Bioinformatics, 31(17), 2015, 2877–2878
doi: 10.1093/bioinformatics/btv271
Advance Access Publication Date: 25 April 2015
Applications Note

Sequence analysis

kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome

Shea N. Gardner[1], Tom Slezak[1] and Barry G. Hall[2,*]

[1]Computations/Global Security, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA and [2]Bellingham Research Institute, Bellingham, WA 98229, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 13, 2015; revised on April 22, 2015; accepted on April 23, 2015

Abstract

Summary: We announce the release of kSNP3.0, a program for SNP identification and phylogenetic analysis without genome alignment or the requirement for reference genomes. kSNP3.0 is a significantly improved version of kSNP v2.

Availability and implementation: kSNP3.0 is implemented as a package of stand-alone executables for Linux and Mac OS X under the open-source BSD license. The executable packages, source code and a full User Guide are freely available at https://sourceforge.net/projects/ksnp/files/

Contact: barryghall@gmail.com

1 Introduction

Next-generation DNA sequencing has led to an explosion in the number of sequenced genomes, especially microbial genomes. SNP analysis of those genomes is important in forensic investigations, strain identification, outbreak tracking, phylogenetic analysis and identifying strain differences that are important to phenotypes such as virulence and antibiotic resistance (Gardner and Hall 2013). The availability of genome sequences of up to thousands of genomes of a single species requires tools that can handle very large datasets in reasonable times (Bertels *et al*., 2014; Gardner and Hall, 2013). Because microbial genomes are subject to massive gene gains and losses, insertion, deletions and rearrangements, alignment of whole microbial genome sequences has proven to be computationally intensive and not applicable to datasets of hundreds or thousands of genomes. The increasing tendency to leave genome sequences at the assembly, or even raw unassembled reads stage has necessitated the development of genome analysis tools that do not require genome alignments, and that can be applied to genomes at all stages of completion.

kSNP v2, released in 2013, is a tool that can identify SNPs in hundreds of microbial genomes without the requirement for genome alignment or a reference genome (Gardner and Hall, 2013). If one or more annotated genomes are included, kSNP can automatically annotate the identified SNPs. kSNP v2 estimates phylogenetic trees by parsimony, Neighbor-Joining and Maximum Likelihood methods, and reports trees with a variety of node labels, including the number of SNPs unique to each node. kSNP can analyze finished (closed) genomes, draft genomes at the assembly stage, genomes at the raw reads stage, or any combination of those stages. kSNP v2 proved to be relatively fast, analyzing 288 genomes of the Rhabdoviridae family in 2.5 h, and analyzing 119 *Escherichia coli-Shigella* genomes in 14.4 h on a Linux cluster or 20.3 h on a desktop iMac.

kSNP v2 has been used for SNP identification and phylogenetic analysis in several recent studies (Forde *et al*., 2014; Raphael *et al*., 2014; Smith *et al*., 2015) and has proven itself a valuable tool for identifying SNPs that predict phenotype (Hall, 2014).

Here we announce the release of kSNP3.0, a significantly improved version of kSNP v2. The kSNP3.0 package also includes a number of new tools and utilities that facilitate the downloading of genomes for kSNP3.0 analysis, preparation of input files, and a variety of post-kSNP run analyses of the output files.

2 Features

2.1 Annotations

kSNP3.0 annotates SNPs by automatically downloading the Genbank annotations of genomes that are identified as annotated in

the input files. The genome sequence input file format of kSNP v2 resulted in annotation being limited to SNPs that were within the chromosome; thus SNPs in plasmids or other replicons were identified as not being in an annotated region. This was a particularly serious problem for genomes such as the *Vibrio* genomes that consist of two chromosomes, in which case almost half of the SNPs would be unannotated, but a considerable amount of annotation was also lost in strains with multiple plasmids. kSNP3.0 now uses a different input format that provides annotation of all replicons.

2.2 Parsimony trees are consensus trees

In kSNP v2 the parsimony tree was a random tree from among the equally most parsimonious trees. kSNP3.0 now reports a consensus of the equally most parsimonious trees.

2.3 Input file format

The input file of genome sequences for kSNP v2 was a single file of all of the genome sequences in fasta format. That format meant that genome sequences had to be stored twice, once as the original file and once as part of the input fasta file. With datasets of hundreds of genomes the input file could be huge, larger than many text editors could open, making it difficult to add or remove genomes from the file. Generating a new dataset that was a subset of an existing set required more duplication and thus more file storage space. The kSNP3.0 input file is just a list of the paths to the original genome files, thus adding or removing genomes or extracting subsets of a data file requires only editing that list. The new format is particularly helpful when dealing with raw read genomes whose file sizes can easily exceed 500 MB. kSNP3.0 automatically detects raw-read files and correctly incorporates them.

2.4 Option to add genomes to an existing run

Adding one or more genomes to a dataset required completely rerunning kSNP v2. kSNP3.0 includes a provision for appending a new genome to an existing run, a process that can take only a few percent of the time required to repeat the run.

2.5 kSNP3.0 is faster

Improvements in speed come from more flexibility in the annotation process. Annotation is a costly process in terms of time (Table 1). kSNP3.0 provides for two annotation modes: standard annotation in which SNPs are annotated based only on the first genome in the list of annotated genomes, and full annotation, in which each SNP is annotated based on each annotated genome in which the SNP is present. Standard annotation is considerably faster and is sufficient for most purposes. kSNP 3.0 identified the same number of SNPs and estimated the same parsimony phylogenetic tree as kSNP v2. Its accuracy is thus the same as that of kSNP v2.

2.6 Utility programs

The kSNP3.0 package includes several small utility programs.

Table 1 Times required for kSNP analysis of 20 *E.coli* genomes

Program	Conditions	Time (h)
kSNP v2	Default (no annotation)	1.04
kSNP3.0	Default (no annotation)	0.89
kSNP v2	Annotation	11.04
kSNP3.0	Standard annotation	2.92
kSNP3.0	Full annotation	11.14

1. Programs that automate the downloading of finished and assembled genomes as files suitable for input into kSNP3.0 (FetchFinishedGenomes and FetchGenomeAssemblies).
2. A program to automatically make a kSNP3.0 input file list from the set of genome files (MakeKSNP3infile).
3. A program that determines the optimum kmer size for the dataset and that calculates FCK, a measure of diversity of sequences in the dataset (Kchooser)
4. Programs that extract locus numbers for SNPs files, that remove node name labels from trees, and that find the SNP loci that map to a particular node on a tree (extract_nth_locus, rm_node_names_from_tree3, select_node_annotations3).

3 Comparison With Other Programs

kSNP3.0 was developed as a solution to the problem of aligning large numbers of microbial genomes. Parsnp (Treangen *et al*., 2014) is an alternative solution to the problem, but Parsnp only aligns the core genomes and requires finished or assembled genomes. Unassembled genomes (raw reads) present a particular problem. Epstein *et al*. (2012) maps raw read seqeunces to a complete (closed) genomes to identify homologous sites, and those sites are concatenated to generate an alignment that is analyzed by the usual methods. That approach only identifies sites that are present in the chosen reference genome. Another method, RealPhy, maps raw reads to a set of several reference genomes, thus increasing the likelihood of using all of the information in the raw-read genomes for analysis. RealPhy depends on accurate mapping of raw reads (or contigs) to the reference genomes, and when some taxa are diverged by > 5–10% the distances to the reference genome are under estimated, leading to incorrect topologies (Bertels *et al*., 2014). Because kSNP 3.0 does not require reference genomes and because it can use raw read files kSNP 3.0 a niche that other programs do not.

Acknowledgements

kSNP3.0 was developed under internal funding at LLNL.

Conflict of Interest: none declared.

References

Bertels,F. *et al*. (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.,* **31**, 1077–1088.

Epstein,B. *et al*. (2012) Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S.* medicae. Plos Genet., **8**, E1002868.

Forde,B.M. *et al*. (2014) The complete genome sequence of *Escherichia coli* Ec958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-St131 clone. *Plos One,* **9**, E104400.

Gardner,S.N. and Hall,B.G. (2013) When whole-genome alignments just won't work: Ksnp V2 Software for alignment-free Snp discovery and phylogenetics of hundreds of microbial genomes. *Plos One,* **8**, e81760.

Hall,B.G. (2014) Snp-associations and phenotype predictions from hundreds of microbial genomes without genome alignments. *Plos One,* **9**, E90490.

Raphael,B.H. *et al.* (2014) Distinguishing highly-related outbreak-associated *Clostridium botulinum* type A(B) strains. BMC Microbiol., **14**, 192.

Smith,T.J. *et al*. (2015) Genomic sequences of six Botulinum neurotoxin-producing strains representing three Clostridial species illustrate the mobility and diversity of Botulinum neurotoxin genes. *Infect. Genet. Evol.*, **30**, 102–113.

Treangen,T.J. *et al*. (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.*, **15**, 524.