Journal of Forensic Research - Open Access

www.omicsonline.org

Research Article
OPEN ACCESS Freely available online
doi:10.4172/2157-7145.1000107

Scalable SNP Analyses of 100+ Bacterial or Viral Genomes

Shea N. Gardner* and Tom Slezak

Lawrence Livermore National Laboratory, Livermore, Computations/Global Security, PO Box 808, L-174, CA 94551

Abstract

With the flood of whole genome finished and draft microbial sequences, analysts need faster, more scalable bioinformatics tools for sequence comparison. An algorithm is described to find single nucleotide polymorphisms (SNPs) in whole genome data. It scales to hundreds of bacterial or viral genomes, and can be used for finished and/or draft genomes available as unassembled contigs. The method is fast to compute, finding SNPs and building a SNP phylogeny in seconds to hours. It identified thousands of putative SNPs from all publicly available Filoviridae, Poxviridae, foot-and-mouth disease virus, Bacillus, and Escherichia coli genomes and plasmids. The SNP-based trees it generated were consistent with known taxonomy and trees determined in other studies. The approach described can handle as input hundreds of megabases of sequence in a single run. The algorithm kSNP is based on k-mer analysis using suffix arrays and requires no multiple sequence alignment.

Keywords: Single nucleotide polymorphism (SNP); Software; Microbial forensics; Genotyping; Phylogeny; Genome analysis

Introduction

Single nucleotide polymorphisms can aid in phylogenetic characterization of bacterial and viral isolates, tracking strains during an epidemic, forensic investigations, and correlating genotype to phenotype. Advanced sequencing technologies deliver dozens or even hundreds of microbial sequences at feasible costs. Bioinformatics for whole-genome analyses may bottleneck our ability to make sense of the flood of sequence data without rapid and scalable algorithms. This work describes a method to find SNPs and build phylogenies for large numbers of finished sequences and/or assembled draft contigs, and presents examples for a number of bacteria and viruses. It can handle many genomes at once, for example, all the available genomes in a viral family or a bacterial genus, and has been used to find thousands or millions of putative SNPs from hundreds of megabases of target sequences. No attempt is made to distinguish sequencing errors from SNPs, although the analysis results can be used to design assays to do so using cost-effective methods such as microarrays or sequencing of short specific regions. We work with all available assembled genomic sequence (from any platform) and do not assume that "raw" data is available that might potentially resolve sequencing or assembly errors on any particular genome. The goal here differs from that of other software like SOAPsnp [1] and others (see references in [1]), in that kSNP finds SNPs from among many assembled genomes and builds a SNP-based phylogeny, while SOAPsnp finds SNPs in unassembled raw sequencing read data from a single genome relative to another assembled reference.

Usually SNP finding begins with a multiple sequence alignment or many pairwise sequence alignments of a set of target sequences. We have been hard pressed to find software to keep pace with the memory required to build accurate alignments for dozens to hundreds of genome-length sequences in a feasible time frame. Instead, here we take advantage of fast, memory-efficient suffix array methods [2] and BLAST+ [3] to find putative SNP loci, and build a tree by either maximum likelihood [4] on a SNP allele matrix or neighbor joining [5] on a SNP hamming distance matrix. We string together these publicly available tools with a few short PERL scripts and Unix commands. Code is available on request from the authors.

Other studies limit the region(s) examined to a few genes or areas with known sequence variation. The kSNP approach scales to examine mutations across whole genomes, which should help to uncover novel regions that correlate with phenotype outside of well-characterized genes or non-coding sequence. It should also be useful in horizontal gene transfer studies, since one can examine SNPs across the entire genome. Although beyond the scope of this paper, microarrays with probes designed for all putative SNPs can be used to experimentally validate SNP alleles, identify sequencing errors, and characterize SNP alleles in unsequenced isolates to place them on a phylogeny (manuscript in preparation).

Methods

The process is diagrammed in Figure 1. First, we enumerate all k-mer oligos in the set of input sequences, or targets; conceptually this is all the subsequences from sliding a window of length k across the targets, stepping by one base. This k-mer enumeration can be efficiently performed with the suffix array code from [2]. We used oligos of k=25. Reverse complements of oligos are added to the list, so each oligo is represented in both directions to account for cases in which sequences have an inversion or report opposing strands, but

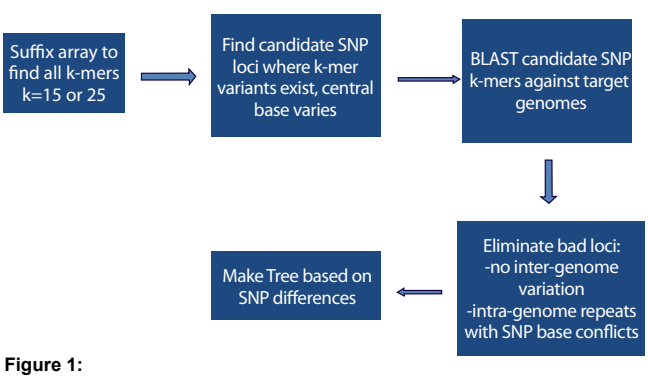

Figure 1:

***Corresponding author:** Shea N. Gardner, Lawrence Livermore National Laboratory, Livermore, CA 94551, Tel: 925-422-4317; Fax: 925-423-6437; E-mail: Gardner26@llnl.gov

Received November 25, 2010; **Accepted** December 27, 2010; **Published** December 30, 2010

Citation: Gardner SN, Slezak T (2010) Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. J Forensic Res 1:107. doi:10.4172/2157-7145.1000107

removing any duplicate oligos. We look for any candidate SNP base in the center, at position 13, by counting the number of oligos with the same up- and down-stream sequence surrounding the central position. For example, when k=25, the surrounding sequence is the 12 bases on both sides of the 13th base. If a surrounding sequence occurs more than once in the list of oligos from all targets, it represents a candidate SNP locus at the 13th position. Although alternative values of k also work, we chose k=25 for most of our calculations since oligos of this length are not frequently repeated within a genome, so it is usually suitable to uniquely characterize a locus. Moreover, it is amenable to SNP assay development such as for microarray probes. We omit cases where all 12 bases on either side of the candidate SNP are a homopolymer repeat. The process of finding candidate SNP k-mers can be parallelized by splitting up the candidate loci by the identity of their first few starting nucleotides. This representation of a SNP locus is based on surrounding sequence information rather than positional information in a genome, and it differs from traditional concepts of a SNP locus. It allows us to consider draft genomes which are available only as contig fragments in which positional information relative to the complete genome is not known.

Next, we prune out the putative SNPs from the candidate loci, eliminating any where the surrounding sequence occurs more than once but with a different central base (i.e. candidate SNP allele) in the same target genome, and also eliminating any where the SNP allele does not vary among the target genomes. We did not eliminate candidates that occur more than once in a genome with the same central base, since this allows us to use draft sequence with imperfect assembly and perhaps to handle unassembled reads that contain repeated sequence. We did this by representing each candidate locus by a blast query string containing two 12-mers with an "N" between them to indicate a variable base. Any sequence that is a reverse complement of another already on the candidate list is now eliminated from the queries, since BLAST will report both plus and minus strand hits. This is important since the targets might contain draft contig data in the minus direction. These candidate loci are the queries BLASTed against a database of the original input sequences. We used the BLAST+ [3] blastn algorithm with the following parameters: -task blastn-short -outfmt '7 std qseq sseq' -word_size 12, and parsed the hits to eliminate those with conflicting mutations at the "N" position within a target sequence or with no variation among targets. If a different value of k is used, some of these settings need to be changed: for example, with k=15, we used –word_size 7 –evalue 100. The BLAST output also gives relevant information about SNP position and orientation in each target genome. We do not require that there be a hit in every target, which is a crucial allowance if we wish to include SNP loci from regions that may be deleted or show greater variation in one of the targets, or are working with draft data in which there might be sequencing gaps. If the 24 bases of sequence surrounding a SNP position (12 on either side) are not present in a target, the locus is considered absent in that target.

The SNP-based Hamming distance between each pair of targets is computed as the number of loci at which their alleles differ. We used the neighbor algorithm of PHYLIP [5] to build an unrooted tree from the pairwise distance matrix, and visualized trees using Dendroscope. [6] We do not claim that these trees are the most phylogenetically accurate, but they are scalable and fast to compute for such large analyses, and seem to give reasonable trees even in cases with many missing loci. We also built SNP matrices by creating a SNP sequence alignment listing the alleles in each genome, and used RAxML v7.2.7 [4] to create a maximum likelihood (ML) tree. The ML trees do not generate accurate trees when there are many clade-specific (e.g. species specific) loci that appear to be missing in other sequences, so ML trees are poor for highly diverse target sets, although they do better for analyses of a single species. The SNPs and Newick tree files are available by request from the authors.

Analyses were performed on several example data sets. These contained complete genomes and plasmids, both finished and draft, downloaded from NCBI nt, reference, and genome projects databases, Baylor College of Medicine Sequencing Center, J. Craig Venter Institute, and the DOE Joint Genome Institute. Sequence data for Escherichia coli and foot-and-mouth disease virus (FMDV) were downloaded in June 2010. Bacillus sequences were downloaded in April 2010. Filoviridae and Poxviridae sequences were downloaded in May 2010. All the data discussed in the results are available by contacting the authors. Calculations were performed on an AMD Opteron node with eight 2.4 GHz processors. A node with 16 GB of available memory was used for the virus runs, and 32 GB of memory for the bacteria.

Results

Table 1 summarizes the kSNP analyses performed. From thousands to millions of SNP loci were found for each target group. Target set sizes ranged from 47 to 184 genomes, and up to 539 MB of sequence data. For the viral and plasmid targets, the analyses completed in seconds or minutes, and the bacterial genome targets completed in 2-14 hours. We will not present an in-depth analysis of the phylogenetics or examine genotype by phenotype correlations, as that is better performed by subject matter experts for each organism. Instead, we will illustrate results of this approach with examples for a wide array of organisms with substantial amounts of sequence data.

Filoviridae

Marburg and Ebola do not share any SNP loci. The Ebola genomes cluster into distinct species groups with a hamming distance tree (Figure 2A (included as supplementary data)). The Marburg Angola sequences also form a single clade, as do the Marburg Ravn sequences. However, the Marburg Uganda and DRC sequences fall into two distinct branches sequences, one very similar to the Ravn

	Number of sequences	Target set size	Number of SNP loci	Time to complete	Number of sequences that cannot be uniquely resolved
Filoviridae	47	884 KB	3,042	45 seconds	0
FMDV	245	2.0 MB	14,060	22 minutes	0
Poxviridae	117	21 MB	29,527	29 minutes	6
Bacillus genus genomes	107	539 MB	1,611,817	14 hrs, 8 minutes	2
Bacillus genus plasmids	113	9.4 MB	9,284	7 minutes	14
Escherichia coli genomes	63	316 MB	342,701	2 hrs, 36 minutes	0
Escherichia coli plasmids	123	7.5 MB	13,443	9 minutes	2

Table 1:

sequences, and the other closely grouped with a South African "Ozolin" and distantly grouped with a handful of sequences including pp3/4 guinea pig variants, Musoke, Popp, and Ci67. There are 271 homoplastic SNPs that do not conform to the branches of the tree. Recombination is known to occur in Filoviridae [7], and may contribute to the presence of homoplastic SNPs. Running the analysis with k=15 gives 4,725 SNP loci (compared to ~3K with k=25), 661 homoplastic loci, and an identical hamming tree.

The ML tree based on the SNPs matrix from the 25-mer analysis mixes the Sudan and Cote D'Ivore, and Reston and Bundibugyo sequences (Figure 2B (included as supplementary data)), while the Marburg sequences are classified according to the same topology as the hamming tree so are not shown. The 25-mer ML tree appears to be less accurate due to the preponderance of loci that appear to be "missing" in some genomes due to sequence variation surrounding the SNP that affects our method of distinguishing loci. The ML tree from the 15-mer analysis (Figure 2C (included as supplementary data)) is much more similar to the hamming tree. Simple hamming distance trees are less susceptible to biases from such "missing" loci than are maximum likelihood trees, so are perhaps more appropriate when the analysis contains highly diverse sequences from different species.

FMDV

Consistent with previous studies of multiple genes [8], members of the same serotypes do not always cluster together, according to the kSNP hamming tree (Figure 2A (included as supplementary data)) or the ML tree (Figure 2B (included as supplementary data)), although the hamming tree conforms to serotypes somewhat more closely than the ML tree, particularly for Asia and C serotypes . Analyses with k=15 are shown. The SNP tree indicates that in serotype O, the O_UKG 2001 sequences are closely related to one another, as are the O_UKG 2007 sequences, but these two groups are not closely related to each other. The O_UKG 2007 sequences cluster with some serotype O and A sequences from South America, while the O_UKG 2001 sequences lie closer to some Asian sequences of serotype O and Asia1. Other serotypes are likewise dispersed across the SNP tree. Only SAT1, SAT2, and SAT3 sequences cluster as a single SAT clade, although the three SAT serotypes are mixed up within the cluster. These analyses point to the difficulty of making a nucleotide-based assay for serotype, since the SNP data are consistent with the known pattern that serotype and genotype are not tightly correlated across much of the FMDV genome. Previous studies have shown that analysis of just the VP1 gene, which codes for the antigenic outer capsid, does cluster the serotypes into distinct lineages. [8].

Using the oligo length of k=25, all 245 target sequences could be uniquely resolved, and ~14K SNP loci were found. With k=15, we found 16,992 SNP loci, and again all genomes could be uniquely resolved. The tree for k=15 clustered all the genomes according to serotype somewhat better than k=25 (not shown, but trees are available from the authors), particularly the Asia1 and C serotypes, although there were a higher fraction of homoplastic loci (SNPs that do not conform to the ML tree) for k=15 than k=25 (26% homoplastic for k=15, 21% homoplastic for k=25).

Poxviridae

The Poxviridae kSNP analyses cluster the 117 genomes by species and strain, as expected (Figure 3 (included as supplementary data)). The phylogeny is virtually identical one for 53 strains recently determined by poxvirus experts at the US Centers for Disease Control and Prevention based on sequence alignments of 9 genes. [9] The variola sequences are split into the major and minor groups as in previous SNP analyses [10], and camelpox and taterapox cluster as nearest neighbors of the variolas. Rabbitpox is very similar to other vaccinia sequences, and their nearest neighbors appear to be horsepox and one strain of cowpox (GRI-90). Slightly more distance branches are the clade of monkeypox sequences and a couple more cowpox strains and ectromelia. Outside the Orthopox branch, the other genomes also cluster by species. This is not a rooted tree, so it should not be used to interpret ancestral versus derived sequences.

Out of the 117 genomes, only three pairs of sequences cannot be resolved using these SNPs: two of the vaccinia Modified Virus Ankara (MVA) sequences VAC_MVA-572 and VAC_MVA-BN; two of the variola sequences from Bangladesh VAR_Bangladesh1974_Shahzaman and VAR_Bangladesh1974_nur_islam; and another two of the variola sequences VAR_India1953_NewDelhi and VAR_Japan1946.

The ML tree (not shown) incorrectly places species relative to one another. This emphasizes the previous point that missing SNPs unduly affect ML trees, so that when diverse sequences of different species are being analyzed together, hamming trees may give more reasonable trees due to the effects of species-specific loci.

Bacillus

Three of the species, anthracis, subtilis, and licheniformis, are very homogeneous within the available genomes for that species, although relatively few licheniformis sequences are currently available (Figure 4 (included as supplementary data)). In contrast, cereus, thuringiensis, and other species show substantially more intraspecific variation, and may result from the challenge of placing a new isolate into a taxonomic group when sequence data is limited. The node containing the anthracis genomes is distinguished by 5,173 species-specific SNPs (all anthracis genomes share the same allele at ~5K SNP loci), which could be useful for developing signatures for specific detection of anthracis. There is one pair of genomes in this set that cannot be resolved based on SNPs: anthracis_Ames_Ancestor and anthracis_A0248. According to the genome project information at NCBI for this strain (genome project ID 33543), "This strain (96-10355; K1256) is a human isolated from USAMRIID, Ohio", and it was sequenced at the Los Alamos National Laboratory. The Ames Ancestor strain is the type strain (0581, A2084, genotype 62, Group A3.b) for *Bacillus anthracis*, and is considered the "gold standard" according to the genome project information at NCBI (genome project ID 10784), and was sequenced at The Institute for Genomic Research in Rockville, MD. So these are different but similar

Cluster Number	Sequence
1	cereus_plasmid_pCER270
1	cereus_AH187_plasmid_pAH187_270
2	A2012_plasmid_pXO1
2	A0248_plasmid_pXO1
3	thuringiensis_miniplasmid
3	thuringiensis_canadensis_plasmid_pBMB2062-4ac
3	thuringiensis_tolworthi_plasmid_pBMB2062
4	Ames_Ancestor_plasmid_pXO2
4	A0248_plasmid_pXO2
4	A2012_plasmid_pXO2
5	cereus_plasmid_pBCXO1
5	cereus_G9241_plasmid_pBCXO1
6	A0193_plasmid_pXO2
6	WNA_USA6153_plasmid_pXO2

Table 2: Bacillus plasmids that cannot be resolved within the indicated clusters.

isolates. All other Bacillus genomes can be uniquely resolved based on SNPs. An analysis of Bacillus 16S rDNA sequences [11] showed a similar relationship among species, with the cereus, anthracis, and thuringiensis clustering into a diverse "cereus group", and the other species such as pumilus, lichenformis, and subtilis on a separate branch.

The Bacillus plasmids do show clear pXO1 and pXO2 clusters (Figure 5 (included as supplementary data)). There are 1057 homoplastic SNPs that do not map to nodes of the tree, highlighting the potentially complex lineages and horizontal gene transfer events that might have occurred. The 14 sequences that cannot be uniquely resolved are shown in Table 2.

E. coli

Whole genome kSNP analyses show that the O157:H7 sequences form a distinct clade. K-12 sequences cluster with DH1 and BW2952, and have as a near neighbor the enterotoxigenic ETEC_H10407 sequence (Figure 6 (included as supplementary data)). The ML tree (shown) and the hamming distance tree (not shown) are virtually identical. As in an analysis by [12] our SNP results indicate that O55_H7_CB9615 clusters closely with the O157:H7 sequences. Diamant et al. [12] created a phylogeny based on the sequences of 7 noncoding regions of approximately 200 bp each, some containing simple sequence repeats. They found that some isolates were not amplified by some of their primers, and others contained no variation among groups of isolates. For example, all the O157:H7 that they studied had completely identical sequences at the loci they examined, so they could not differentiate among them. In contrast, from our SNP analyses based on whole genome sequences, the O157:H7 are very similar but there are loci that enable discrimination at a finer scale, enabling isolate level discrimination of the available genomes. In fact, all the available E. coli genomes can be resolved based on putative SNPs.

We created a tree for the available E. coli genomes using an in silico application of the assays described in [12], predicting amplicons from their primers and aligning and building a tree for the sequences of the combined 7 regions using Dialign with default parameters [13], shown in Figure 7 (included as supplementary data). Essentially, we simulated amplification and sequencing using the primers from Diamant for the sequenced genomes. This tree based on the 7 regions from Subramanian et al. shows the O157:H7 genomes as three separate clusters, and it differs substantially from the SNP tree. Some of the differences may be due to gaps and errors in draft genomes, but real differences such as the absence of a given region or variations in primer binding sites also affect the relationships.

For the plasmid data (Figure 8), only one pair of sequences cannot be uniquely resolved: 517-2H1_plasmid_pLEW517 and plasmid_pLEW517. These have different lengths and are collected from different strains, but we found no SNP differences between them.

Discussion

The SNPs uncovered by kSNP are putative, since some may be a result of sequencing errors and may need further validation, for example, by additional sequencing or SNP microarrays. For SNP analysis of a single gene or other relatively short set of sequences that one can comfortably align, SNP identification from an alignment is a better option, as it will uncover clustered SNPs within a 12 base proximity. The kSNP k-mer approach described here is intended for larger scale applications where there might hundreds of genomes with lengths up to the 5-10 megabase range. Indeed, the availability of more than one genome is an essential requirement of this method. For viruses that are too divergent to align well and if most variation is at a scale larger than single nucleotide differences, this approach might be valuable as a preliminary method to cluster sequences into a preliminary phylogeny as a guide to those subsets for which alignment might be feasible. For the plasmid analyses we have included, it may not be ideal, or even accurate, to analyze all plasmids together and draw them as part of the same tree since a common ancestor may be very distant, and some branches may not contain any of the same loci as a distant branch. However, this method enables a fast, first-pass clustering, since cluster-specific loci are identified which serve to group the sequences into related sets and suggest relationships within those clusters. A subject matter expert can then separately examine branches of interest or pull out those SNP loci that differentiate key branches.

One application of the kSNP approach is to determine SNP alleles that characterize a node. For example, there are 5,173 SNP loci for which all anthracis species share the same allele which differs from the alleles in non-anthracis Bacillus species. These node-distinguishing SNPs may be useful for developing detection or genotyping assays. Another possible application for these SNP data is to guide decisions as to how to allocate efforts for genome sequencing: A SNP microarray designed from the output of kSNP can yield data to generate a phylogeny for multiple unsequenced isolates. Using a microarray to detect SNP variants at known loci (from SNP analysis of available genomes) in an unsequenced isolate may be a relatively cost effective method to place the isolate on a phylogeny, and may help to determine whether it is of sufficient interest (e.g. novel) to merit the expense of sequencing. While this will not uncover novel SNP loci, it can suggest how similar an isolate is to other isolates at known SNP loci. We have also used kSNP to analyze a set of genomes before and after adding newly sequenced genomes to the mix, to rapidly determine if those new genomes contributed any novel SNP loci.

As reviewed in [14] and succinctly stated by [12], multilocus analysis enables one to "dilute the bias of individual loci". Since the approach described here scales to entire genomes, resulting phylogenies should be less affected by regions that have undergone strong selection, deletions, or horizontal gene transfer (HGT) than other methods that rely on only a handful of genes. For the E. coli genomes, we found that a tree based on simulated amplification and sequencing of 7 regions selected by Diamant et al. [12] differed substantially from a tree based on whole-genome SNP analysis. The tree based on 7 regions did not cluster the O157:H7 sequences together, but broke them into 3 distant groups, in contrast to the whole-genome SNP tree which tightly clustered all the O157:H7 sequences as a single group.

SNPs from HGT regions should show up as homoplastic SNPs that are inconsistent with the phylogenetic relationships of the whole genome SNP tree. Although beyond the scope of this work, SNPs in HGT regions might be distinguished from mutations or sequencing errors if HGT SNPs appear as blocks of SNPs in proximity on the genome that are consistent with an alternative tree. In other work, we are using results of kSNP as a starting point for such analyses (unpublished).

This method is an improvement from a previous approached developed by one of the authors [15] since that method demanded a consensus sequence for the initial input. Although with the previous method we were able to build a BLAST-based consensus for some target sets, and it was certainly more feasible than an approach

requiring a multiple sequence alignment, it still scaled poorly for dozens of genomes. We found that poor accuracy of the consensus affected our ability to identify putative SNPs, and we missed SNPs from any region that was not present in the reference genome around which we built the consensus.

Nevertheless, the method as described here will miss some SNPs that are less than 12 bases apart (or $\frac{k-1}{2}$, if other values of k are used). Adjacent mutations will be missed in many cases by this method. But it can find nearby SNPs in some cases. For example, if only the following two oligos are present in the targets, then neither of the positions in bold will be found as candidate SNPs:

ACTTTGT**C**ATCA**A**TCGAATCGGAGA

ACTTTGT**G**ATCA**G** TCGAATCGGAGA

But if in addition, either one of the following oligos is also present in the enumerated list of k-mers and their reverse complements

ACTTTGT**C**ATCA**G** TCGAATCGGAGA

ACTTTGT**G**ATCA**A** TCGAATCGGAGA

then the candidate SNP at the 13th position should be found, although it will be counted as two separate loci for each surrounding sequence variant. Thus, SNPs in tight linkage disequilibrium which are less than 12 bases apart can be missed by the method as described. Adding a "fuzzy" search that allows mismatches is a possible improvement to address this shortcoming, but it is more complicated and will reduce scalability of the algorithm. The proximity of SNPs that can be found is also affected by the choice of k in the initial oligo enumeration. Shorter k enables us to find SNPs in closer proximity to one another, but also increases the chance that the surrounding sequence will be present more than once in a target with a different central nucleotide, and thus be thrown out of the pool of putative SNPs. Shorter k may be appropriate for some highly variable targets like viruses, as appears to be the case for FMDV and Filoviridae for which more SNP loci were found and the resulting tree clustered genomes more consistently with serotype or species designations.

It is possible that this approach could also be applied to identify protein differences at a large, whole-proteome scale by enumerating peptide k-mers. The suffix array algorithms, BLAST, and k-mer computations should work on the amino acid as well as the nucleotide alphabet. The value of k would need to be shorter than that for nucleotide sequence analysis, and would depend on the pattern length of peptide conservation and variation. An improvement we are currently implementing to improve speed and scalability to handle gigabases of raw, unassembled short reads from Illumina™ (San Diego, CA) or SOLID™ (Life Technologies, Carlsbad, CA) is to replace the BLAST+ step with MUMmer [16].

In conclusion, we describe kSNP, an approach for rapid, scalable SNP analysis of up to hundreds of bacterial or viral genomes, of either draft or finished quality. While the method will not find all SNPs in the data, particularly those that are in very close proximity, it will find a large number of them which can be used to build a phylogeny based on whole-genome analysis rather than being limited to a few genes. An advantage of scalable whole genome analysis is to avoid bias that might be present in some smaller regions that may have undergone horizontal gene transfer, strong selection, sequencing errors, or other processes. Applications of this kSNP approach could be to find sets of SNPs that map to a branch of interest or correlate with a notable phenotype, or to identify key node-distinguishing SNPs from which one can design SNP assays such as microarrays, PCR, or targeted sequencing.

Auspices

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Acknowledgments

Many thanks to David Hysom for writing "sa" to incorporate the sarray functions from [2], since using sa for other applications was the inspiration for this approach to SNP finding. This work was funded by The National Biodefense Analysis and Countermeasures Center (NBACC) of the Department of Homeland Security. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the National Biodefense Analysis and Countermeasures Center (NBACC), Department of Homeland Security (DHS), or Battelle National Biodefense Institute (BNBI).

References

1. Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19: 1124-1132.
2. McIlroy TM, McIlrow MD Sarray a collection of Suffix-array functions. 1997, Copyright (C) Lucent Technologies: http://www.cs.dartmouth.edu/~doug/sarray/.
3. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
4. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.
5. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6, in Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
6. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.
7. Wittmann TJ, Biek R, Hassanin A, Rouquet P, Reed P, et al. (2007) Isolates of Zaire ebolavirus from wild apes reveal genetic lineage and recombinants. Proc Natl Acad Sci U S A 104: 17123-17127.
8. Knowles NJ, Samuel AR (2003) Molecular epidemiology of foot-and-mouth disease virus. Virus Research 91: 65-80.
9. Emerson GL, Li Y, Frace MA, Olsen-Rasmussen MA, Khristova ML, et al. (2009) The Phylogenetics and Ecology of the Orthopoxviruses Endemic to North America. PLoS ONE 4: e7666.
10. Li Y, Carroll DS, Gardner SN, Walsh MC, Vitalis EA, et al. (2007) On the origin of smallpox: correlating variola phylogenics with historical smallpox records. Proc Natl Acad Sci U S A 104: 15787-15792.
11. Porwal S, Lal S, Cheema S, Kalia VC (2009) Phylogeny in Aid of the Present and Novel Microbial Lineages: Diversity in Bacillus. PLoS ONE 4: e4438.
12. Diamant E, Palti Y, Gur-Arie R, Cohen H, Hallerman EM, et al. (2004) Phylogeny and Strain Typing of Escherichia coli, Inferred from Variation at Mononucleotide Repeat Loci. Appl Environ Microbiol 70: 2464-2473.
13. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol 3: 6.
14. Pearson T, Okinaka RT, Foster JT, Keim P (2009) Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect Genet Evol 9: 1010-1019.
15. Gardner SN, Wagner MC (2005) Software for Optimization of SNP and PCR-RFLP genotyping to discriminate many genomes with the fewest assays. BMC Genomics 6: 73.
16. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.