Evaluating the VIPER pedigree visualisation: detecting inheritance inconsistencies in genotyped pedigrees

Evaluating the VIPER Pedigree Visualisation: Detecting Inheritance Inconsistencies in Genotyped Pedigrees

Trevor Paterson[1]
The Roslin Institute
University of Edinburgh

Martin Graham[2]
School of Computing
Edinburgh Napier University

Jessie Kennedy[3]
School of Computing
Edinburgh Napier University

Andy Law[4]
The Roslin Institute
University of Edinburgh

ABSTRACT

VIPER (Visual Pedigree Explorer) is a tool for exploring large complex animal pedigrees and their associated genotype data. The tool combines a novel, space-efficient visualisation of the pedigree structure with an inheritance-checking algorithm. This allows users to explore the apparent errors within the genotype data in the full context of the family and pedigree structure. Ultimately, the aim is to develop an interactive software application that will allow users to identify, confirm and then remove errors from the pedigree structure and scored genotypes.

This paper describes an evaluation of how VIPER displays the different scales and types of data set that can occur, along with a description of the further interface functionality necessary to meet the challenges such data presents. This is followed by an examination of a range of possible pedigree genotype errors by replicating these errors in controlled simulated data sets and showing how they are manifested in the VIPER interface and observed by a domain expert. The data sets used include both real and artificially generated data, the advantage of the latter being that they produce known effects in the visualization which the domain expert can then interpret as being useful or unhelpful as they see fit.

Keywords: Pedigree Visualisation, Error Visualisation, Interface Design, Utility Evaluation

Index Terms: H5.2. [Information Interfaces and Presentation]: User Interfaces

1 INTRODUCTION

Genotyped pedigree data underpins many forms of genetic analyses that are performed by breeders and biologists to identify, map and select economically or biologically important genes or heritable traits. For techniques such as linkage analysis, genotype scores for polymorphic markers across the genome are analysed in the context of the pedigree structure and Mendelian laws of inheritance. These statistical analyses are critically sensitive to any errors in the data that exhibit as 'inheritance inconsistencies', i.e. patterns of inheritance for alleles that are not consistent with the asserted parent-child relationships recorded in the pedigree. Any such errors must be identified and cleansed from the data before downstream analyses. Error cleansing constitutes a complex, labour-intensive expert task, particularly given the scale of modern genotyping studies, where populations of several thousand animals may be genotyped for tens of thousands of markers.

We have previously described a simplistic prototype tool for assisting data cleansing [7] that combines the ResSpecies genetic consistency-checking algorithm with a tabular display of genotypes for individuals within a pedigree. The tool highlights inconsistent genotypes and allows the interactive removal of identified erroneous data points. Critically, however, the tabular display format does not allow the user easily to explore patterns of errors in the context of the family structures in the pedigree 'tree'. The ability to explore inheritance patterns in this context is critical for pinpointing the exact data points in error, particularly in the case of incomplete datasets where the inheritance algorithm will infer logically consistent (missing) data from the existing data points and thus 'move' reported errors down to the lowest possible point in the pedigree.

In [4] we evaluated pre-existing pedigree visualisation tools and demonstrated that none were suitable to assist the biologist in exploring errors in the complex pedigree structures and associated genotypes found in experimental data sets. We further presented the design and development of a novel pedigree visualisation method (the "sandwich" visualisation) as part of a new interactive visualisation tool (VIPER). The design was derived from an analysis of the requirements and working practices of experienced biologists and consideration of the pros and cons of existing pedigree, graph and matrix visualisation techniques.

This paper presents a two stage evaluation of the interactive visualisation provided in VIPER. An initial evaluation was performed on a small number of test datasets and used to identify and implement any critical features or improvements required for a functionality evaluation. A second, in-depth evaluation performed by expert biologists tested the effectiveness of the visualisation in helping to identify a variety of representative error states deliberately introduced into simulated pedigree and genotype data sets. The paper then concludes with a discussion of the results and further work.

2 BACKGROUND

2.1 Pedigree and Genotype data

Roslin Bioinformatics provides the web-based ResSpecies data system (www.resspecies.org) for recording and analysing animal pedigree-genotype data. The experimental pedigrees available in ResSpecies, particularly those from the five major farmed species (Chicken, Turkey, Pig, Cow and Sheep) exemplify the variety in structure and scale of study populations currently encountered and these pedigrees have been used (after anonymisation) for the generation of test datasets employed in the evaluation.

Pedigrees stored in ResSpecies range in size from 45 to 11000 individuals, but more typical sizes range from 100 to 2500. The structure of each particular pedigree reflects the design of the breeding experiment, e.g. inbreeding versus outbreeding, with the number of generations varying between 2 and 11. Similarly the

[1]e-mail: trevor.paterson@roslin.ed.ac.uk
[2]e-mail: m.graham@napier.ac.uk
[3]e-mail: j.kennedy@napier.ac.uk
[4]e-mail: andy.law@roslin.ed.ac.uk

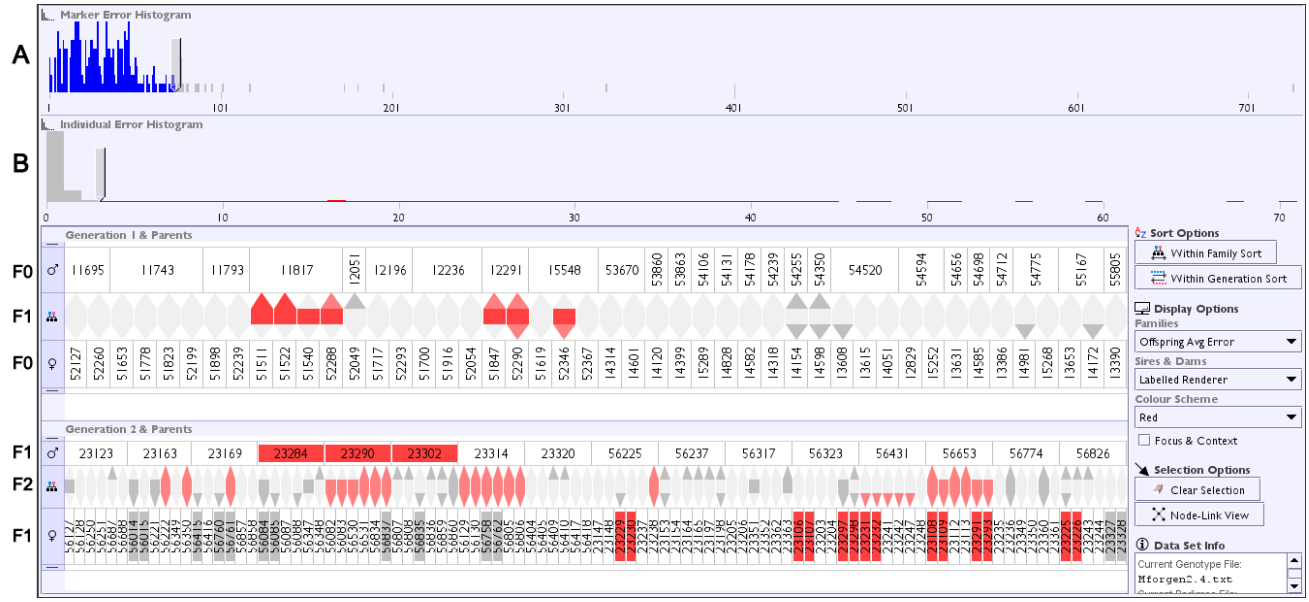

Figure 1. VIPER prototype as described in [4]. A three generation pedigree scored for 281 markers is shown in an aggregated family view. Family groups are shown as hexagons (F1, F2) sandwiched between F0 and F1 parents. The sensitivity of error reporting is controlled through the two histograms with sliders. Slider A filters out information about markers above the selected error threshold (75 here) and slider B alters the colourisation sensitivity (raising the reporting threshold to 3 errors here).

number of founder animals in each pedigree varies between 2 and 1200; founders may be introduced throughout a breeding program, not just in 'generation 0', and the proportion of founders used in a study varies greatly from 0.3% to 55%. The proportion of males recorded in a pedigree ranges from 0.7% to 94%, whilst the proportion of females varies between 2 and 98%. Some pedigrees, particularly fowl studies, may not record the sex of animals not kept for breeding, resulting in up to 98% of individuals unsexed, however, more typically only a few percent of animals are unsexed. Sexing becomes a particular issue when identifying the inheritance pattern of sex-linked markers. The shape of pedigrees (i.e. the number/proportion of individuals per generation) also varies, with some experiments using very few individuals in earlier generations, but generating large numbers in the final generations. The choice of mate selection is also study dependant, with some studies crossing a single individual (often a male) with multiple partners, even across generations.

This great variety in pedigree structure differs from typical human pedigrees, and the imbalances, multiple and cross generational pairings, and in some cases the size of families, present additional challenges for a successful pedigree visualisation which would allow the user to trace inheritance patterns from ancestors to descendants and siblings.

Inheritance studies use genotypes scored for any number of detectable genetic markers distributed across the genome of the study organism. A specific marker genotype is scored for each individual in the pedigree by detecting the (paternally and maternally inherited) allele pair. This allows the inheritance pattern of alleles to be traced through the pedigree structure. Current large scale genotype studies are based on SNP-chip technology, i.e. the identification of bi-allelic Single Nucleotide Polymorphisms, allowing genotypes to be concisely represented by pairing single nucleotide characters (ACGT) or '-' for null sex-linked alleles. Earlier genotype studies typically assayed fewer, more variable genetic markers, with multiple 'named' alleles, by a variety of less automated techniques. The potential scale of SNP-chip datasets reflects the availability of tens of thousands of SNP markers for study organisms.

Real experimental datasets commonly contain missing genotype data. Whilst this may occur sporadically due to lost samples or

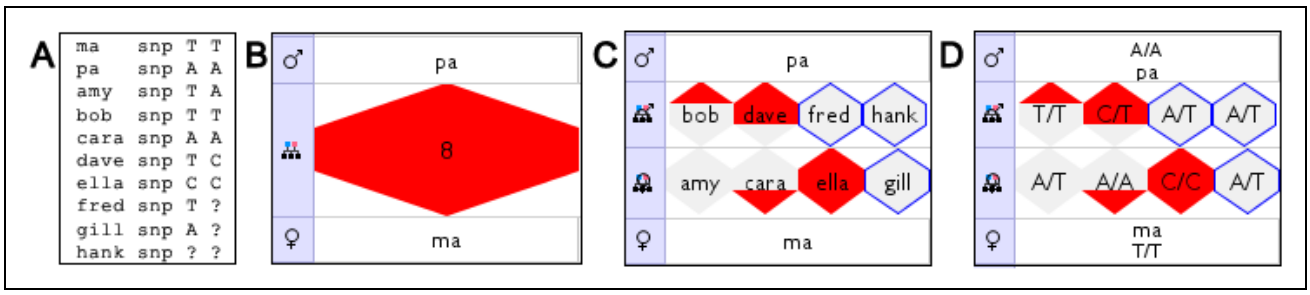

Figure 2. Detail of a single family shown in VIPER, illustrating the reporting of errors and the application of inference. (A) Input genotypes for one family, for one marker 'snp', note that 3 individuals have incomplete data. (B) Family glyph for 8 offspring, 'sandwiched' between the parents, exhibiting all 3 error types. (C) Individual glyphs in data overview, 3 individuals are highlighted as 'incomplete'. (D) Individual glyphs in single marker detail labelled with actual genotypes, note the 3 'inferred' genotypes.

failed assays, missing data frequently reflects a systematic decision not to analyse samples for some individuals or generations which may be considered uninformative. The ResSpecies inheritance-checking algorithm infers inherited genotypes for missing data points using the principles of Mendelian inheritance, which may result in either completely resolved or partially resolved (e.g. 'T/?') genotypes. More complex partial inferences are possible for multi-allelic markers (e.g. [T or A]/[T or A or C]) but for simplicity this study only uses bi-allelic markers. Sex-linked inheritance patterns are observed for markers located on the sex chromosomes, where the heterogametic sex has a single allele for these loci. If sex-linkage is known in advance a null allele may be recorded in a dataset (e.g. 'A/-'), otherwise the genotype would typically be erroneously scored as homozygous (e.g. 'A/A'), with the inheritance pattern of an unrecognized sex-linked marker exhibiting a distinctive error pattern.

2.2 VIPER

VIPER adopts a 'family-centric' sandwich view in its pedigree visualisation [4]. Briefly, the technique lays out the pedigree by generations. Within each generation, the top row of the "sandwich" represents the male parents (sires) whilst the bottom row contains the female parents (dams). The offspring are grouped into families (those sharing a common set of parents) within the cells between the respective mate pairs (see Figure 1). The user can toggle between an aggregated family overview (a single set of statistics per family), or the display of all individuals separately within each family or 'mating pair' (see Figure 2B and 2C). In addition, the user can re-sort parents or offspring using a variety of data properties, change colour schemes and highlight selected individuals or families together with their ancestors and descendants by click/ctrl-click selection. This representation provides the family-centric visualisation necessary to view and assess errors in the context of a pedigree structure.

Reported inheritance inconsistencies may be categorised into three types: genotypes where no allele is inherited from the sire, where no allele is inherited from the dam, or where a novel, non-parental allele is detected. A single genotype may exhibit any or all of these error categories. The three categories of error are represented in the offspring row as the component parts of a hexagonal glyph, with the tips acting as stylised arrows oriented either up or down. These tips point to the sire and dam rows with colour coding for the sire or dam errors, and the 'mid-stripe' of the hexagon is coloured for novel allele errors. In the sire and dam rows, the combined error count for the sire or dam is used to colour the representation of an individual (see Figures 1and 2). A simple discrete four-level colour-coding is used to indicate the proportion of erroneous markers associated with an individual, from white (no errors) through light, mid and heavy colour shading for increasing error rates.

Two histograms with integrated slider widgets are used to control the sensitivity of error reporting (see top of Figure 1). They both give a summary view of genotype errors, but one is binned by error count per marker and the other by error count per individual. The first histogram reports the number of markers (y-axis) with a given error count (x-axis) across the individual set, whilst the second histogram reports the number of individuals (y-axis) with a given error count (x-axis) across the marker set.

In keeping with the principles of dynamic querying [1], the histogram sliders interactively and quickly allow the user to 1) filter out from the analysis markers above the selected error threshold and 2) to alter the thresholds for colouring an individual's error display in the sandwich view. This filtering allows the biologist to home in on problematic mating pairs by either removing markers with errors above a threshold (i.e. 'very bad' markers) or controlling the heatmap sensitivity by number of errors per individual. 'Very Bad' markers can be removed from the analysis because the inheritance algorithm is applied independently to each marker, however, bad individuals cannot simply be removed from the pedigree.

Further, removing markers using the top histogram dynamically changes the error counts that are shown in the second histogram. The effect is similar, but not exactly the same, as that seen with multiple histograms in the Attribute Explorer [10] - there, filtered out items were coloured differently rather than omitted altogether.

3 METHODOLOGY

The evaluation was performed in two stages. The first stage validated the ability of VIPER to handle and display the types of datasets to be used in the second stage - a utility evaluation [9] in which we tested the visualisation's ability to faithfully represent pedigree genotype data sets of the necessary scale, and to cope with and indicate errors and omissions within them. This is distinct from usability or efficiency testing, as the primary aim here is to ensure the necessary functionality of the system is present. This allowed the identification of critical features for subsequent implemention to support data browsing and error localisation.

The second evaluation stage involved testing the visualisation's capability for displaying differing error types commonly found in real-world pedigree genotype data sets. Each separate 'error type' under test was evaluated by creating an appropriate permuted pedigree and genotype data file pair, and then browsing the data visualized in VIPER to verify whether the pattern of inheritance inconsistencies revealed could be used to deduce the underlying, causative error. Simulated data sets were used in order that each error type could be evaluated independently, without the confounding effect of multiple overlapping and interfering errors as found in real data sets.

The two stages of the evaluation were performed by the two authors from Roslin with extensive experience in analysing and error-cleaning genotyped pedigree datasets. Such an evaluation might not have the numbers of other usability or utility inspection methods, but the expertise of the domain experts in helping assess the visualisation is the overriding factor here [6], an assessment by novices to the domain would not glean near as much information.

One biologist created and anonymised the data sets, and monitored the other biologist exploring each data set in one-to-one sessions lasting an hour. The ease and accuracy with which errors were identified was qualitatively scored, and any issues with the interface or comments about desirable improvements were recorded. These observations formed the basis for deciding which additional information and functionality it is essential to present to the user, what modifications might be beneficial but not essential, and any general usability and navigation issues. The approach in whole has a similarity with expert reviews [11] but the experts here are domain rather than visualization experts, as they are the only ones who can truthfully assess whether the necessary functionality is present and correct.

The majority of the evaluations used a moderately-sized, anonymised chicken pedigree comprising 1792 individuals, the details of which are shown below in Table 1:

Table 1. Statistics for anonymized chicken pedigree.

Generation	Male	Female	Total
F0	28	48	76
F1	16	102	118
F2	0	1598	1598
			1792

Table 2. Representative results for synthetic large pedigree files examined in VIPER, showing cut-off levels of resolution for a standard 1280x1024 monitor.

Individuals	Generations	Families / Generation	Usability Limitations
10 000	30	10	Vertical scrolling accommodates 'any number' of generations.
10 000	3	100	Families display reasonably, but individual offspring icons too small to display genotype labels and distinguish error glyph reliably.
5 000	5	50	Families display reasonably, but individual offspring icons too small for genotype labelling.
5 000	3	250	Families at limit of usable resolution for standard monitors, and individual offspring icons too small for labelling and error glyphs.

Alternately pedigrees with controlled numbers of individuals, generations and families per generation were generated *de novo* using a parameterisable script. Dummy genotype data files for pedigrees were similarly created using a suite of creation and permutation scripts, and desired errors were introduced into either the pedigree or genotype files manually or with further editing scripts.

The simulated genotype data sets reflect the data types found in current large scale studies based on bi-allelic SNPs, including sex-linked markers. A suite of scripts was used to create and then systematically corrupt synthetic genotype data, and to partially erase data from the F1 generation to simulate incomplete data coverage. Initially consistent genotype data was generated using seven different randomly seeded markers. Each marker had bi-allelic SNP alleles C and T, with 5 different C:T heterozygosity ratios (1:1,1:2,1:3,1:4,1:5), 1 mammalian style male sex-linked pair (C, T, Y-null), and 1 female (avian style) sex-linked pair (C, T, W-null). Consistent datasets were generated for 7, 70 and 350 markers by seeding with each marker one, ten or fifty times. The 70 marker genotype dataset proved to be adequate for revealing the expected error pattern for the majority of error types in the data overview.

4 First Stage Evaluation

VIPER's ability to handle and display representative data sets was initially validated with regards to three particular aspects of the data in question. Firstly, the ability to handle pedigrees of a range of sizes; secondly, the effect of incomplete data which requires inferencing over the missing data, and thirdly the ability to report systematic errors in sex-linked markers.

4.1 Size and Structure of Pedigrees

A wide range of animal pedigrees extracted from the ResSpecies data source were tested to confirm the layout and display capabilities of VIPER over a realistic range of experimental pedigrees size and structures. In addition, in order to test the limits of display resolution and usability a number of pedigrees were created with controlled numbers of total individuals, generations and families per generation, as listed in Table 2.

In summary, it was demonstrated that the visualisation can cope with any realistic number of generations and over 200 families per generation at the overview level, although the labelling of parent names becomes problematic with over 100 families. However, available space constrains the ability to distinguish the properties of individual offspring where there are a large number of families in a generation (50 to 100) or a large number of offspring in a family. None of the experimental pedigrees available in ResSpecies exceed these thresholds. Display limitations could be ameliorated with higher specification monitors, but in the authors experience the target user group for VIPER, e.g. animal breeders, often lack high specification desktop hardware aand monitors.

4.2 The effect of incomplete data and genotype inference

As described above, in addition to reporting genotypes that are inconsistent with Mendelian transmission, the ResSpecies inheritance algorithm infers missing genotype data by recursively applying allele transmission that must necessarily be true from known data points. As a consequence of the algorithm traversing the pedigree from founders (F0) down through descendants (to F2 here), errors are reported as low down the pedigree as possible,

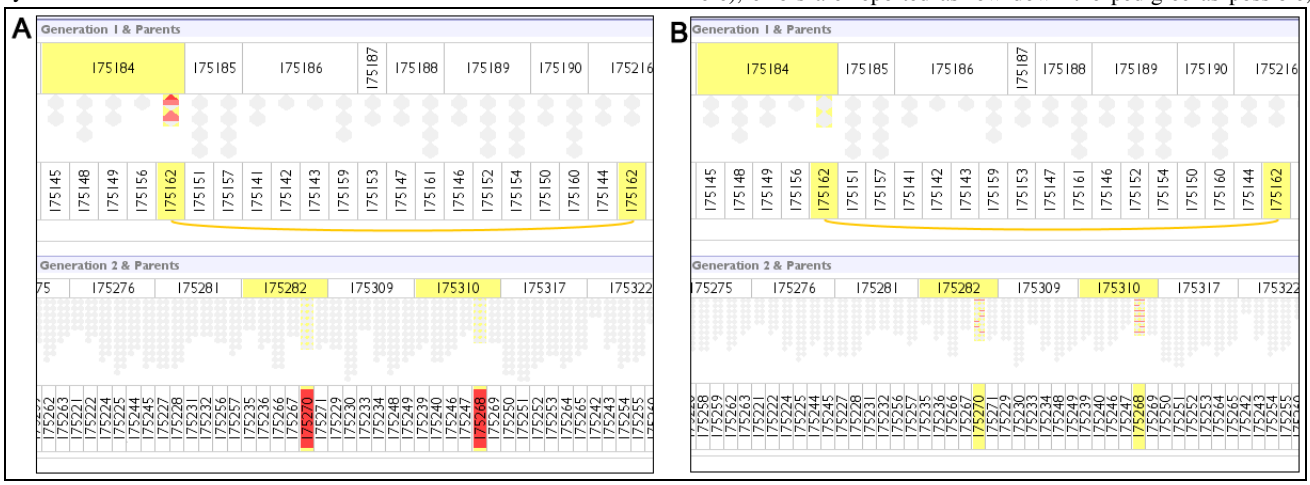

Figure 3. Comparison of complete genotypye data set (A) with data set with 50% of F1 genotypes removed (B). Two offspring of (175162x175216) are wrongly assigned to sire 175184. In (A) these offspring report failure to inherit from the supposed father, but in (B) this is obfuscated due to missing data and the error is now reported in the progeny of the wrongly assigned litter.

and particularly in the context of missing data and genotype inference, errors can be reported in individuals (siblings or descendants) removed from the actual source error. The obfuscating effect of this was apparent when synthetic data sets were examined, where a proportion of genotypes were erased from the intermediate generation (F1) individuals, see Figure 3.

4.3 Sex-linked Markers

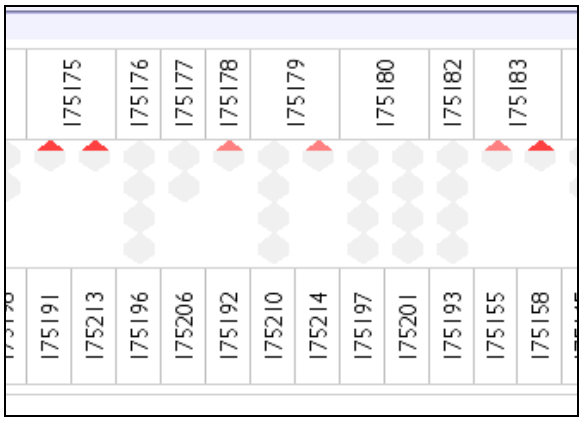

Figure 4. Multiple individuals report 'nil from sire' errors (red upper hexagon) due to unrecognized sex-linkage. All of the affected individuals are in fact male in this case, and cannot inherit a sex-linked allele from their father.

A common systematic error found in real datasets arises when unrecognized sex-linked markers are analysed. Typically this arises in mammals when the genotype assay scores males as homozygous for an allele, whereas in fact they should be heterozygous for the 'y-null' (absent) allele; the effect is opposite in most birds with heterozygous 'z-null' females unrecognized. As can be seen in Figure 4 this causes a gross systematic error to be reported, immediately apparent as a preponderance of '*nil from sire*' errors (for mammals). However the sex segregation of this effect is not readily apparent as the sex of individuals is not represented in the sandwich view.

5 Improvements to VIPER

This initial evaluation of VIPER identified several features which were required prior to performing the second stage of the evaluation. These improvements are illustrated in Figure 5.

In order to support the exploration of individuals in any selected large family a 'Detail View' window was implemented to complement the overview of the entire pedigree that was already present – one of the standard practices for solving such problems in Information Visualisation, as documented in [3]. This 'Detail View' can show a detailed representation of families with hundreds of individuals.

In order to expose the degree of genetic inference in the data (which occurs because of data incompleteness) a second colourmap was implemented that shows the degree of inference across the pedigree via the intensity of the border colour on an individual or family representation. When data is visualised for a single marker (as in Figures 2C, 2D and 5), a (single-state) blue border indicates that an individual genotype has been derived by genetic inference. Clashes between the dual colour highlighting used in the sandwich view to report inference and error rate are limited, because the inheritance algorithm does not infer 'erroneous' genotypes from incomplete data.

In order to expose the sex of individuals in a family, and hence assist the identification of sex-linked inheritance problems, a further level of colouring was rejected as it would reduce the pre-attentive 'pop-out' [5] that the coloured error display currently enjoyed. Instead the (optional) partitioning of offspring by sex was implemented, spatially separating the male and female offspring into different rows – in effect creating a 'club sandwich' view. In essence, a separate visual attribute, spatial positioning, is being used to communicate low-count categorical data attributes of the offspring, rather than overload the colour channel. Standard pedigree layouts use shape to represent gender in pedigree diagrams [2] but spatial positioning is a more powerful visual communicator, especially when it is desirable to split a set of objects into groups.

The initial VIPER prototype provides only an 'overview' of summary information about inheritance errors averaged across all markers. This summary view adequately exposes many types of systematic errors resulting from wrong pedigree information or sample mis-identification, but it does not allow the discrimination of more sporadic errors, nor can the user explore the actual reported genotypes for a given marker. This deficiency was

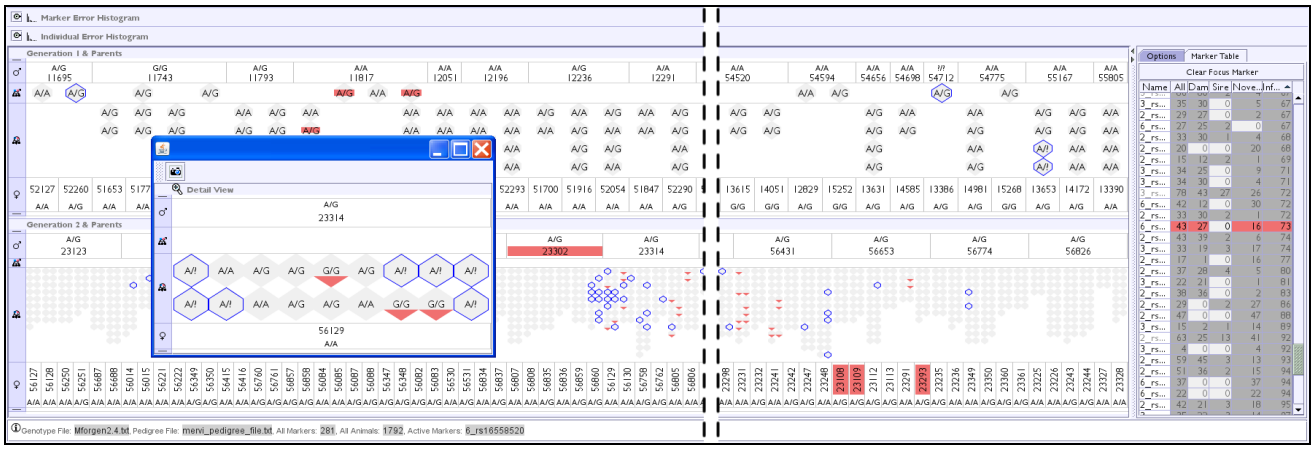

Figure 5. Improved VIPER prototype following the first evaluation step (with the same data analyzed as in Figure 1). Individual offspring can now be separated by sex (male above female). The 'Marker Table' tab allows sorting of markers by error metrics, and selection of any particular marker for display in isolation. Here the display shows the recorded or inferred genotypes for the marker highlighted in red in the 'Marker Table'. Incomplete data is highlighted via a blue border, and a 'Detail View' window allows inspection of a family in full detail.

addressed by adding a sortable 'Marker Table' to allow the user to select individual marker genotype data to explore. Markers can be sorted according to their name, counts of reported error types (sire, dam, novel allele or all) and degree of inference. A 'focal marker' can be selected in the table, allowing specific genotype information for that marker to be overlaid in the sandwich visualisation – as seen in Figure 5. Further, by scrolling the mouse wheel on the marker table or using the keyboard the focal marker and the resulting display in the sandwich visualisation can be rapidly changed. The marker table uses the same error colouring scheme as the sandwich view and the error counts and colouring are similarly tied in to the filtering operations of the histograms. In this way the histograms, sandwich view and marker table now form an example of a coordinated multiple view visualisation.[8]

The single marker display is essentially identical to the overview, but adds the actual or inferred genotype to the labelling of individuals, allowing the user to analyse in detail the inheritance patterns of alleles in the pedigree. Note that when viewing the data for single markers, both the (red fill) 'heatmaps' on the error glyphs and the (blue border) inference highlights become 'binary' (on/off) indicators.

With these improvements applied to the VIPER prototype, it was now felt that the second stage of the evaluation could proceed.

6 SECOND STAGE EVALUATION

The second stage of the evaluation explored the effect of introducing controlled errors into real and artificial pedigree genotype data sets, and whether they would be represented by the visualisation in a form recognisable to a domain expert.

6.1 Pedigree Errors

Real datasets frequently contain errors in the asserted pedigree structure, which might be caused by the mis-identification of animals, incorrectly assigned paternity or errors in record keeping. Furthermore sample mis-identification or contamination can result in apparent pedigree errors.

In order to evaluate whether the VIPER visualisation adequately exposes the possible kinds of pedigree errors found in real datasets various pedigree disruptions were engineered in the categories given in Table 3. Where appropriate these permutations were performed in separate generations (F0, F1, F2) and upon both same sex and different sex pairs of individuals. Furthermore, to evaluate the potential effects of inference on the observed inheritance patterns, genotype files were derived with 50 or 100% of F1 genotypes erased.

Table 3. Categories of pedigree permutations explored in VIPER. All permutations were identified in the sandwich overview apart from No.11, where inconsistent sex breaks the pedigree, causing a fatal error on pedigree file loading.

1.	Alter Father of Individual
2.	Alter Father of Family
3.	Alter Father of Litter
4.	Alter Father of Sire Sibs
5.	Alter Mother of Individual
6.	Alter Mother of Family
7.	Alter Mother of Litter
8.	Alter Parents of Individual
9.	Alter Parents of Family
10.	Alter Parents of Litter
11.	Alter Sex of Individuals

Pedigree files drawn from the categories in Table 3 were explored in VIPER using the 70 marker test genotype dataset (described above) and with the 100%, then 50% then 0% erased F1 genotypes. The ease with which the error types were located and identified was assessed, and the influence of genotype inference on the inheritance pattern was considered. The categories listed in Table 3 were all successfully explored, and only particular notes recorded here. Permutations of type 3 clearly demonstrate the importance of the blue-border inference highlight, to draw attention to the obfuscating effect of inference over missing data. As described above (Figure 3) the reported errors are pushed down to F2 when 50 or 100% genotypes are erased, making diagnosis more difficult. The improved VIPER prototype draws attention to the lack of genotype data for the wrongly assigned littermates, alerting the user to the possibility that the errors reported in F2 may be propagated from the F1 generation (see Figure 6).

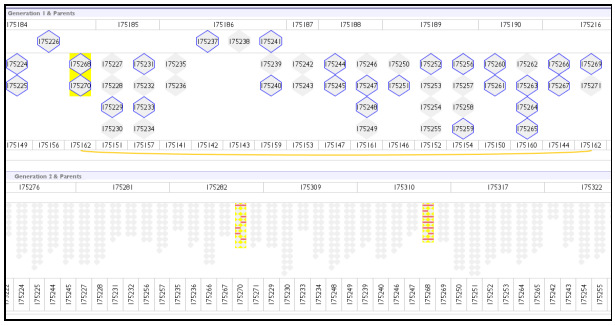

Figure 6. The data set shown in Figure 3B which has 50% of F1 genotypes deleted is reanalyzed. The wrongly assigned F1 littermates (175270 and 175268) are highlighted in yellow. The addition of blue-border highlighting of inferred genotype data points throughout the F1 progeny draws the user's attention to the possibility of error propagation by the algorithm. As seen in Figure 3, the algorithm reports errors in the F2 progeny rather than in the mis-assigned F1 parents.

6.2 Genotype Errors

Genotyping assays can give rise to systematic or sporadic errors. Unreliable assays may give rise to unusable data with very high error frequencies, but a low rate of sporadic 'wrong calls' cannot be discounted for any assay. Errors in sample or data handling may again be systematic or sporadic, and hence might give rise to inconsistency patterns resembling systematic pedigree errors, or to more random, less tractable patterns.

Various types of errors were introduced into hitherto consistent (error-free) genotype files, as categorized in Table 4. Where appropriate errors were introduced to individuals in different generations (F0, F1, F2), and, in order to demonstrate the potential effects of inference on the observed inheritance patterns, alternate data versions created with 50 or 100% of the F1 genotypes erased. The permutations and genotype mixings were also done on both same sex and different sex pairs of individuals.

Figure 7 shows analysis of a representative example error of type 3 in Table 4, and models the case where samples (or genotyping results) have been swapped between two unrelated generation 1 individuals. The inheritance checking algorithm reports multiple apparent inheritance inconsistencies for the misidentified samples, and their offspring. However, when generation 1 genotype data is incomplete, the genetic inference of missing data has the consequence of spreading reported errors through a sister of the misidentified individual to several of her nieces (see Figure 7B).

Table 4. Categories of genotype permutations explored in VIPER. All permutations (1-14) apart from (7) were identified in the sandwich overview visualisation. Although offspring can be sorted by litter information, there is as yet no suitable visualisation for litter mates; consequently attention is not drawn to errors restricted to a particular litter.

1.	Exchange Complete Genotypes between Individuals
2.	Exchange Some Genotypes between Individuals
3.	Swap All Genotypes from One into a Different Individual
4.	Swap Some Genotypes from One into a Different Individual
5.	Mix Some Random Genotypes into Individual
6.	Regenotyped Family with Novel Father
7.	Regenotyped Litter with Novel Father
8.	Regenotyped Full Sire Sib Set with Novel Father
9.	Regenotyped Individual with Novel Father
10.	Score Sex Linked Marker as Homozygous
11.	Swap Non Sibling IDs between generations
12.	Swap Non Sibling IDs in same generation
13.	Swap Sire Sibling IDs in same generation
14.	Swap Siblings IDs

In summary, identification of these various systematic ID / genotype / parentage swaps proved tractable for experienced geneticists using the sandwich pedigree layout. In particular the ability to select and highlight an individual and its ancestors and descendants allows inheritance patterns to be traced.

6.3 Large Genotype Data Files

The memory efficiency of data loading and processing has not yet been addressed, but the ability of the current prototype VIPER to load and display increasingly large marker datasets on relatively low specification hardware is examined here. With a 1000 individual 5 generation pedigree, VIPER running with 1G RAM on Win32 could load and display genotype datasets for many hundred markers, but with 1000 markers the program exceeded available memory whilst instantiating the genotype objects. Using Linux64 2.5G RAM architecture the memory limit was not reached until 2500 markers were loaded, although processing was slow with 2000 markers. Below these memory thresholds the program behaved as expected, providing the filterable summary overview, and allowing selection of marker by marker genotype views. Even allowing for possible memory optimisations to the data model and the inheritance algorithm the scale of current SNP-chip datasets suggests that a data pre-processing and segmentation controller will be required to handle very large datasets, and to guide the user through exploration of problematic markers.

7 Conclusions

The VIPER prototype has been evaluated for the display, exploration and identification of errors in genotyped pedigree datasets, using a range of synthetic datasets which incorporate a wide variety of pedigree and genotype errors, and introduce degrees of data erasure to mimic data incompleteness. The first evaluation exposed a number of critical features that were implemented prior to the full functional evaluation (as described above: the single marker view and table, the 'Detail View' for large families, the heatmap of genotype 'incompleteness' and the ability to sort siblings by sex). The results of the second functional evaluation confirmed the ability to discriminate the vast majority of single error types in pedigree and genotype datasets.

Findings from the evaluations can be split up into two categories: what we learnt about VIPER in particular and what we

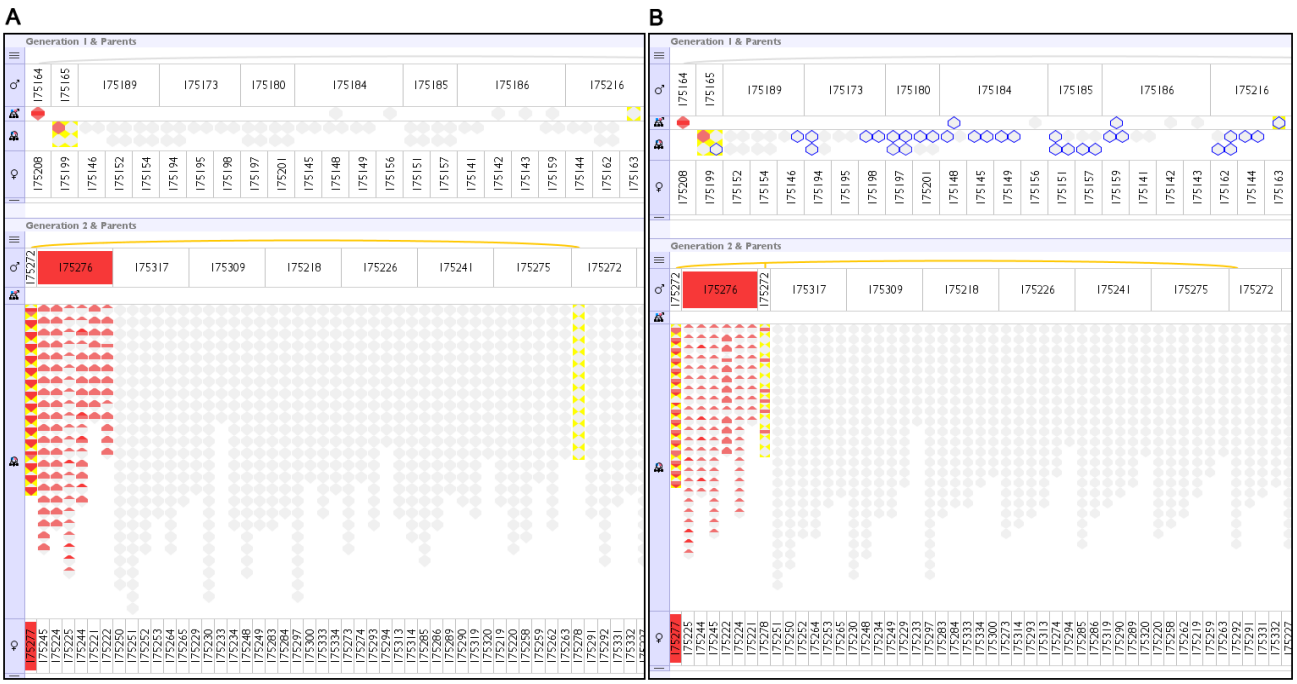

Figure 7. (A) Genotype dataset corrupted to swap two samples from generation 1 (175277 female / 175276 male). Both samples report multiple inheritance inconsistencies of all three types: nil from sire, nil from dam and novel alleles. Generation 2 offspring from these individuals report failure to inherit from their misidentified parent. In (B) incomplete genotype information for generation 1 individuals causes inference by the genetic algorithm (blue borders), which has the effect of propagating the reporting of errors to offspring of 175278, a sister of the mis-sampled 175277.

discovered about the process of testing an application with domain experts in this manner. For VIPER in particular, the space-efficient layout of the pedigree population in generational layers, organized by mating pairs (families) allows realistically large pedigree datasets to be explored, and the ability to toggle between a summary family view and detailed view of individual offspring provides a workable compromise between a simplified overview and individual detail. The mechanism for highlighting the ancestors (parents) and descendants (children) of an individual allows the user to trace inheritance patterns across the pedigree.

The display of error frequency via a heatmap imposed on directional glyphs (nil from sire, nil from dam, novel allele) not only directs the user to error locations, but provides evidence about the nature or source of the error. The display of a heatmap reflecting the completeness of genotype data for an individual is critical for considering how the reporting of an error may have been propagated down the generations by the inheritance algorithm.

In general, we found that testing the application with data sets of the size and complexity that crop up in the everyday working practices of these domain experts was essential; it validated that the visualization could cope with data it could expect to encounter in practice. Not having a visualization that can cope with representative data would negate most, if not all, of the advantage of later bringing in real users to interact with it. Note that we say representative; as well as having real data in the form of ResSpecies pedigrees, we also generated artificial data sets to test the effect of particular combinations of data size and granularity on the visualization.

These artificially generated data sets have the advantage that we know what they should look like in the visualization if all goes to plan. Trying to analyse whether real data sets have rendered properly would depend on a working knowledge of that particular data set, which is a catch-22 when considering that gaining such knowledge of that data is why we wish to visualize it in the first place.

This held true into the second evaluation stage where known errors were introduced into real and generated pedigree genotypes. Again, visualising an existing data set known to have errors in would have required deep knowledge of that particular data set to see if VIPER was communicating those errors properly. By artificially injecting controlled errors, both pedigree and genotype, into clean data sets we can quickly ascertain whether the visualization is communicating the presence of error and then, according to the domain expert, whether that communication makes sense. There is also the bonus that such data sets will make handy training data sets for new users to the prototype in the future. Once the visualization has been verified as having the functionality necessary to correctly inform an expert user we can then revert to the real data and real users mantra

8 FUTURE WORK

The implementation of the ability to display actual genotypes and errors on a marker by marker basis, by selecting from a table sortable by marker properties (error rate and completeness) was essential for the next stage of VIPER development, namely data cleaning functionality. Following the approach of our earlier GenotypeChecker application [7] we will implement hypothesis testing functions with which the user will be able to test the effect of removing candidate errors. By temporarily removing or masking selected problematic data points (genotypes, individuals or pedigree relationships) and then reapplying the inheritance checking algorithm, the identity of causative errors can be confirmed. By these means the user will be able to incrementally identify and remove the minimal set of bad data points that must be removed to create a completely consistent dataset.

REFERENCES

[1] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic Queries for Information Exploration: An Implementation and Evaluation," In *Proc. ACM CHI* (Monterey, California, USA), pp. 619-626, 3-7 June 1992. ACM Press.

[2] R. L. Bennett, K. A. Steinhaus, S. B. Uhrich, C. K. O'Sullivan, Robert G. Resta, D. Lochner-Doyle, D. S. Markel, V. Vincent, and J. Hamanishi, "Recommendations for Standardized Human Pedigree Nomenclature," *American Journal of Human Genetics*, 56(3):745-752, March 1995.

[3] A. Cockburn, A. Karlson, and B. B. Bederson, "A Review of Overview+Detail, Zooming, and Focus+Context Interfaces," *ACM Computing Surveys*, 41(1), December 2008.

[4] M. Graham, J. Kennedy, T. Paterson, and A. Law, "Visualising Errors in Animal Pedigree Genotype Data," *Computer Graphics Forum*, 30(3):1011-1020, June 2011.

[5] C. G. Healey, K. S. Booth, and J. T. Enns, "High-speed Visual Estimation using Pre-attentive processing," *ACM Transactions on Human-Computer Interaction*, 3(2):107-135, June 1996.

[6] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware, "User Studies: Why, How, and When?," *IEEE Computer Graphics and Applications*, 23(4):20-25, July/August 2003.

[7] T. Paterson and A. Law, "GenotypeChecker: An interactive tool for checking the inheritance consistency of genotyped pedigrees.," *Animal Genetics*, In Press(??):??-??, 2011.

[8] J. C. Roberts, "State of the Art: Coordinated & Multiple Views in Exploratory Visualization," In *Proc. Coordinated and Multiple Views in Exploratory Visualisation* (Zurich, Switzerland), pp. 61-71, 2 July 2007. IEEE Computer Society Press.

[9] G. Robertson, "Beyond Time and Errors - Position Statement," In *Proc. BELIV* (Florence, Italy), 5 April 2008. ACM Press.

[10] R. Spence and L. Tweedie, "The Attribute Explorer: information synthesis via exploration," *Interacting with Computers*, 11(2):137-146, December 1998.

[11] M. Tory and T. Möller, "Evaluating Visualizations: Do Expert Reviews Work?," *IEEE Computer Graphics & Applications*, 25(5):8-11, Sept-Oct 2005.