us valuable feedback.

We are trying to get a consensus on what tools scientists like you need and would use in order to best direct our efforts in the development of Germinate 3 in relation to the storage and analysis of QTL based data.

The results will form part of a PhD however your identity details **will not be disclosed**.

If you have not used Germinate 3 you can see it in action at the following links,

http://bioinf.scri.ac.uk/germinate_pea
http://bioinf.scri.ac.uk/germinate_cpc
http://bioinf.scri.ac.uk/germinate_grasses

Many thanks in advance for agreeing to fill out this questionnaire. If you have any queries please do not hesitate to contact me on paul.shaw@scri.ac.uk or by phone +44 (0)1382 562731 ext. 2638.

Paul Shaw, Genetics Programme, Plant Bioinformatics Group, SCRI, Dundee, DD2 5DA, Scotland.

Ok enough of that, let's get started.

Name:		**Email:**	
Organisation:		**Date:**	

1. Background

A. What would you describe your main area of work and what species do you work with?

B. Do you work in…?

	A University or Institute
	A Private Company

C. How many plant lines do you routinely deal with?

	Up to 100
	More than 100 but less than 1,000
	More than 1,000 but less than 10,000
	More than 10,000

If it's not confidential feel free to indicate more precise numbers here:

D. How many markers do you routinely deal with?

	Up to 100
	More than 100 but less than 1,000
	More than 1,000 but less than 10,000
	More than 10,000

If it's not confidential feel free to indicate more precise numbers here:

E. How many phenotypic scores do you have across all your data?

	Up to 1,000
	More than 10,000 but less than 100,000
	More than 100,000 but less than 1 million
	More than 1 million

If it's not confidential feel free to indicate more precise numbers here:

F. Tick the following that apply to you. Tick as many as you like we won't tell!

	I don't mind registering to use useful tools.
	I would rather use a desktop application than a web-based tool.
	I would actively participate in a community orientated knowledgebase should such features exist in an online application.
	I think it's important to get as much of my data as possible available to the wider community however some of my data needs to remain private for a period of time.
	I think that the visual look and feel of a web site is important.
	I would be put off using a web resource that looked old and out of date.

	The ability to export data in particular import formats for analysis tools is important to me.
	I would rather only see raw data and no data visualizations.
	Data visualizations are important to me.
	Any database tools need to be installed on my desktop machine; I have no access to Linux based database / web servers.
	I want to get my data online but I don’t have the in house capability. I would be interesting in outsourcing this task to someone with these skills.

2. What Questions Need Answered

Give details of any specific questions relating to the area of QTL analysis that you would want to be able to ask of your genotypic and phenotypic data.

Examples of questions that you may want to ask may include “I want to see all markers in a defined region that are polymorphic” or “I want to identify potential candidate markers based on a defined statistical analysis of my data”.

A. First Question

B. Second Question

C. Third Question

3. **Functionality**

What features would you want to see in relation to QTL analysis that we could perform within Germinate 3?

Examples of features may include the ability to generate genome scans from your data or the ability to multiple analyses using different methods. Others may include integration with statistical packages such as R, Genstat or Minitab or the ability to generate interactive visualizations of your data.

Please be as specific as you can.

D. Most important feature

E. Second most important feature

F. Third most important feature

4. **Anything Else?**

If there is anything else you need to get off your chest, good or bad, we need your feedback so go ahead and enter it here.

Thanks for taking the time to complete this questionnaire; your feedback is much appreciated!

Paul

Appendix 3 Edinburgh Napier Informed Consent Form

Informed Consent Form

Visualizing Genetic Transmission Patterns in Plant Pedigrees

Edinburgh Napier University requires that all persons who participate in research studies give their written consent to do so. Please read the following and sign it if you agree with what it says.

1. I freely and voluntarily consent to be a participant in the research project on the topic of plant pedigree visualization techniques to be conducted by Paul Shaw, who is a PhD student in the Edinburgh Napier School of Computing.

2. The broad goal of this research study is to explore the use of visualization techniques to further enhance the use of pedigrees in experimental plant genetics. Specifically, I have been asked to perform a series of short tasks and answer a questionnaire which should take no longer than 45 minutes to complete.

3. I have been told that my responses will be anonymised. My name will not be linked with the research materials, and I will not be identified or identifiable in any report subsequently produced by the researcher.

4. I also understand that if at any time during the study I feel unable or unwilling to continue, I am free to leave. That is, my participation in this study is completely voluntary, and I may withdraw from it at any time without negative consequences.

5. In addition, should I not wish to answer any particular question or questions, I am free to decline.

6. I have been given the opportunity to ask questions regarding the study and my questions have been answered to my satisfaction.

7. I have read and understand the above and consent to participate in this study. My signature is not a waiver of any legal rights. Furthermore, I understand that I will be able to keep a copy of the informed consent form for my records.

Participant’s Signature Date

I have explained and defined in detail the research procedure in which the respondent has consented to participate. Furthermore, I will retain one copy of the informed consent form for my records.

Researcher’s Signature Date

Appendix 4 Initial user testing questionnaire

Pedigree Visualization Study

The aim of this study is to get end-user feedback on the use of a new software tool for visualizing large plant pedigrees. We are specifically interested in users' cognition of our visualization abstraction of plant lines as nodes modelled as a directed acyclic graph and if this brings advantages over currently used techniques.

As part of this study you will be asked to answer a number of questions and perform a series of tasks using our simple prototype pedigree viewer. There are no right and wrong answers so please answer as truthfully as possible!

Part 1 Pre-Study Questionnaire

This short questionnaire is just to get a bit of background about your experience and work in this area. The answers you give here, like every question in this survey, are anonymous.

1. What is your job title? E.g. Geneticist, Plant Breeder, Student, Bioinfomatacist.

2. What qualifications and length of experience do you have in your field?

3. Do you use pedigree data in the course of your work?

4. Do you use any tools in the course of your work to handle or visualize pedigree data?

Part 2 – Pedigree and Interface Components

As part of this study we are using Microsoft's Expression Studio to record your interaction with the visualization tool, this will be used in subsequent data analysis.

1. In your opinion, what do white graph nodes represent in this display?

2. What are the parents and grandparents for "Ayr"

3. What are the children of line "Sebastian"?

4. What are the grandparents of the single progeny of "Oxbridge"

5. How many divisions are there for phenotype "Plant Growth Habit"?

6. What are the great-grandparents of the single progeny of “Oxbridge”

7. What is the breeders code and AFP number for “Hart”

8. In phenotype “Time of Ear Emergence” what is the value for lines “Scarlett” and “Vegas”

Part 3 - Post Study Questionnaire Part A

Statements	Very difficult	difficult	Neutral	Easy	Very easy
1. Finding parents of a specified line.	□	□	□	□	□
2. Distinguishing phenotype classes.	□	□	□	□	□
3. Tracing a lines lineage.	□	□	□	□	□

4. Understanding information being conveyed.	□	□	□	□	□
5. Relating colour coding to phenotypic values.	□	□	□	□	□
6. Finding progeny of a specified line.	□	□	□	□	□
7. Accessing background information on plant lines.	□	□	□	□	□
8. Clarity of relationships between nodes/lines.	□	□	□	□	□
9. Finding specific lines.	□	□	□	□	□
10. Maintaining your bearings in the visualization.	□	□	□	□	□
11. Navigation round the visualization (zooming and panning)	□	□	□	□	□
12. Overall ease of use.	□	□	□	□	□

Post Study Questionnaire Part B

You should have now filled in the pre-study questionnaire and carried out the predefined tasks using the Helium user interface. Having had a few minutes to think about your experiences using this tool; please answer the following series of questions. Feel free to expand on your answers here using a separate sheet if necessary.

1. Was there anything in the Helium interface that you found to be confusing?

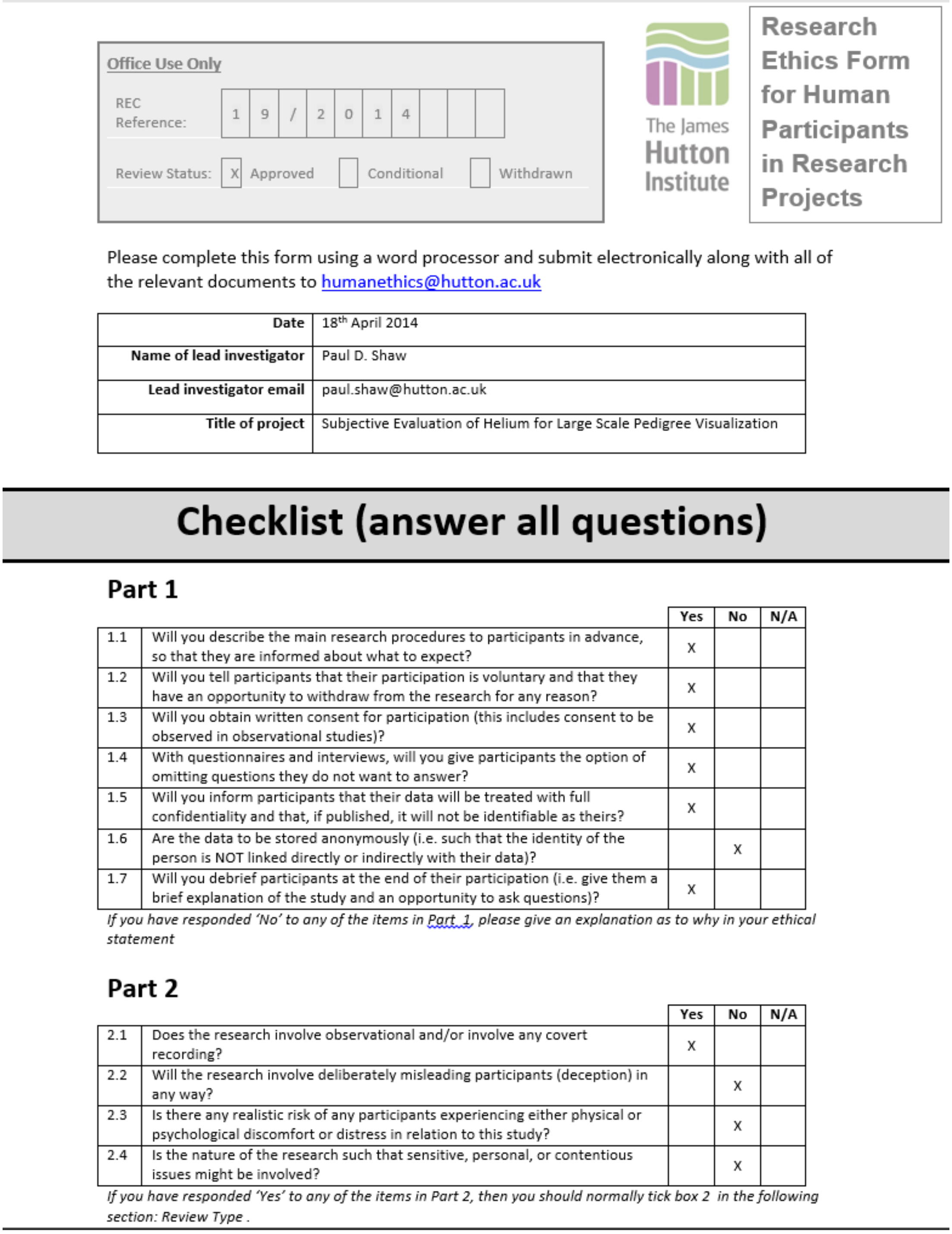
2. Was there anything in the Helium interface that you found to be particularly clear to understand?

3. Are there any other features that you would like to see that would make this tool more useful for your work?

4. Are there any other general comments that you would make about this interface?

Thank you for your time in helping with this evaluation!

Appendix 5 James Hutton Institute Human Ethics Committee Application



Office Use Only

REC Reference:	1	9	/	2	0	1	4			

Review Status: ☒ Approved □ Conditional □ Withdrawn

The James Hutton Institute

Research Ethics Form for Human Participants in Research Projects

Please complete this form using a word processor and submit electronically along with all of the relevant documents to humanethics@hutton.ac.uk

Date	18th April 2014
Name of lead investigator	Paul D. Shaw
Lead investigator email	paul.shaw@hutton.ac.uk
Title of project	Subjective Evaluation of Helium for Large Scale Pedigree Visualization

Checklist (answer all questions)

Part 1

		Yes	**No**	**N/A**
1.1	Will you describe the main research procedures to participants in advance, so that they are informed about what to expect?	X		
1.2	Will you tell participants that their participation is voluntary and that they have an opportunity to withdraw from the research for any reason?	X		
1.3	Will you obtain written consent for participation (this includes consent to be observed in observational studies)?	X		
1.4	With questionnaires and interviews, will you give participants the option of omitting questions they do not want to answer?	X		
1.5	Will you inform participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	X		
1.6	Are the data to be stored anonymously (i.e. such that the identity of the person is NOT linked directly or indirectly with their data)?		X	
1.7	Will you debrief participants at the end of their participation (i.e. give them a brief explanation of the study and an opportunity to ask questions)?	X		

If you have responded 'No' to any of the items in Part 1, please give an explanation as to why in your ethical statement

Part 2

		Yes	**No**	**N/A**
2.1	Does the research involve observational and/or involve any covert recording?	X		
2.2	Will the research involve deliberately misleading participants (deception) in any way?		X	
2.3	Is there any realistic risk of any participants experiencing either physical or psychological discomfort or distress in relation to this study?		X	
2.4	Is the nature of the research such that sensitive, personal, or contentious issues might be involved?		X	

If you have responded 'Yes' to any of the items in Part 2, then you should normally tick box 2 in the following section: Review Type .

Part 3

		Yes	No	N/A
3.1	Are there any other ethical issues within the proposed research that you are aware of? (for example, potential conflicts of interest; the use of artefacts; environmental impact)		X	
3.2	Does your project involve people under 18 years of age?		X	
3.3	Does your project involve people with learning or communication difficulties?		X	
3.4	Is your project likely to involve people involved in illegal activities?		X	
3.5	Does your project involve people belonging to a vulnerable group, other than those listed above?		X	
3.6	Does your project involve patients and/or hospital staff ?		X	
3.7	Does your project require access to personal information about participants from other parties (e.g. teachers, employers), databanks or files?		X	
3.8	Do you plan to conceal your own identity during the course of the research?		X	

If you have responded 'Yes' to any of the items in Part 3, then you should normally tick box 2 in Review Type.

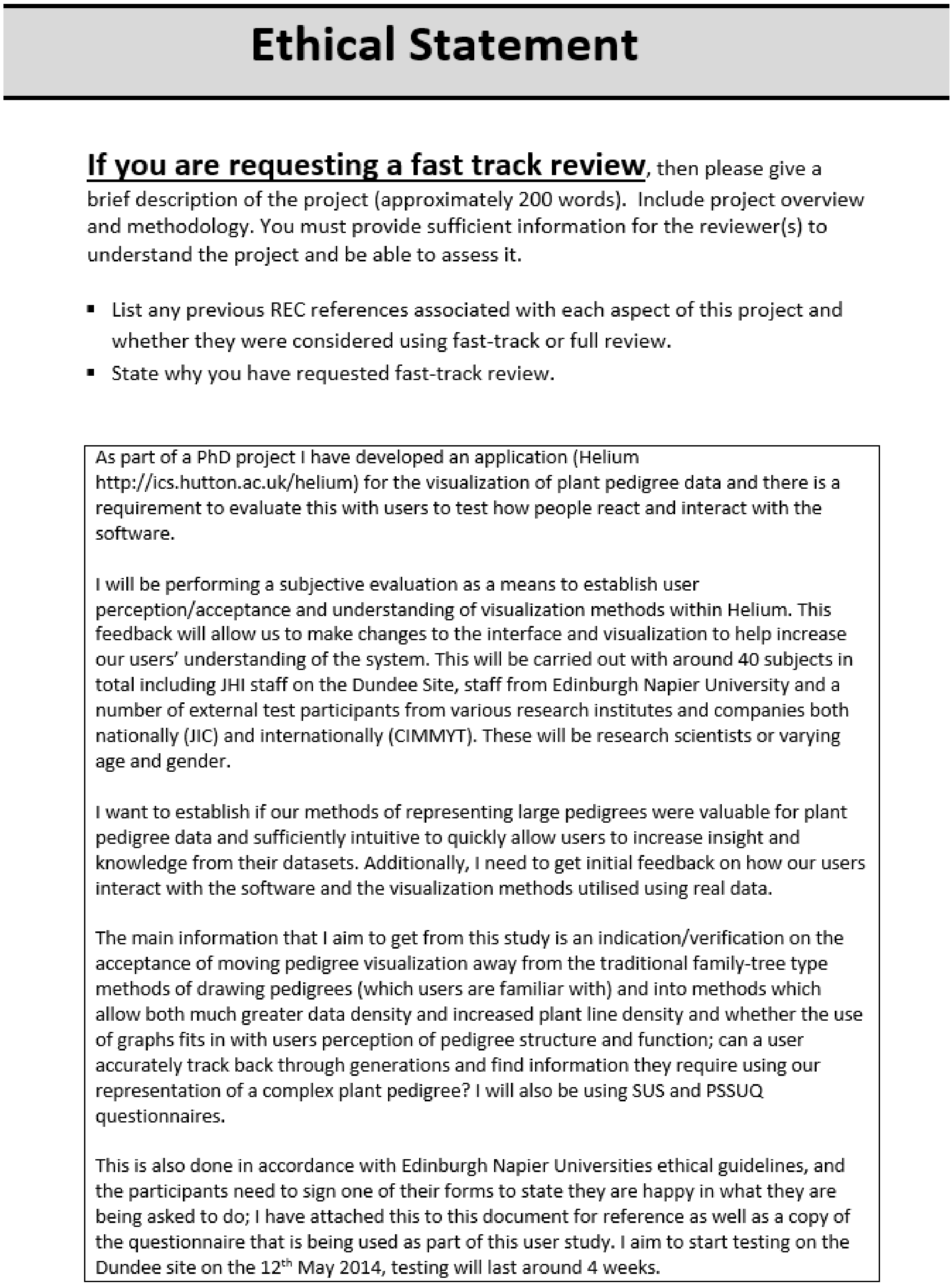
There is an obligation on the lead investigator to bring to the attention of the Research Ethics Committee (REC) any issues with ethical implications not clearly covered by the above checklist.

Review Type (tick 1 box only)

Please tick either box 1 or 2 below and provide the required details in support of your application.

Please tick

1	I consider that this project has no significant ethical implications to be brought before the Research Ethics Committee and request a fast-track review on the basis of the information provided in the ethical statement.	X
2	I consider that this project may or does have significant ethical implications that should be given a full review by the Research Ethics Committee, and/or it will be carried out with children or other vulnerable populations.	



Ethical Statement

<u>If you are requesting a fast track review</u>, then please give a brief description of the project (approximately 200 words). Include project overview and methodology. You must provide sufficient information for the reviewer(s) to understand the project and be able to assess it.

- List any previous REC references associated with each aspect of this project and whether they were considered using fast-track or full review.
- State why you have requested fast-track review.

As part of a PhD project I have developed an application (Helium http://ics.hutton.ac.uk/helium) for the visualization of plant pedigree data and there is a requirement to evaluate this with users to test how people react and interact with the software.

I will be performing a subjective evaluation as a means to establish user perception/acceptance and understanding of visualization methods within Helium. This feedback will allow us to make changes to the interface and visualization to help increase our users' understanding of the system. This will be carried out with around 40 subjects in total including JHI staff on the Dundee Site, staff from Edinburgh Napier University and a number of external test participants from various research institutes and companies both nationally (JIC) and internationally (CIMMYT). These will be research scientists or varying age and gender.

I want to establish if our methods of representing large pedigrees were valuable for plant pedigree data and sufficiently intuitive to quickly allow users to increase insight and knowledge from their datasets. Additionally, I need to get initial feedback on how our users interact with the software and the visualization methods utilised using real data.

The main information that I aim to get from this study is an indication/verification on the acceptance of moving pedigree visualization away from the traditional family-tree type methods of drawing pedigrees (which users are familiar with) and into methods which allow both much greater data density and increased plant line density and whether the use of graphs fits in with users perception of pedigree structure and function; can a user accurately track back through generations and find information they require using our representation of a complex plant pedigree? I will also be using SUS and PSSUQ questionnaires.

This is also done in accordance with Edinburgh Napier Universities ethical guidelines, and the participants need to sign one of their forms to state they are happy in what they are being asked to do; I have attached this to this document for reference as well as a copy of the questionnaire that is being used as part of this user study. I aim to start testing on the Dundee site on the 12th May 2014, testing will last around 4 weeks.

If you are requesting a full review, then please provide responses to the points listed below in a separate document.

- Project title.
- Purpose of the project and its academic rationale.
- Brief description of methods and measurement procedure(s).
- Participants: recruitment methods; number; age; gender; exclusion/inclusion criteria.
- Recruitment, information, consent arrangements, debriefing.
- Please attach copies of all intended sheets/forms and associated correspondence.
- A clear and concise statement of the ethical considerations raised by the project and how you intend to deal with them.
- Estimated start date and duration of the project.

You must provide sufficient information for the reviewer(s) to understand the research protocol.

Information and Signature

Have you prepared and included all the necessary documents for submission?

	Yes	No	N/A
REC approval form fully completed	X		
Consent form	X		
Debrief form/information			X
Recruitment materials			X
Information sheet(s)			X
Instructions for participants	X		

If you have ticked 'no' in any of the boxes, it is very likely that your application is not yet complete so you should not submit it until all the relevant documentation is prepared and all boxes are ticked as either 'yes' or 'not applicable'.

(LEAD INVESTIGATOR to sign)	Name (please enter)	Date (please enter)
I confirm that I have familiarised myself with the appropriate ethical guidelines, and have used this form to bring any ethical considerations in this research project to the attention of the REC.	Paul Shaw.	18/04/2014

Appendix 6 Second round user testing questionnaire

Helium User Evaluation

The aim of this study is to get end-user feedback on the use of a new software tool for visualizing large plant pedigrees. This is not a test of you, but of how people interact and use the software. The feedback gained from this testing will be used to improve the visualization interface we have developed.

We are specifically interested in users' cognition of our visualization abstraction of plant lines as nodes modelled as a directed acyclic graph and inclusion of varying data types and if this brings advantages over currently used techniques and alternative pedigree tools.

As part of this study you will be asked to answer a number of questions and perform a series of tasks using the Helium pedigree viewer.

All answers are anonymous and you are encouraged to ask questions as we go through the testing.

Part 1 Pre-Study Questionnaire

This short questionnaire is to collect some background information about your experience and work in this area.

5. ***Did you participate in the 1st round of evaluation for this tool in April 2013?*** **YES / NO**

6. ***What is your job title?***
    a. Geneticist
    b. Plant Breeder
    c. Bioinformatician

   d. Student
   e. Other (please specify)

7. ***What is your highest level of academic qualification?***
   a. PhD
   b. MSc
   c. BSc
   d. Other (please specify)

8. ***What length of experience do you have in your field?*** **Years Months**

9. ***Do you use pedigree data in the course of your work?***
   YES / NO

10. ***If you answered YES above, roughly how often do you use pedigree data?***
    a. Every day.
    b. Every week.
    c. Every month.
    d. Every year.

11. ***Do you consider the current level of lineage errors in recorded pedigree data to be a cause for concern?***
    YES / NO

12. ***Do you use any tools in the course of your work to handle or visualize pedigree data? If so please enter them here and indicate how frequently you use them.***

Part 2 Pedigree and Interface Components

As part of this study we are using Microsoft's Expression Studio to record your interaction with the visualization tool, this will be used in subsequent data analysis but will not be available to anyone outside of this study.

9.3 Section A

1. What are the parents and grandparents for the line "Ayr"?

2. What are the children of the line "Sebastian"?

3. What are the great-great grandparents for the line "Agenda"?

4. What are the grandparents of the progeny of "Oxbridge"?

5. What are the "ear_shape" characters of the progeny of "Oxbridge"?

6. What are the a) most and b) least abundant divisions for the phenotype "plant_growth_habit"?

    a. Most abundant

    b. b. Least abundant

7. What is the breeder's code and AFP numbers for the line "Hart"?

AFP **Breeders Code**

8. Using the phenotype “Time of Ear Emergence” what are the values for the lines “Scarlett” and “Vegas”?

Scarlett **Vegas**

9.3.1 *Section A Feedback*

Question	**Description**	**Very Difficult**	**Difficult**	**Neutral**	**Easy**	**Very Easy**
1	Finding parents of a line.	○	○	○	○	○
2	Tracing a lines lineage.	○	○	○	○	○
3	Relating colour coding to phenotypic values.	○	○	○	○	○
4	Finding progeny of a specified line.	○	○	○	○	○
5	Clarity of relationships between lines.	○	○	○	○	○
6	Finding specific lines.	○	○	○	○	○

9.4 **Section B**

9. Which line which has category “medium” for phenotype “ear_length” has been used to derive most subsequent varieties? And how many grandchildren of this line have missing data that are not intermediate crosses.

10. Using the line “chariot” as the base line for genetic similarity calculations how many lines are within 90% similarity to ”chariot”?

11. Using the results from question 9 above are there any obvious errors within the similarity data?

9.4.1 *Section B Feedback – Standard SUS (System USability Scale) Questions*

Strongly

agree Strongly disagree

Question	Description	1	2	3	4	5
1	I think that I would like to use this system.	○	○	○	○	○
2	I found the system unnecessarily complex.	○	○	○	○	○
3	I thought the system was easy to use.	○	○	○	○	○
4	I think that I would need the support of a technical person to be able to use this system.	○	○	○	○	○
5	I found the various functions in the system were well integrated.	○	○	○	○	○
6	I thought there was too much inconsistency in this system.	○	○	○	○	○
7	I would imagine that most people would learn to use this system very quickly	○	○	○	○	○
8	I found the system very cumbersome to use.	○	○	○	○	○
9	I felt very confident using the system.	○	○	○	○	○
10	I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

Part 3 Post Study Questionnaire

You should have now filled in the pre-study questionnaire and carried out the predefined tasks using the Helium. Having had a few minutes to think about your experiences using this tool; please answer the following series of questions. Feel free to expand on your answers here using a separate sheet if necessary.

9.5 Section C

5. Was there anything in the Helium interface that you found to be confusing?

6. Was there anything in the Helium interface that you found to be particularly clear to understand?

7. Are there any other questions that you would be able to ask using Helium that you have not already tried?

8. Are there any other general comments that you would make about this interface?

Finally, having reflected on your use of Helium and filled in any general comments how do you feel about your experience in using this pedigree visualization tool?

Section D - Standard PSSUQ (Post STUDY SYSTEM USABILITY) QUESTIONS

Strongly agree

Strongly disagree

Question	Description	1	2	3	4	5	6	7	NA
1	Overall I am satisfied with how easy it is to use this system.	○	○	○	○	○	○	○	○
2	It was simple to use this system.	○	○	○	○	○	○	○	○
3	I could effectively complete the tasks and scenarios using this system.	○	○	○	○	○	○	○	○
4	I was able to complete the tasks and scenarios quickly using this system.	○	○	○	○	○	○	○	○
5	I was able to efficiently complete the tasks and scenarios using this system.	○	○	○	○	○	○	○	○
6	I felt comfortable using this system.	○	○	○	○	○	○	○	○
7	It was easy to learn to use this system.	○	○	○	○	○	○	○	○
8	I believe I could become productive quickly using this system.	○	○	○	○	○	○	○	○
9	The system gave error messages that clearly told me how to fix problems.	○	○	○	○	○	○	○	○
10	Whenever I made a mistake using the system, I could recover easily and quickly.	○	○	○	○	○	○	○	○
11	The information (such as on-line help, on screen messages and other documentation) provided with this system was clear.	○	○	○	○	○	○	○	○
12	It was easy to find the information I needed.	○	○	○	○	○	○	○	○
13	The information provided for the system was easy to understand.	○	○	○	○	○	○	○	○

14	The information was effective in helping me complete the tasks and scenarios.	○	○	○	○	○	○	○	○
15	The organisation of information on the system screens was clear.	○	○	○	○	○	○	○	○
16	The interface of this system was pleasant.	○	○	○	○	○	○	○	○
17	I liked using the interface of this system.	○	○	○	○	○	○	○	○
18	This system has all the functions and capabilities I expect it to have.	○	○	○	○	○	○	○	○
19	Overall, I am satisfied with this system.	○	○	○	○	○	○	○	○

Thank you for your time in helping with this evaluation!

Appendix 7 Second round user testing raw data Part 1

User	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	yes	geneticist	phd	34	yes	month	yes
2	yes	bioinformatician	msc	2	yes	month	yes
3	no	breeder	phd	6	yes	daily	yes
4	no	breeder	phd	15	yes	weekly	yes
5	no	genebank manager	phd	25	yes	month	no
6	no	research manager	phd	20	no	NA	yes
7	no	geneticist	phd	2	yes	month	yes
8	no	geneticist	phd	10	yes	month	no
9	no	breeder	phd	30	yes	month	no
10	no	geneticist	phd	25	yes	month	yes
11	no	geneticist	phd	13	yes	month	yes
12	no	bioinformatician	msc	10	no	NA	yes
13	yes	statistician	phd	6	no	NA	yes
14	yes	geneticist	phd	30	yes	month	yes
15	no	geneticist	phd	15	yes	month	yes
16	no	breeder	phd	18	yes	month	yes
17	yes	geneticist	phd	20	yes	month	yes
18	no	breeder	phd	18	yes	weekly	yes
19	no	breeder	bsc	7	yes	month	yes
20	no	cytogeneticist	phd	10	no	NA	NA
21	no	bioinformatician	phd	12	yes	month	yes
22	yes	genetisist	phd	8	yes	weekly	yes
23	no	bioinformatician	phd	5	yes	month	yes
24	yes	genetisist	phd	10	yes	month	no
25	no	genetisist	phd	32	yes	month	no
26	no	genetisist	phd	4	yes	month	yes
27	yes	genetisist	phd	12	yes	month	yes
28	no	genetisist	phd	22	yes	month	no

Appendix 8 Second round user testing raw data Part 2 Section A

Pedigree and interface components. 1 represents correct response, 0 incorrect response to question.

User	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1	1	1	0	1	1	1	1	1
2	1	1	1	1	0	1	1	1
3	1	1	0	0	1	1	1	0
4	1	1	1	1	1	1	1	1
5	1	1	1	0	1	1	1	1
6	1	1	1	0	1	1	1	1
7	0	0	0	0	0	1	1	1
8	0	0	1	1	1	1	1	1
9	1	1	0	1	1	1	1	1
10	0	0	1	1	1	1	1	1
11	1	1	0	0	1	1	1	1
12	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1
14	0	1	1	1	1	1	1	1
15	1	1	0	1	1	1	1	0
16	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	0	1
20	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1
22	1	1	1	0	1	1	1	0
23	1	1	1	1	1	1	1	1
24	1	1	0	0	1	1	1	0
25	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	0
27	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1
TOTAL /28	24	25	21	21	26	28	27	23
AVERAGE	0.86	0.89	0.75	0.75	0.93	1.00	0.96	0.82

Appendix 9 Second round user testing raw data Part 2 Section A Likert Responses

User testing second round. Feedback Likert scale. Very difficult to very easy 1-5.

User	Q1	Q2	Q3	Q4	Q5	Q6	SD	MODE	AVERAGE
1	4	4	3	4	4	5	0.6	4	4.00
2	5	5	4	5	5	5	0.4	5	4.83
3	5	3	3	4	4	5	0.8	5	4.00
4	5	5	4	5	4	5	0.5	5	4.67
5	5	5	3	5	2	5	1.2	5	4.17
6	5	4	3	4	4	5	0.7	4	4.17
7	5	5	5	5	5	5	0	5	5.00
8	5	5	4	5	5	5	0.4	5	4.83
9	4	4	4	4	4	4	0	4	4.00
10	5	4	3	4	4	5	0.7	4	4.17
11	5	5	2	5	4	5	1.1	5	4.33
12	4	4	2	4	4	5	0.9	4	3.83
13	5	5	3	5	5	5	0.7	5	4.67
14	5	4	3	4	4	4	0.6	4	4.00
15	5	4	2	4	4	4	0.9	4	3.83
16	5	5	4	4	4	5	0.5	5	4.50
17	4	4	4	4	4	4	0	4	4.00
18	4	4	3	4	4	5	0.6	4	4.00
19	5	5	4	5	5	5	0.4	5	4.83
20	5	5	3	5	4	5	0.8	5	4.50
21	5	4	3	4	4	5	0.7	4	4.17
22	5	5	2	5	4	5	1.1	5	4.33
23	4	4	4	4	4	5	0.4	4	4.17
24	5	5	4	5	4	5	0.5	5	4.67
25	4	3	5	4	4	5	0.7	4	4.17
26	5	4	4	4	5	4	0.5	4	4.33
27	4	4	3	4	4	5	0.6	4	4.00
28	4	4	4	4	4	5	0.4	4	4.17
AVERAGE	4.68	4.36	3.393	4.393	4.1429	4.821	16.4		
SD	0.47	0.61	0.817	0.488	0.5803	0.383	0.76		
Upper Limit (0.95)	4.86	4.6	3.72	4.59	4.37	4.97			
Lower Limit(0.95)	4.49	4.36	3.07	4.2	3.91	4.67			
Margin of Error	0.18	0.24	0.33	0.19	0.23	0.15			
Mode	5	4	3	4	4	5			

Appendix 10 Second round user testing raw data Part 2 Section B

Advanced questions. 1 represents correct response, 0 incorrect response to question.

User	Q9	Q10	Q11
1	1	1	1
2	1	1	1
3	0	1	1
4	1	1	1
5	1	1	1
6	0	1	1
7	1	1	0
8	1	1	1
9	0	1	1
10	0	1	1
11	1	1	1
12	1	1	1
13	0	0	0
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	1
19	1	1	1
20	0	0	1
21	1	1	1
22	1	1	1
23	1	0	1
24	1	1	1
25	1	1	1
26	1	1	0
27	1	1	1
28	1	1	1
Total (/28)	22	25	25
AVERAGE	0.79	0.89	0.89

Q9	CORRRECT	22
	INCORRECT	6
Q10	CORRRECT	25
	INCORRECT	3
Q11	CORRRECT	25
	INCORRECT	3

Appendix 11 Second round user testing raw data Part 2 Section B - SUS

SUS test NON-CODED

User	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	3	2	5	2	5	2	5	4	5
2	1	5	1	5	1	4	1	5	2	5
3	4	4	4	3	4	2	4	1	3	4
4	1	5	1	4	1	5	1	5	2	5
5	2	4	1	4	1	5	1	5	2	3
6	2	4	3	4	3	4	3	5	2	3
7	1	5	1	2	2	5	1	5	1	3
8	1	5	1	5	1	5	1	2	3	5
9	1	4	2	4	1	5	2	4	2	5
10	2	3	2	3	2	2	2	4	2	3
11	1	4	2	4	1	5	2	4	2	5
12	1	5	2	4	2	4	2	5	3	4
13	1	5	1	5	1	5	1	5	2	5
14	1	4	2	4	2	4	2	4	2	4
15	2	4	2	3	3	5	1	5	2	5
16	1	5	2	5	2	5	2	5	3	5
17	1	5	1	5	1	5	1	5	1	5
18	1	4	1	5	1	5	1	5	1	5
19	1	5	1	3	1	5	1	5	2	5
20	2	5	1	4	3	4	1	5	3	5
21	2	4	2	3	3	5	1	5	2	5
22	1	5	1	4	1	5	1	5	2	5
23	1	5	2	5	2	5	2	5	3	5
24	2	4	2	3	3	5	1	5	2	5
25	2	3	2	3	2	2	2	4	2	3
26	1	4	2	4	1	5	2	4	2	5
27	4	4	4	4	4	2	4	1	3	4
28	1	5	2	4	2	4	2	5	3	4

Appendix 12 Second round user testing raw data Part 2 Section B – SUS - ENCODED

SUS test CODED ODD N-1 EVEN 5-N convert to 0-4 scale with 4 being most positive.

TS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	TOTAL	TOTAL * 2.5
1	4	2	3	4	3	4	3	4	1	4	32	80
2	4	4	4	4	4	3	4	4	3	4	38	95
3	1	3	1	2	1	1	1	0	2	3	15	37.5
4	4	4	4	3	4	4	4	4	3	4	38	95
5	3	3	4	3	4	4	4	4	3	2	34	85
6	3	3	2	3	2	3	2	4	3	2	27	67.5
7	4	4	4	1	3	4	4	4	4	2	34	85
8	4	4	4	4	4	4	4	1	2	4	35	87.5
9	4	3	3	3	4	4	3	3	3	4	34	85
10	3	2	3	2	3	1	3	3	3	2	25	62.5
11	4	3	3	3	4	4	3	3	3	4	34	85
12	4	4	3	3	3	3	3	4	2	3	32	80
13	4	4	4	4	4	4	4	4	3	4	39	97.5
14	4	3	3	3	3	3	3	3	3	3	31	77.5
15	3	3	3	2	2	4	4	4	3	4	32	80
16	4	4	3	4	3	4	3	4	2	4	35	87.5
17	4	4	4	4	4	4	4	4	4	4	40	100
18	4	3	4	4	4	4	4	4	4	4	39	97.5
19	4	4	4	2	4	4	4	4	3	4	37	92.5
20	3	4	4	3	2	3	4	4	2	4	33	82.5
21	3	3	3	2	2	4	4	4	3	4	32	80
22	4	4	4	3	4	4	4	4	3	4	38	95
23	4	4	3	4	3	4	3	4	2	4	35	87.5
24	3	3	3	2	2	4	4	4	3	4	32	80
25	3	2	3	2	3	1	3	3	3	2	25	62.5
26	4	3	3	3	4	4	3	3	3	4	34	85
27	1	3	1	3	1	1	1	0	2	3	16	40
28	4	4	3	3	3	3	3	4	2	3	32	80
										MEAN	32.438	81.071
										MEDIAN	34	85
										SD	5.996	14.991

Appendix 13 Second round user testing raw data Part 2 Section D – PSSUQ

PSSUQ Questions 1 - 9

TS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
1	1	1	1	1	2	2	2	3	NA
2	2	1	1	1	1	1	1	1	NA
3	5	5	6	6	5	5	6	7	NA
4	2	1	2	1	1	1	2	1	NA
5	1	1	1	1	1	1	2	1	NA
6	3	3	3	4	3	3	3	3	6
7	1	1	NA	1	NA	1	1	3	NA
8	1	1	3	1	2	2	2	1	NA
9	2	2	3	3	3	3	4	3	NA
10	2	2	2	3	2	2	2	2	NA
11	1	1	1	1	1	1	1	1	NA
12	2	3	4	4	5	2	2	3	NA
13	2	2	3	3	2	2	1	1	NA
14	2	2	2	3	3	2	2	2	NA
15	1	2	3	3	3	2	3	3	NA
16	1	1	2	1	1	2	3	2	NA
17	1	1	2	1	1	1	2	3	2
18	1	1	1	1	1	1	2	2	NA
19	1	1	1	1	1	1	1	1	NA
20	1	1	2	1	1	2	2	NA	NA
21	1	1	1	1	1	1	1	1	NA
22	2	3	4	4	5	2	2	3	NA
23	2	1	1	1	1	1	1	1	NA
24	1	1	1	1	2	2	2	3	NA
25	1	2	3	3	3	2	3	3	NA
26	1	1	2	1	1	2	3	2	NA
27	2	2	3	3	2	2	1	1	NA
28	2	2	2	3	3	2	2	2	NA
MEAN	1.61	1.64	2.14	2.07	2.04	1.82	2.11	2.11	0.29

PSSUQ Questions 10-19

TS	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
1	2	3	3	3	2	1	1	1	3	1
2	1	1	1	1	1	1	1	1	1	1
3	6	NA	5	5	6	7	7	6	6	7
4	NA	NA	2	NA	NA	1	1	1	2	2
5	NA	1	1	1	1	1	1	1	4	1
6	5	NA	4	3	2	3	2	2	NA	3
7	NA	1	1	1	1	1	1	1	2	1

8	NA	NA	2	1	4	2	2	2	4	2
9	2	NA	3	3	NA	3	3	3	NA	2
10	NA	NA	3	2	2	3	2	2	NA	2
11	NA	NA	2	2	1	1	1	1	3	2
12	4	4	3	2	2	2	1	2	NA	2
13	1	NA	2	1	1	1	1	1	1	1
14	NA	NA	2	2	2	2	2	2	3	2
15	3	NA	3	2	3	3	2	2	2	2
16	NA	NA	1	3	2	1	2	2	3	1
17	NA	NA	2	2	2	2	2	2	2	2
18	NA	2	2	2	2	2	1	2	1	2
19	1	NA	1	NA	NA	1	1	1	1	1
20	3	NA	2	3	1	1	1	1	NA	3
21	NA	NA	2	2	1	1	1	1	3	2
22	4	4	3	2	2	2	1	2	NA	2
23	1	1	1	1	1	1	1	1	1	1
24	2	3	3	3	2	1	1	1	2	1
25	3	NA	3	2	NA	3	2	2	2	2
26	NA	NA	1	3	2	1	2	2	3	1
27	1	NA	2	1	1	1	1	1	1	1
28	NA	NA	2	2	2	2	2	2	3	2
MEAN	1.39	0.71	2.21	1.96	1.64	1.82	1.64	1.71	1.89	1.86

Appendix 14 Second round user testing raw data Part 2 Section D – PSSUQ – INTERFACE STATISTICS AND RENUMBERED

PSSUQ Questions 1-11

TS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
1	1	1	1	2	2	3	NA	2	3	3	2
2	2	1	1	1	1	1	NA	1	1	1	1
3	5	5	6	5	6	7	NA	6	NA	5	6
4	2	1	1	1	2	1	NA	NA	NA	2	NA
5	1	1	1	1	2	1	NA	NA	1	1	1
6	3	3	4	3	3	3	6	5	NA	4	2
7	1	1	1	1	1	3	NA	NA	1	1	1
8	1	1	1	2	2	1	NA	NA	NA	2	4
9	2	2	3	3	4	3	NA	2	NA	3	NA
10	2	2	3	2	2	2	NA	NA	NA	3	2
11	1	1	1	1	1	1	NA	NA	NA	2	1
12	2	3	4	2	2	3	NA	4	4	3	2
13	2	2	3	2	1	1	NA	1	NA	2	1
14	2	2	3	2	2	2	NA	NA	NA	2	2
15	1	2	3	2	3	3	NA	3	NA	3	3
16	1	1	1	2	3	2	NA	NA	NA	1	2
17	1	1	1	1	2	3	2	NA	NA	2	2
18	1	1	1	1	2	2	NA	NA	2	2	2
19	1	1	1	1	1	1	NA	1	NA	1	NA
20	1	1	1	2	2	NA	NA	3	NA	2	1
21	1	1	1	1	1	1	NA	NA	NA	2	1
22	2	3	4	2	2	3	NA	4	4	3	2
23	2	1	1	1	1	1	NA	1	1	1	1
24	1	1	1	2	2	3	NA	2	3	3	2
25	1	2	3	2	3	3	NA	3	NA	3	NA
26	1	1	1	2	3	2	NA	NA	NA	1	2
27	2	2	3	2	1	1	NA	1	NA	2	1
28	2	2	3	2	2	2	NA	NA	NA	2	2

PSSUQ Questions 12-16 + Interface Quality Statistics

TS	Q12	Q13	Q14	Q15	Q16	OVERALL	SQUAL	IQUAL	INTQUAL
1	1	1	1	3	1	1.75	1.67	2.00	1.67
2	1	1	1	1	1	1.13	1.17	1.17	1.00
3	7	7	6	6	7	5.80	5.67	5.40	6.33
4	1	1	1	2	2	1.62	1.33	2.33	1.33
5	1	1	1	4	1	1.53	1.17	1.80	2.00
6	3	2	2	NA	3	3.47	3.17	4.33	2.00
7	1	1	1	2	1	1.60	1.33	2.20	1.33

8	2	2	2	4	2	2.43	1.33	4.00	2.67
9	3	3	3	NA	2	3.23	2.83	4.25	3.00
10	3	2	2	NA	2	2.85	2.17	4.50	2.00
11	1	1	1	3	2	2.00	1.00	3.75	1.67
12	2	1	2	NA	2	3.20	2.67	4.50	1.50
13	1	1	1	1	1	2.20	1.83	3.60	1.00
14	2	2	2	3	2	3.00	2.17	5.00	2.33
15	3	2	2	2	2	3.27	2.33	5.40	2.00
16	1	2	2	3	1	2.71	1.67	5.00	2.33
17	2	2	2	2	2	2.80	1.50	5.00	2.00
18	2	1	2	1	2	2.67	1.33	5.20	1.33
19	1	1	1	1	1	2.29	1.00	5.50	1.00
20	1	1	1	NA	3	3.00	1.40	5.40	1.00
21	1	1	1	3	2	2.71	1.00	6.25	1.67
22	2	1	2	NA	2	3.87	2.67	6.17	1.50
23	1	1	1	1	1	2.44	1.17	4.67	1.00
24	1	1	1	2	1	3.13	1.67	5.83	1.33
25	3	2	2	2	2	4.00	2.33	8.50	2.00
26	1	2	2	3	1	3.43	1.67	7.50	2.33
27	1	1	1	1	1	3.13	1.83	6.40	1.00
28	2	2	2	3	2	4.00	2.17	8.50	2.33

Exit 0; ☺