Olorin: combining gene flow with exome sequencing in large family studies of complex disease

James A. Morris* and Jeffrey C. Barrett
Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK

ABSTRACT

Motivation: The existence of families with many individuals affected by the same complex disease has long suggested the possibility of rare alleles of high penetrance. In contrast to Mendelian diseases, however, linkage studies have identified very few reproducibly linked loci in diseases such as diabetes and autism. Genome-wide association studies have had greater success with such diseases, but these results explain neither the extreme disease load nor the within-family linkage peaks, of some large pedigrees. Combining linkage information with exome or genome sequencing from large complex disease pedigrees might finally identify family-specific, high-penetrance mutations.

Results: Olorin is a tool, which integrates gene flow within families with next generation sequencing data to enable the analysis of complex disease pedigrees. Users can interactively filter and prioritize variants based on haplotype sharing across selected individuals and other measures of importance, including predicted functional consequence and population frequency.

Availability: http://www.sanger.ac.uk/resources/software/olorin
Contact: olorin@sanger.ac.uk

Received on June 20, 2012; revised on October 4, 2012; accepted on October 8, 2012

1 INTRODUCTION

Next generation sequencing has rapidly become the standard approach for identifying mutations responsible for Mendelian diseases (Bamshad et al., 2011). Although software and file formats for the processing of raw sequence data are relatively robust (Danecek et al., 2011; Li et al., 2009), there is currently a lack of easy-to-use software for downstream analysis of these data. For some study designs, such as focused analysis of fully penetrant de novo mutations or autosomal recessive inheritance, exome sequence data can be analysed and filtered relatively simply. Increasingly, however, sequence-based approaches are being applied to complex diseases, which are unlikely to follow a simple genetic model, such as autism (Neale et al., 2012), and to more complicated scenarios, such as large pedigrees with incomplete penetrance. These studies require new tools to enable the diverse community of researchers working on such families to interactively and comprehensively analyze next generation sequence data. Figure 1 shows how our new program, Olorin, integrates within-family linkage analysis with exome sequencing in a user-friendly package.

*To whom correspondence should be addressed.
obtain additional information, such as whether a particular individual has been sequenced, via a mouseover popup box. To begin filtering variants, the user first needs to select individuals to be used in searching for shared genomic segments by clicking on them in the pedigree (Fig. 2).

2.2.2 Initial variant filtering After selecting individuals, the user can customize the analysis via a filtering dialog (Fig. 2). First, they set the minimum number of individuals required to share a segment. This enables searches for variants of incomplete penetrance if the threshold is set below the total number of affected individuals in the pedigree. Next, the user can select which information fields from the VCF will be included for subsequent filtering and display. A population frequency cut-off can also be specified at this point if (as is often the case) the study design is focused on variants expected to be rare in healthy individuals.

2.2.3 Dynamic variant filtering Olorin populates an analysis table (Fig. 2) with variants found in the shared segments. This table can be sorted on any column, and variants in the table can be filtered out in real time using a number of filtering tools (Fig. 2), which are dynamically generated based on the user-selected data fields. Olorin can show variants discovered in any or all of these individuals, depending on the genetic model under consideration.

2.2.4 Predicted variant effects Because the ‘consequence’ strings in the VCF information field contain a wealth of parseable information, Olorin supports further processing of two variant consequence string formats: the UK10K analysis pipeline format and the Ensembl Variant Effect Predictor format (McLaren et al., 2010). Because each variant can have multiple consequences, Olorin automatically selects and displays only the most damaging effect for each variant, showing the remainder via a popup box.

3 IMPLEMENTATION

Olorin is written in Java and will work on any platform with Java 1.6 or later installed. The interactive pedigree is drawn using the PedVizAPI (Fuchsberger et al., 2008). The genome-wide sharing plots are generated using source code from the visualization tool, IdeogramBrowser (Müller et al., 2007).

ACKNOWLEDGEMENTS

The authors would like to thank Luke Jostins and Karola Rehnström for their help in discussions regarding features and the user interface.

Funding: Wellcome Trust grant number WT098051.

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. et al. (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics, 30, 97–101.

Bamshad,M.J. et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet., 12, 745–755.

Danecek,P. et al. (2011) The variant call format and VCFtools. Bioinformatics, 27, 2156–2158.

Fuchsberger,C. et al. (2008) PedVizAPI: a Java API for the interactive, visual analysis of extended pedigrees. Bioinformatics, 24, 279–281.

Li,H. et al. (2009) The sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078–2079.

McLaren,W. et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26, 2069–2070.

Müller,A. et al. (2007) Visualization of genomic aberrations using Affymetrix SNP arrays. Bioinformatics, 23, 496–497.

Neale,B.M. et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature, 485, 242–245.