



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Interim report on updated microarray probes for the LLNL Burkholderia pseudomallei SNP array

S. Gardner, C. Jaing

March 30, 2012

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

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

**Interim report on updated microarray probes for the LLNL
Burkholderia pseudomallei SNP array**

**Project Title: Forensic Analysis and Characterization of *Burkholderia* Isolates using
Burkholderia SNP Genotyping Microarray**

IAA No: HSHQPM-10-X-00099/P00001

Shea Gardner and Crystal Jaing

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

Principal Investigator and Correspondent

Crystal Jaing

925-424-6574, jaing2@llnl.gov

LLNL-TR-543366

March 12, 2012

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

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

Introduction

The overall goal of this project is to forensically characterize 100 unknown *Burkholderia* isolates in the US-Australia collaboration. We will identify genome-wide single nucleotide polymorphisms (SNPs) from *B. pseudomallei* and near neighbor species including *B. mallei*, *B. thailandensis* and *B. oklahomensis*. We will design microarray probes to detect these SNP markers and analyze 100 *Burkholderia* genomic DNAs extracted from environmental, clinical and near neighbor isolates from Australian collaborators on the *Burkholderia* SNP microarray. We will analyze the microarray genotyping results to characterize the genetic diversity of these new isolates and triage the samples for whole genome sequencing. In this interim report, we described the SNP analysis and the microarray probe design for the *Burkholderia* SNP microarray.

Methods

*k*SNP analysis

With the flood of whole genome finished and draft microbial sequences, we designed kSNP as a more scalable bioinformatics tools for sequence comparison than multiple genome alignment (1). It scales to hundreds of bacterial or viral genomes, and can be used for finished and/or draft genomes available as unassembled contigs and a limited number of genomes as unassembled reads. The method is fast to compute, finding SNPs and building a SNP phylogeny in seconds to hours. The SNP-based trees it generates are consistent with known taxonomy and trees determined in other studies. The approach can handle as input hundreds of megabases to gigabases of sequence in a single run. The algorithm kSNP is based on k-mer analysis using suffix arrays and requires no multiple sequence alignment. A SNP locus is represented by the surrounding sequence, e.g. when $k=25$, a SNP locus is indicated by the 12 conserved bases up- and down-stream of the variable central SNP base, where conservation is required among only 2 or more of the input genomes. This representation of a SNP locus is based on surrounding sequence information rather than positional information in a genome. It differs from traditional concepts of a SNP locus, and 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. It is also useful for viruses in which there may be highly divergent and poorly alignable regions among a large group of sequences, and conserved regions only exist among small subgroups of sequences. There are currently many users of kSNP at NCBI, FDA, and CDC.

Phylogenetic Trees

Phylogenetic trees based on neighbor joining of SNP hamming distances (# SNP differences between every pair of genomes) are shown in Figures 2, 3, and 4. Hamming trees are more robust to “missing” or cluster specific loci than other types of trees like maximum likelihood of the SNP allele sequences, and therefore perform better for divergent sets of sequences like viruses. In Figures 2-4, the numbers of SNP alleles shared by the leaves of a branch are shown at each node, and the count of strain specific SNPs alleles are given in brackets or as numbers following the strain name. Homoplastic SNP loci are those in which the pattern of shared alleles does not conform to any of the branches of this tree. They depend not only on the branch topology, but also on the placement of the root and direction of descendant/ancestor relationships. Because alternative trees are possible, the exact number and identities of those

SNPs considered to be homoplastic will depend on a particular tree. Since we did not include outgroup sequences to root the trees, we counted the number of homoplastic SNPs for every possible rooting of the tree, and selected the best root as the one that gave the fewest homoplastic loci. Ties for the best root were decided randomly.

Microarray probe design

Microarray probes were designed for every SNP. Probe design strategy maximized sensitivity and specificity based on extensive prior lab testing on a NimbleGen microarray platform. Probes were 32-40 bases long, with the SNP at the 13th position from the 5' end. Probe candidates with hybridization free energy below $\Delta G = -43$ kcal/mol were shortened until either their ΔG exceeded -43 kcal/mol or they reached the minimum 32 bases. Thus, we attempted to equalize binding energy to the extent possible within the allowable length range. Probes were designed around the SNP on both the plus and minus strands, for all 4 possible SNP alleles, and all surrounding sequence variants. Because of the asymmetry of the SNP on the probe for the plus and minus strands (i.e. the probe is at the 13th position, not the probe center) the plus and minus probes provide good, orthogonal information for identifying the correct SNP allele in a sample hybridized to the SNP array. The probes are longer than the length k , so there is often variation in probe sequence outside of the conserved sequence defining the SNP locus, and probes for all sequence variations outside of the k -mer region were designed. Thus, there are often more than 8 probes per SNP locus. On the array, only the subset of probes for observed alleles were included. For example, if a SNP was bi-allelic, then only the probes for the 2 variants in the available genomes were included.

Results

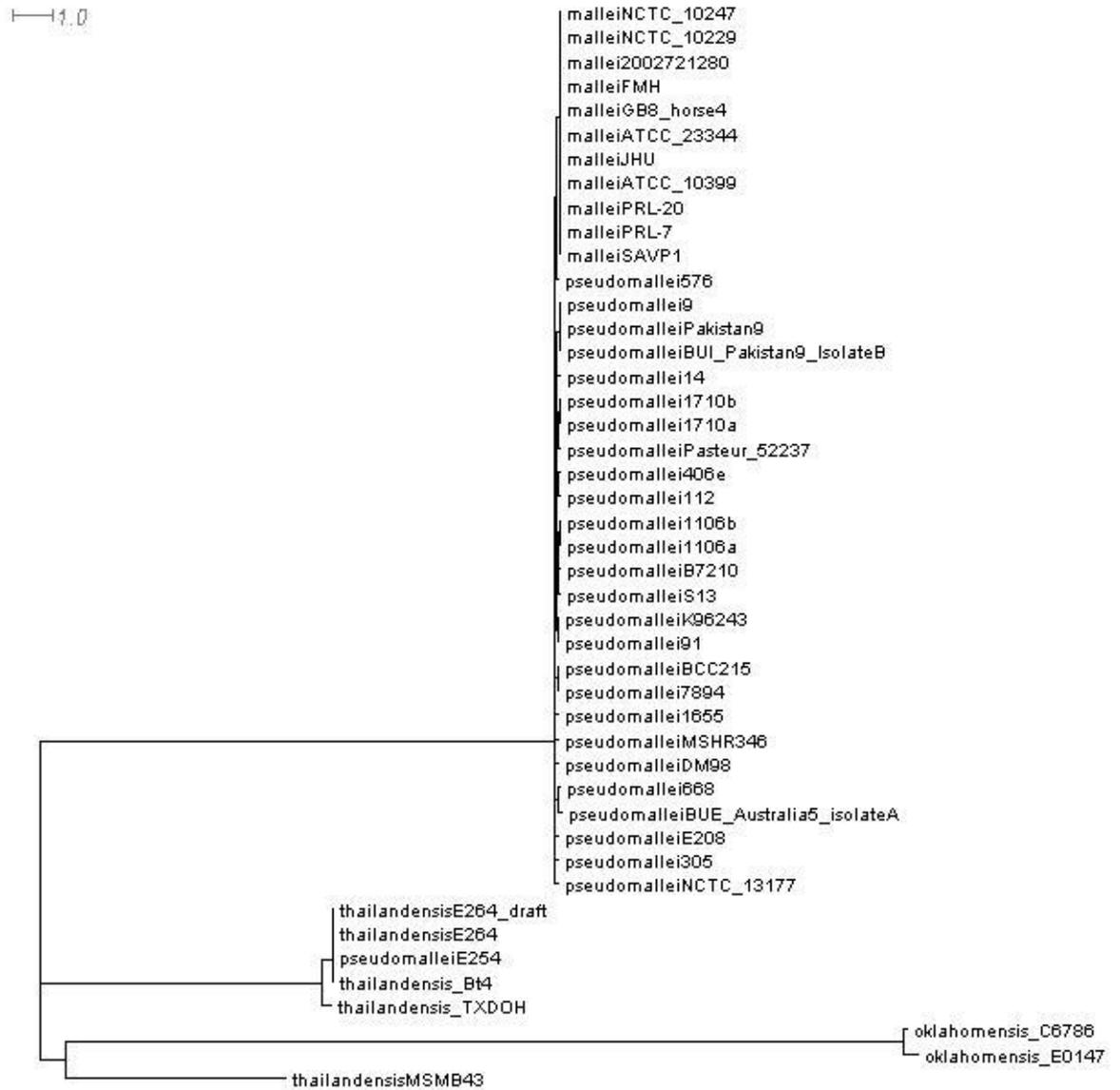


Figure 1. *Burkholderia* SNP maximum likelihood tree using SNPs from *B. pseudomallei*, *B. mallei*, *B. thailandensis* and *B. oklahomensis*.

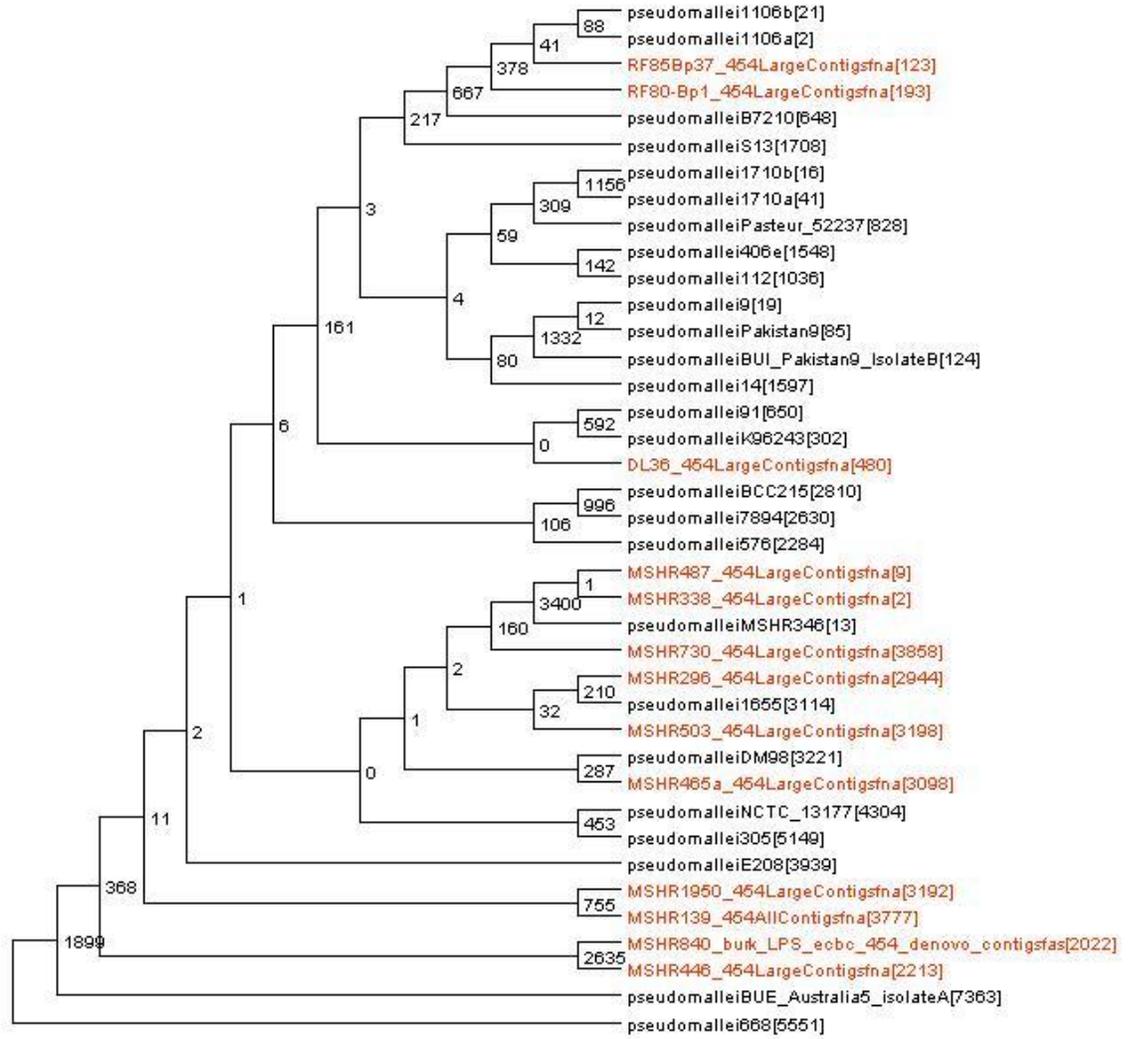


Figure 2. *B. pseudomallei* SNP phylogenetic tree. The sequences in red are draft genomes provided by Northern Arizona University. The number of SNPs for each strain or branch is listed in the brackets.

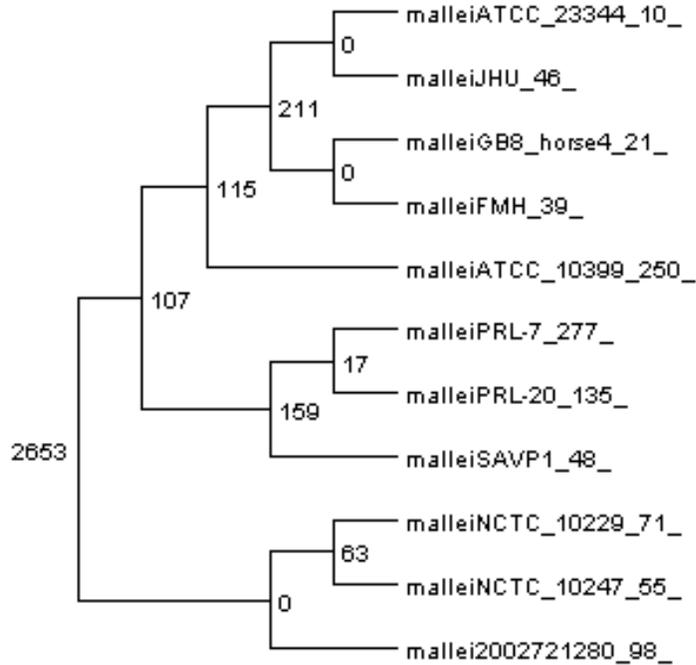


Figure 3. *B. mallei* SNP phylogenetic tree labeled with SNP loci counts for each branch and strain.

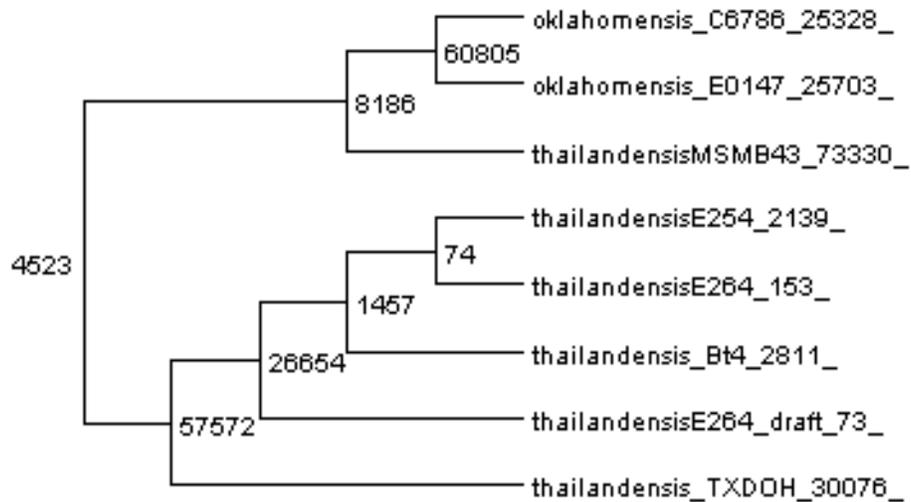


Figure 4. *B. thailandensis* and *B. oklahomensis* SNP phylogenetic tree labeled with SNP loci counts for each branch and strain.

Probes were included for all kSNP loci in the 39 *Burkholderia pseudomallei* genomes available at the time the array was designed. In addition, probes were included for all kSNP loci among the 11 *B.mallei* genomes available. Probes for a random selection of 500 SNPs that are species specific from the branch for *B.mallei* were also included, i.e. SNPs shared between all *B.mallei* strains and different from other *Burkholderia* species. The closest species to *B.mallei* and *B.pseudomallei* are *B.thailandensis* and *B.oklahomensis*, so in the space remaining on the array probes were included for each major branch of *B.thailandensis* and *B.oklahomensis* subtypes. We used probes for a random selection of 500 SNP loci from each major branch of the *B.thailandensis* and *B.oklahomensis* SNP sub-trees in the *Burkholderia* kSNP analysis, but no strain specific SNPs (i.e. SNP alleles present in only one genome). These probes should inform a user if a sample is *B.thailandensis* or *B.oklahomensis*, and the major subtype. The number of probes for each of the *Burkholderia* species included on the SNP array is shown in Table 1.

Table 1. Number of SNPs identified and probes designed for the *Burkholderia* species

Species	# SNPs	# homoplastic SNPs	Number probes
<i>B. pseudomallei</i>	150,680	66,007	666,348
<i>B. mallei</i> strain genotyping	1,908	308	7,423
<i>B. mallei</i> species identification	--	--	5,180
<i>B. thailandensis</i> and <i>B. oklahomensis</i> species and subtype	177,260	--	40,960
Total after removing duplicates			718,354

References

1. Gardner, S. and Slezak, T. (2010) Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. *J Forensic Res*, **1**, 107, doi:110.4172/2157-7145.1000107.