Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

SAMFIRE: multi-locus variant calling for time-resolved sequence data

C. J. R. Illingworth [1,*]

[1] Department of Genetics, University of Cambridge, Cambrdge CB2 3AS, United Kingdom

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: An increasingly common method for studying evolution is the collection of time-resolved short-read sequence data. Such datasets allow for the direct observation of rapid evolutionary processes, as might occur in natural microbial populations and in evolutionary experiments. In many circumstances, evolutionary pressure acting upon single variants can cause genomic changes at multiple nearby loci. SAMFIRE is an open-access software package for processing and analysing sequence reads from time-resolved data, calling important single- and multi-locus variants over time, identifying alleles potentially affected by selection, calculating linkage disequilibrium statistics, performing haplotype reconstruction, and exploiting time-resolved information to estimate the extent of uncertainty in reported genomic data.
Availability and Implementation: C++ code may be found at https://github.com/cjri/samfire/.
Contact: chris.illingworth@gen.cam.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Time-resolved genome sequence data provides direct information about evolutionary processes. At low recombination rates, populations evolve via clonal competition, whereby linkage effects spanning multiple loci determine the outcome of adaptation (Neher and Shraiman, 2009). The Samfire software extracts and processes multi-locus information from time-resolved genomic data, so as to shed light on underlying patterns of evolution.

Samfire includes a sequence processing pipeline **filter** that performs quality control on short read data in aligned SAM format. Similar to a previous approach to variant calling (Watson *et al*., 2013), individual reads are filtered by median base quality, trimming sequence in order to achieve this, subject to a minimum read length. Individual nucleotides that fall below a given quality threshold are then removed, being replaced by blank "-" alleles, to indicate a lack of available information. Duplicate reads may be removed, before paired-end reads are joined, any interval between paired reads being spanned by blank alleles. The pipeline **sl_traj** then calls single-locus polymorphisms, filtering by the minimum number of reported variant alleles, minimum observed allele frequency, and a statistical measure, assessing the probability of a polymorphism arising through sequencing error. Noting that across time, variants can arise or die out in a population through mutation and selection, observed allele counts are reported for polymorphic loci across all time-points, even if a polymorphism is found only in a single sample; we refer to these collected observations as the trajectory of an allele. Alternatively, observation in multiple samples may be specified as a condition for calling a polymorphism.

Having identified polymorphisms, the pipeline **sl_noise** utilises the temporal resolution of the data to estimate the extent of noise inherent in the variant frequencies. Excluding trajectories in which an exceptional amount of allele frequency change is observed, changes in allele frequencies in other trajectories are assumed, conservatively, to arise from noise alone. Under this assumption, a Dirichlet multinomial distribution is characterised, obtaining the optimal extent of variance required to account for the observed deviations of each trajectory from constancy. While approximate Bayesian methods may be used to characterise evolutionary behaviour without such fitting (Foll *et al*., 2014), estimation of a likelihood function allows for techniques such as model selection to be used in discriminating between potential evolutionary scenarios.

In an optional step within the code, the pipeline **sl_neutrality** allows the identification of trajectories exhibiting potentially non-neutral evolutionary behaviour, using a simple model of allele frequency change to identify trajectories potentially under constant or time-dependent selection. Where multi-locus models of selection are evaluated, the

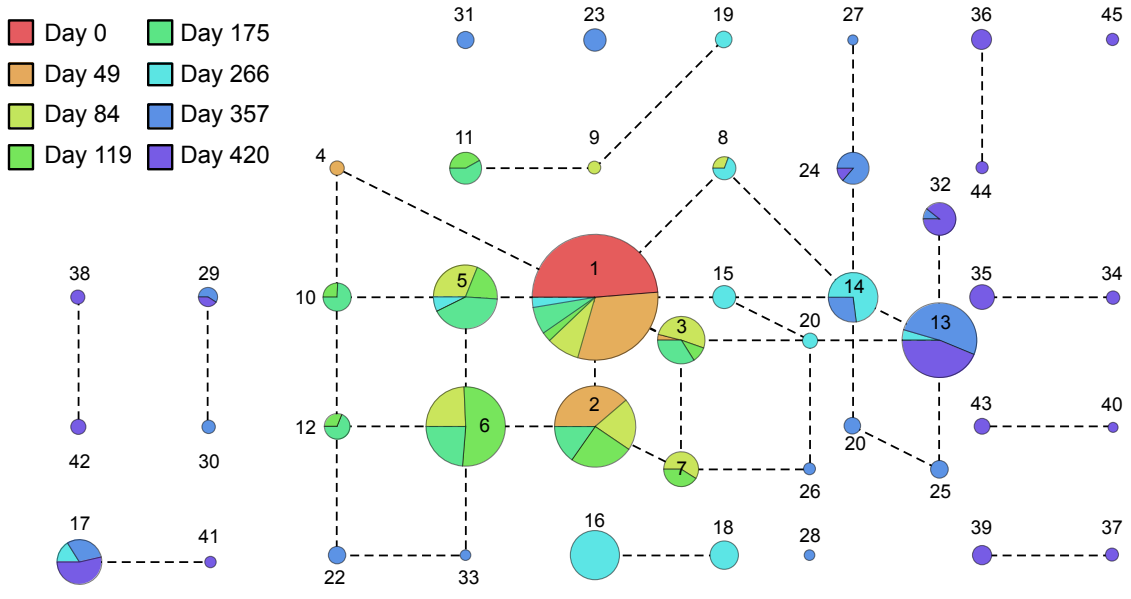

Fig. 1. Increasing genetic diversity of the V3 loop of HIV envelope protein observed via calling of time-resolved multi-locus haplotypes. Numbered pie charts show the proportion of each haplotype observed from viral data collected from a single patient. The area of a segment in a single chart is proportional to the fraction of sequences observed to have the given multi-locus haplotype at each time. Dotted lines connect multi-locus haplotypes that differ by a single mutation

reduction of the model space to a smaller number of loci can improve computational efficiency (Illingworth *et al*., 2014).

Having identified single-locus polymorphisms, the pipeline **call_ml** identifies multi-allele variants spanning these loci. Sets of loci that could be spanned by paired-end reads are efficiently found. Multi-locus variants are filtered by the minimum number of observations of a variant and variant frequency, then reported across all time-points. Optionally within this pipeline, full haplotypes may be reconstructed from the multi-locus reads via a rule-based procedure (see Supporting Information). While other approaches to haplotype reconstruction have used model selection to infer a minimum number of haplotypes required to explain a dataset (Töpfer *et al*., 2013; Fischer *et al*., 2014), we here adopt a maximal approach, constructing a set of haplotypes that are guaranteed to explain all of the filtered multi-locus sequence observations without further error. The pipeline **ml_noise** infers the extent of noise in the observed multi-locus sequence data, again optimising a Dirichlet multinomial likelihood function. Time-resolved data has elsewhere been used to infer haplotype frequencies under a dynamic model of adaptation (Illingworth, 2015). Given variants spanning both multiple, and single loci, the pipeline **ld_calc** calculates estimates of linkage disequilibrium between loci.

Figure 1 shows an example of data generated by applying SAMFIRE to within-host data from time-resolved sampling of HIV in a single patient (Gall *et al*., 2013), here showing the expansion of viral diversity over time as expressed in one set of 14-locus partial haplotypes. The code is highly flexible, allowing for multiple user choices in evaluating data. With the increase in the collection of genomic data, we hope that Samfire will be of value to researchers aiming to build a greater understanding of rapid evolutionary processes.

Funding

CI was supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (Grant Number 101239/Z/13/Z).

References

Fischer, A., *et al*. (2014) High-Definition Reconstruction of Clonal Composition in Cancer, *Cell Reports*, **7**, 1740-52.

Foll, M., *et al*. (2014) Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective, *PLoS Genetics*, **10**, e1004185.

Gall, A., *et al*. (2013) Restriction of V3 region sequence divergence in the HIV-1 envelope gene during antiretroviral treatment in a cohort of recent seroconverters, *Retrovirology*, **10**, 8.

Illingworth, C. J. R., *et al*. (2014) Identifying Selection in the Within-Host Evolution of Influenza Using Viral Sequence Data. *PLoS Comp. Biol.*, **10**, e1003755.

Illingworth, C. J. R. (2015) Fitness inference from short-read data: within-host evolution of a reassortant H5N1 influenza virus, *Mol. Biol. Evol.*, **11**, 3012-26.

Neher, R. A., and Shraiman, B. I. (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proc. Natl. Acad. Sci.*, **10**, e1003755.

Töpfer, A., *et al*. (2013) Probabilistic Inference of Viral Quasispecies Subject to Recombination, *J. Comp. Biol.*, **20**, 113-123.

Watson, S. J., *et al* (2013) Viral population analysis and minority-variant detection using short read next-generation sequencing, *Phil. Trans. Roy. Soc B*, **368**, 20120205.