TrioVis: a visualization approach for filtering genomic variants of parent–child trios

Ryo Sakai1,2,*, Alejandro Sifrim1,2, Andrew Vande Moere3 and Jan Aerts1,2
1Department of Electrical Engineering-ESAT, KU Leuven, SCD-SISTA, Leuven 3001, Belgium, 2Future Health Department, iMinds, Leuven 3001, Belgium and 3Department ASRO, KU Leuven, Research[+]Design, Leuven 3001, Belgium

ABSTRACT
Summary: TrioVis is a visual analytics tool developed for filtering on coverage and variant frequency for genomic variants from exome sequencing of parent–child trios. In TrioVis, the variant data are organized by grouping each variant based on the laws of Mendelian inheritance. Taking three Variant Call Format files as input, TrioVis allows the user to test different coverage thresholds (i.e. different levels of stringency), to find the optimal threshold values tailored to their hypotheses and to gain insights into the global effects of filtering through interaction.

Availability: Executables, source code and sample data are available at https://bitbucket.org/biovizleuven/triovis. Screencast is available at http://vimeo.com/user6757771/triovis.
Contact: ryo.sakai@esat.kuleuven.be

Received on December 30, 2012; revised on April 12, 2013; accepted on May 5, 2013

1 INTRODUCTION
Recent advances in massively parallel sequencing technologies, especially sequencing of the entire protein-coding portion of the genome (exome), have introduced new strategies for identifying Mendelian disease genes (Gilissen et al., 2012). Analysis of parent–child trios is one of the strategies for identifying single pathogenic mutations among the thousands to millions of genomic variants. By sequencing, the patient as well as his or her parents, variants can be filtered based on consistency or inconsistency according to the laws of Mendelian inheritance.

Although filtering based on inheritance pattern seems to be straightforward, distinguishing true variation from artefacts and false negatives while retaining sensitivity is a challenging task because of the sequencing error rate and the interdependency of sequencing quality for multiple samples. A previous study because of the sequencing error rate and the interdependency false negatives while retaining sensitivity is a challenging task straight-forward, distinguishing true variation from artefacts and coverage threshold based on the overall coverage and their

...
researcher aims to minimize the number of inconsistent calls while keeping the number of consistent calls high.

The variant frequency sliders (Fig. 1C) visualize the distribution of variants based on variant frequency values. These sliders can be used by the researchers to adjust the ranges for variant frequency for genotyping variants for that sample. By default, any variants with variant frequency >90 are considered alternative homozygous, and any variants with variant frequency between 20 and 89 are considered alternative heterozygous. Any variants <20 are filtered out. The coverage sliders (Fig. 1D) set the coverage thresholds for each sample individually. These sliders also represent the distribution of variants based on coverage values. Finally, the histogram view (Fig. 1E) represents the distribution of consistent and inconsistent variants in stacked bar graphs with coverage values between 1 and 20 for the selected sample. Hovering the mouse over the stacked bar graph highlights cells in the main table, showing where these variants are represented. This view aids the researcher to calibrate the coverage threshold for the selected sample.

The variant data can be investigated under two assumptions: with the ‘migration’ assumption, any variant below the coverage threshold is considered homozygous reference; when this assumption is inactive, variants below the coverage threshold are considered invalid and discarded from the combined set of variants. Filtered results can be exported using the ‘export VCF’ button and saved as VCF files. The researcher can also select the ‘migration’ assumption, any variant below the coverage threshold for the selected sample. Hovering the mouse over the stacked bar graph highlights cells in the main table, showing where these variants are represented. This view aids the researcher to calibrate the coverage threshold for the selected sample.

The variant data can be investigated under two assumptions: with the ‘migration’ assumption, any variant below the coverage threshold is considered homozygous reference; when this assumption is inactive, variants below the coverage threshold are considered invalid and discarded from the combined set of variants. Filtered results can be exported using the ‘export VCF’ button and saved as VCF files. The researcher can also select the ‘migration’ assumption, any variant below the coverage threshold for the selected sample. Hovering the mouse over the stacked bar graph highlights cells in the main table, showing where these variants are represented. This view aids the researcher to calibrate the coverage threshold for the selected sample.

The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073.

Wang, K. et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164.