Bioinformatics, 31(14), 2015, 2377–2379
doi: 10.1093/bioinformatics/btv135
Advance Access Publication Date: 8 March 2015
Applications Note

OXFORD

Genome analysis

Reducing the search space for causal genetic variants with VASP

Matthew A. Field[1,*], Vicky Cho[2], Matthew C. Cook[1,3], Anselm Enders[1,4], Carola G. Vinuesa[1], Belinda Whittle[2], T. Daniel Andrews[1,†] and Chris C. Goodnow[1,5,†]

[1]Department of Immunology, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia, [2]Australian Phenomics Facility, Australian National University, Canberra, ACT 2601, Australia, [3]Department of Immunology, The Canberra Hospital, Canberra, ACT 2605, Australia, [4]Rammaciotti Immunisation Genomics Laboratory, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia and [5]Immunogenomics Group, Immunology Research Program, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
Associate Editor: Alfonso Valencia

Received on November 2, 2014; revised on February 10, 2015; accepted on March 2, 2015

Abstract

Motivation: Increasingly, cost-effective high-throughput DNA sequencing technologies are being utilized to sequence human pedigrees to elucidate the genetic cause of a wide variety of human diseases. While numerous tools exist for variant prioritization within a single genome, the ability to concurrently analyze variants within pedigrees remains a challenge, especially should there be no prior indication of the underlying genetic cause of the disease. Here, we present a tool, variant analysis of sequenced pedigrees (VASP), a flexible data integration environment capable of producing a summary of pedigree variation, providing relevant information such as compound heterozygosity, genome phasing and disease inheritance patterns. Designed to aggregate data across a sequenced pedigree, VASP allows both powerful filtering and custom prioritization of both single nucleotide variants (SNVs) and small indels. Hence, clinical and research users with prior knowledge of a disease are able to dramatically reduce the variant search space based on a wide variety of custom prioritization criteria.

Availability and implementation: Source code available for academic non-commercial research purposes at https://github.com/mattmattmattmatt/VASP.

Contact: matt.field@anu.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

While exome sequencing has been successfully utilized in the discovery of causal variation using small numbers of unrelated individuals (Ng *et al.*, 2010) it is becoming clear that this approach is insufficient to reliably identify the genetic causes in many cases (Yang *et al.*, 2013). Increasingly, genetic variation information from related individuals is being employed to reduce the search space for causal variants, by both the prioritization of variants common to affected individuals and the exclusion of benign variants shared between affected and unaffected individuals. The effective analysis of sequenced pedigrees requires new tools capable of combining variant specific and pedigree wide annotations with powerful filtering options to dramatically reduce the causal variation search space. Current tools focus on either progressively removing variants based

on criteria deemed unlikely to be causal (Li *et al.*, 2012) or by focusing on variants matching specific inheritance models [compound heterozygotes (Kamphans *et al.*, 2013); autosomal dominant (Koboldt *et al.*, 2014)]. Here, we present variant analysis of sequenced pedigrees (VASP), a tool that integrates and summarizes information across a sequenced pedigree without making any assumptions regarding disease inheritance further providing the full integrated pedigree variation information for subsequent prioritization. VASP is standalone software written in Perl derived from our original variant detection pipeline (Andrews *et al.*, 2012).

2 Methods

VASP integrates variant information from each pedigree member and aggregates this information on a per variant basis, describing, among other things, the likely inheritance pattern while simultaneously incorporating external annotation. Furthermore, VASP uses parental allele segregation patterns to determine phasing of variants and identifies genomic blocks common to affected pedigree individuals. VASP and other tools that aggregate information across a pedigree are critical for successful causal variant detection as they automate a complex task too labor intensive to perform manually.

2.1 Input

2.1.1 Input files

Two input files are required; a pedigree (PED) file representing pedigree structure and a variant call format (VCF) file containing variant information. The third required argument is either a variant effect predictor (VEP) annotation file (McLaren *et al.*, 2010) or the path to a local VEP installation when no annotation file is readily available. Optional input files include a text file with the path to individual binary sequence alignment/map (BAM) files used for determining zygosity and inheritance patterns. VASP is able to support any genome build providing all input files refer to the same reference genome.

2.1.2 Input filters

A key feature of VASP is flexible options with prioritization and ordering of variants based on stipulated criteria being specified at the outset. Filtering options include specific inheritance patterns, genomic region, gene name, population allele frequency, phasing information, number of affected/unaffected variant individuals and certain polyphen2 and sorting intolerant from tolerant (SIFT) categories. Combining parameters results in a dramatic reduction in the length of candidate variant lists—and the remaining variants will better correspond to hypotheses about the genetic basis of a particular disease.

2.1.3 Input variant set

Variant callers typically employ a quality score as a cutoff, above which lie a set of presumed high-quality variant calls. In reality, variants above this cutoff often include false positives, whereas variants below this cutoff include true positives (Weisenfeld *et al.*, 2014). To minimize and potentially avoid this problem, the union of all variants called within the pedigree is used as the initial input variant set. This approach allows variant calls at a particular genomic position to be reconciled across the pedigree, potentially correcting calls lying near the cutoff for a single individual.

2.2 External variant annotation

Individual variants are annotated with VEP data including population frequency information from dbSNP (Sherry *et al.*, 2001) and the 1000 genomes project (Genomes Project *et al.*, 2010) as well as SIFT (Kumar *et al.*, 2009) and PolyPhen2 (Adzhubei *et al.*, 2010) scores for estimating the functional effect of missense mutations.

2.3 Pedigree-wide annotation

2.3.1 Genome phasing and de novo mutations

Whenever possible the parental allele inherited by each child is determined, and this information is clustered into genomic blocks of presumed shared inheritance to obtain genome-phasing information. Variants demonstrating segregation differences between affected and unaffected children are further prioritized. In addition, any variant exhibiting Mendelian inconsistencies is annotated, with special attention paid to putative *de novo* mutations (i.e. non-mutant, unaffected parents and a heterozygous offspring).

2.3.2 Disease inheritance patterns and compound heterozygosity

For each variant in the pedigree the inheritance is determined and annotated accordingly. The zygosity of particular variants is preferentially determined from raw sequence data obtained from BAM files using SAMtools (Li et al., 2009) but failing this the required genotype field (GT) and optional allele depth (AD) tags from the VCF file are utilized. Compound heterozygote genes are also annotated, defined as genes containing at least one heterozygous SNV or indel inherited from each parent with unaffected and affected siblings not sharing identical heterozygous variants. These variants must further be heterozygous in all affected individuals and not be homozygous in any unaffected individuals. These compound heterozygous genes are further prioritized in cases where each parent contributes rare or novel alleles.

2.3.3 Gene variability statistics

For each gene three measures of variability are reported; total number of variants, total number of unique variant coordinates and percentage of total transcript bases found variant. Increased gene variability may be relevant to particular diseases but also may be indicative of read alignment issues (often due the presence of gene duplicates) or may indicate the gene is functionally redundant and thus not functionally constrained.

2.4 VASP output and ordering

VASP reports contain all variants detected in at least one pedigree member and categorises variants as either novel, rare (0–2% population frequency), no frequency (known variant but no frequency data available) or common (>2% frequency). For each variant VASP reports both pedigree-wide information (such as inheritance pattern or phasing data) as well as variant-specific information (such as population frequency or polyphen score). By default VASP reports are sorted progressively on four measures: variant category (novel, rare, no frequency and common), the number of variant affected samples (in descending order), the number of unaffected variant samples and lastly the variant population frequency.

3 Results

VASP makes no assumptions regarding the underlying disease transmission mechanism, an apparent strength when compared with similar software (Supplementary Table S1). Instead, VASP provides powerful filters with the aim of allowing researchers to harness their additional knowledge of the disease to generate reduced variant lists

suitable for manual interrogation. One current limitation of VASP is that it can only be run on the command line.

Five pedigrees (Supplementary Table S2) were analyzed to calculate variant segregation statistics with pedigree G1 (Supplementary Figure S1) variant lists (Supplementary Table S3) taken forward to illustrate the effect of various filtering strategies (Supplementary Table S4). To date VASP has been used to analyze 45 pedigrees and found strong candidate causal variants in 15 of these (33.3%). These 15 pedigrees exhibit a wide array of disease transmission mechanisms including autosomal dominant and recessive inheritance, *de novo* mutations, compound heterozygosity and more complex multi-gene cases. This variety in transmission mechanisms within this relatively small group sharing similar diseases illustrates the importance of flexible pedigree analysis software.

We present VASP, a flexible tool for identifying putative causal variants from pedigree sequence data. Through aggregation of data for genetic variants across pedigree members, VASP allows powerful, custom variation prioritization, taking advantage of external datasets and prior knowledge of disease incidence and inheritance patterns. With this tool users have the opportunity to greatly narrow the number of candidate causal variants, using custom criteria, to a size suitable for manual interrogation.

Acknowledgement

We thank the National Computational Infrastructure (Australia) for access to significant computation resources and technical expertise.

Funding

This work was supported by National Health and Medical Research Council Australia Fellowship 585490, National Institutes of Health [grant number U19 AI100627] and Bioplatforms Australia.

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Andrews,T.D. *et al.* (2012) Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biol.*, **2**, 120061.

Genomes Project,C. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Kamphans,T. *et al.* (2013) Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PloS One*, **8**, e70151.

Koboldt,D.C. *et al.* (2014) Exome-based mapping and variant prioritization for inherited Mendelian disorders. *Am. J. Human Genet.*, **94**, 373–384.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,M.X. *et al.* (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**, e53.

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

Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Weisenfeld,N.I. *et al.* (2014) Comprehensive variation discovery in single human genomes. *Nat. Genet.*, **46**, 1350–1355.

Yang,Y. *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, **369**, 1502–1511.