BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1175–1176
doi:10.1093/bioinformatics/btt749

Genome analysis

Advance Access publication January 5, 2014

FamAnn: an automated variant annotation pipeline to facilitate target discovery for family-based sequencing studies

Jianchao Yao[1,*,†,‡], Kelvin Xi Zhang[2,*,†], Melissa Kramer[1], Matteo Pellegrini[3] and W. Richard McCombie[1,*]

[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA, [2]Department of Biological Chemistry, Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095, USA and [3]Department of Molecular, Cellular, and Developmental Biology, University of California, Los Angeles, CA 90095, USA

Associate Editor: Inanc Birol

ABSTRACT

Summary: FamAnn is an automated variant annotation pipeline designed for facilitating target discovery for family-based sequencing studies. It can apply a different inheritance pattern or a *de novo* mutations discovery model to each family and select single nucleotide variants and small insertions and deletions segregating in each family or shared by multiple families. It also provides a variety of variant annotations and retains and annotates all transcripts hit by a single variant. Excel-compatible outputs including all annotated variants segregating in each family or shared by multiple families will be provided for users to prioritize variants based on their customized thresholds. A list of genes that harbor the segregating variants will be provided as well for possible pathway/network analyses. FamAnn uses the *de facto* community standard Variant Call Format as the input format and can be applied to whole exome, genome or targeted resequencing data.

Availability: https://sites.google.com/site/famannotation/home

Contact: jianchaoyao@gmail.com, kelvinzhang@mednet.ucla.edu, mccombie@cshl.edu

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

Received on December 11, 2012; revised on December 9, 2013; accepted on December 20, 2013

1 INTRODUCTION

Recent advances in DNA sequencing technology have led to a resurgence of family-based studies for the discovery of genetic variants, in particular single nucleotide variants and small insertions and deletions, harbored in causal genes that underlie Mendelian and complex diseases. By sequencing exomes or genomes of selected individuals from families, a handful of disease-causing or associated genes have been identified (Bamshad *et al.*, 2011; Boileau *et al.*, 2012; Sullivan *et al.*, 2012). The number of families selected in each study can vary from one to several hundreds, and each family or group of families may follow a different inheritance pattern. Moreover, the sequenced individuals from each family may be present as trios or extended pedigrees. As a result, an easy-to-use automated pipeline would be beneficial for systematically selecting variants segregating in each family and investigating variants shared across families, as well as annotating variants to facilitate user's customized prioritization for target discovery. A number of open-source tools (Lyon and Wang, 2012; Supplementary Table 1) have been developed for annotating and prioritizing variants, but few of them can compare multiple families simultaneously to identify variants recurrently present across families. Most of the current tools do not provide an easy-to-use output so that users can prioritize variants or genes based on their customized thresholds.

Here, we present FamAnn, an automated variant annotation pipeline designed for facilitating disease variants or genes discovery for family-based sequencing studies. The advantages of our pipeline are severalfold. It selects and annotates variants segregating in each family and shared across families. Families with different inheritance patterns can be analyzed simultaneously by indicating the corresponding genetic model in the metadata file. A model for *de novo* mutations discovery is provided as well for users who are interested in identifying *de novo* mutations in trio studies. It is easy-to-use, and one Perl command is sufficient to generate Excel-compatible outputs that retain all annotated variants. Users with limited bioinformatics skills can apply various thresholds, such as allele frequency cutoffs, directly on the output to prioritize variants. When a variant hits multiple transcripts and hence may have different types of functional effects, it outputs all the effects for the same variant to avoid missing critical biological information. We provide functionalities offered by different bioinformatics resources, such as ENCODE annotation, frequency checking in public databases, pathogenicity prediction and conservation scores. Finally, FamAnn can be applied to all types of sequencing data, such as whole-exome sequencing, genome sequencing or targeted resequencing, and can be used to annotate and prioritize any variant calls generated in the *de facto* community standard Variant Call Format (VCF) (Danecek *et al.*, 2011).

2 METHODS

FamAnn was developed using Perl and can be used as a standalone application on diverse hardware and operating systems where standard Perl

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Present address: Group of Computational Genomics & Genetics, Department of Informatics IT, Merck & Co., Inc., Boston, MA 02210, USA.

modules are installed. It uses snpEff (Cingolani *et al*., 2012) or Variant Effect Predictor (VEP) (McLaren *et al*., 2010)-annotated VCF files as inputs where each line corresponds to one genetic variant with annotated genomic location and coding effect. Multiple genetic models are provided for selecting variants segregating in each family, such as autosomal dominant and recessive. There is also a model for *de novo* mutation discovery for trio studies. In addition, a general model that identifies variants shared by affected individuals but absent in unaffected individuals is provided for users who do not want to make any genetic assumption. FamAnn will generate an Excel-compatible output that lists all variants shared by multiple families and variants segregated in each family. If the total number of variants is >1 million, we will split the file into subsets of 1 million variants so that each subset can be analyzed in Excel. FamAnn will also generate a list of genes in TEXT format, which includes all the genes harboring variants shown in the variant output file. Users can use this list as an input for possible pathway or network analyses, such as input for Ingenuity Pathways Analysis.

In the variant output file, a variety of annotations are provided for each variant. For example, loss-of-function mutation annotation is provided based on snpEff function annotation prediction. Compound heterozygous mutations in the affected individuals are identified in the trio studies. To identify variants in regulatory regions such as enhancers or promoters, FamAnn retrieves ENCODE annotations by using histone modification tracks obtained from University of California, Santa Cruz Table Browser, such as mono- and tri-methylation of histone H3 lysine 4 (H3K4me1, H3K4me3) and acetylation of histone H3 lysine 27 (H3K27ac). In addition to histone modification annotation, FamAnn also marks variants that fall in the predicted enhancers by using the DNaseI hypersensitivity track (DNase clusters). To facilitate prioritizing variants based on their allele frequencies in public databases, FamAnn extracts variant frequencies from the 1000 Genomes Project (http://www.1000genomes.org/) and Exome Variant Server (http://evs.gs.washington.edu/EVS/). To predict whether a variant is in a duplicated region defined by the Segmental Duplication track obtained in the University of California, Santa Cruz Table Browser, FamAnn marks it as 'yes' in the 'Segmental_dups' column in the output file if the variant is present in a duplicated region. To predict protein disruption and conservation, FamAnn uses the dbNSFP database (Liu *et al*., 2011) to aggregate scores of SIFT (Kumar *et al*., 2009), PolyPhen-2 (Adzhubei *et al*., 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz *et al*., 2010), GERP++ (Cooper *et al*., 2010), PhyloP (Cooper *et al*., 2005) and SiPhy (Garber *et al*., 2009) in the common output file. Therefore, users may prioritize their variants in Excel using the variant output file and customized filtering procedures, such as the recurrent frequency of the variant present in multiple families, the cutoff of allele frequency in the 1000 Genomes project or mutation type of each variant.

To facilitate the use of our pipeline, all the input datasets and the annotation tracks will be stored in one directory. To run FamAnn in the same directory, users need to generate a metadata file in TEXT format to include the family IDs and sequenced individual IDs and their affected status and the model they want to apply to each family. The detailed framework can be found in Figure 1, and a manual of our pipeline can be found in the section 'FamAnn Manual' in the Supplementary Information.

3 RESULTS

To evaluate the performance of FamAnn, we tested it on both real exome-sequenced family datasets (Boileau *et al*., 2012) and synthetic whole-genome-sequenced family datasets. Consistent results between the expected and observed outcomes demonstrate the accuracy of our pipeline. The details of these tests can be found in the section 'Performance Evaluation for FamAnn' in the Supplementary Information.

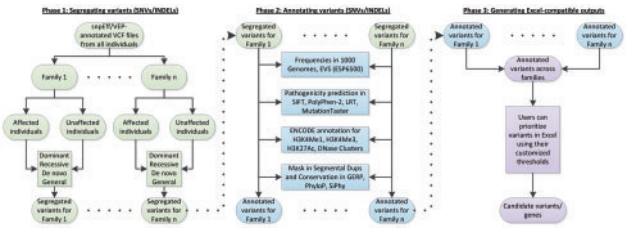

Fig. 1. A framework for annotating variants to facilitate target discovery for family-based sequencing studies by using FamAnn

4 CONCLUSION

In summary, FamAnn offers a combination of unique advantages in variant annotation to facilitate target discovery for family-based sequencing studies. It can be applied to variant discovery for Mendelian or complex disease studies in which whole exome or genome or targeted resequencing is performed.

ACKNOWLEDGEMENTS

The authors thank Dr Dianna M. Milewicz and Dr Dong-Chuan Guo for generously sharing the real exome-sequenced VCF files.

Funding: A grant from T. and V. Stanley.

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.

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

Boileau,C. *et al*. (2012) TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. *Nat. Genet*., **44**, 916–921.

Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res*., **19**, 1553–1561.

Cingolani,P. *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Cooper,G.M. *et al*. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*., **15**, 901–913.

Cooper,G.M. *et al*. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.

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

Garber,M. *et al*. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.

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.

Liu,X. *et al*. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat*., **32**, 894–899.

Lyon,G.J. and Wang,K. (2012) Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Med*., **4**, 58.

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

Schwarz,J.M. *et al*. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.

Sullivan,P.F. *et al*. (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet*., **13**, 537–551.