SANDIA REPORT

SAND2013-8906
Unlimited Release
Printed October 2013

Pathogenicity Island Mobility and Gene Content

Kelly Porter Williams

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2013-8906
Unlimited Release
Printed October 2013

Pathogenicity Island Mobility and Gene Content

Kelly Porter Williams
Systems Biology
Sandia National Laboratories
P.O. Box XXX
Livermore, California 94551-MS9291

Abstract

Key goals towards national biosecurity include methods for analyzing pathogens, predicting their emergence, and developing countermeasures. These goals are served by studying bacterial genes that promote pathogenicity and the pathogenicity islands that mobilize them. Cyberinfrastructure promoting an island database advances this field and enables deeper bioinformatic analysis that may identify novel pathogenicity genes.

New automated methods and rich visualizations were developed for identifying pathogenicity islands, based on the principle that islands occur sporadically among closely related strains. The chromosomally-ordered pan-genome organizes all genes from a clade of strains; gaps in this visualization indicate islands, and decorations of the gene matrix facilitate exploration of island gene functions. A “learned phyloblocks” method was developed for automated island identification, that trains on the phylogenetic patterns of islands identified by other methods. Learned phyloblocks better defined termini of previously identified islands in multidrug-resistant *Klebsiella pneumoniae* ATCC BAA-2146, and found its only antibiotic resistance island.

ACKNOWLEDGMENTS

This research was fully supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. I thank Cathy Branda and other managers for supporting my research program, and my Sandia research collaborators, in particular, Joe Schoeniger, Corey Hudson, Robert Meagher, Zach Bent and Owen Solberg.

CONTENTS

FIGURES

NOMENCLATURE

Kpn2146	*Klebsiella pneumoniae* ATCC BAA-2146

1. INTRODUCTION

Key goals towards national biosecurity include improved methods for identifying pathogens (natural or engineered), predicting their emergence, and developing vaccines and therapeutics. These goals could be achieved, in part, through a comprehensive analysis of bacterial genes that promote pathogenicity and the mobile DNA elements (pathogenicity islands) where such genes typically reside. Our current knowledge of bacterial pathogenicity genes is extremely limited, and typically based on painstaking laboratory study. Cyberinfrastructure such as an island database will advance this field and moreover provide a short list of potential pathogenicity genes in a new pathogen, enabling deeper bioinformatic island analysis that may allow positive identification of novel pathogenicity genes.

Multiple islands can accrue throughout a genome to combinatorially enhance or modulate pathogenicity. Diverse islands can even appear in tandem arrays at genomic integration hotspots, a configuration promoting inter-island recombination events that may produce new islands with novel gene combinations. An island database will enable an in-depth analysis of island genomic sites and arrangements, as well as their phylogenetic distributions, to shed light on how island mobility, evolution, and functional cooperation promote pathogenicity.

Comparing genomes within a bacterial species or genus divides their genes into either core (present in all strains in a species or genus group) or accessory (present in some strains and not others) components. Accessory genes often occur in clusters as mobile DNA segments termed islands, that can move between closely- or even distantly-related strains [1]. This mobility can often be ascribed to well known mechanisms: islands may move in the form of bacteriophage virus particles or through conjugation. Island genes themselves may in principle be divided into selfish genes (for island mobility and maintenance functions), and payload genes (that alter the phenotype of the island's bacterial host, improving host fitness through exploitation of an ecological niche). Genes that promote bacterial pathogenicity are a typical example of such payload genes, and islands that contain these are called pathogenicity islands (PAIs) [1]. Below, phyogenomic methods, i.e., phylogeny-based analyses of multiple whole genomes, are developed for visualizing and automatically determining genomic islands.

Island gene characterization presents several challenges. Although some island genes are large and show clear relationships to genes of known function, more often they are small and "hypothetical", i.e., without clear homology to known genes. A likely reason for this is that island genes should evolve more rapidly than core genes, because they are subject to major selective sweeps when they enter a new bacterial host, and because they tend to mobilize under stress conditions where DNA replication and repair can be unusually error prone and recombination can be unusually promiscuous. But there may be additional reasons for the difficulty in characterizing island genes. They may be undergoing decay if their function has not been required in recent hosts. In a more intriguing hypothesis counter to Jacob's "tinkerer" model of evolution [2], islands may be random sequence cauldrons where new (pathogenicity) proteins are forged. The proposed research program is expected to reveal genomic, post-genomic and other bioinformatic hallmarks that will allow positive identification of island payload genes in general, and therefore indicate candidates for novel pathogenicity genes. Structure determination for such candidates may suggest pathogenicity mechanisms, and additionally help determine whether they are highly diverged versions of known protein folds or instead represent novel folds that may have evolved de novo.

2. PHYLOGENOMIC VISUALIZATION OF GENOMIC ISLANDS

One method for finding genomic islands is based on the phylogenomic principle that islands occur sporadically among closely related strains. The rich visualization below puts the total gene content of forty *Brucella* genomes before the eye at once, and highlights islands as darker regions. Browser-based forms of such visualizations allow mouseover labeling to explore taxon and gene names, and switching between different color-coding schemes to emphasize different gene properties.

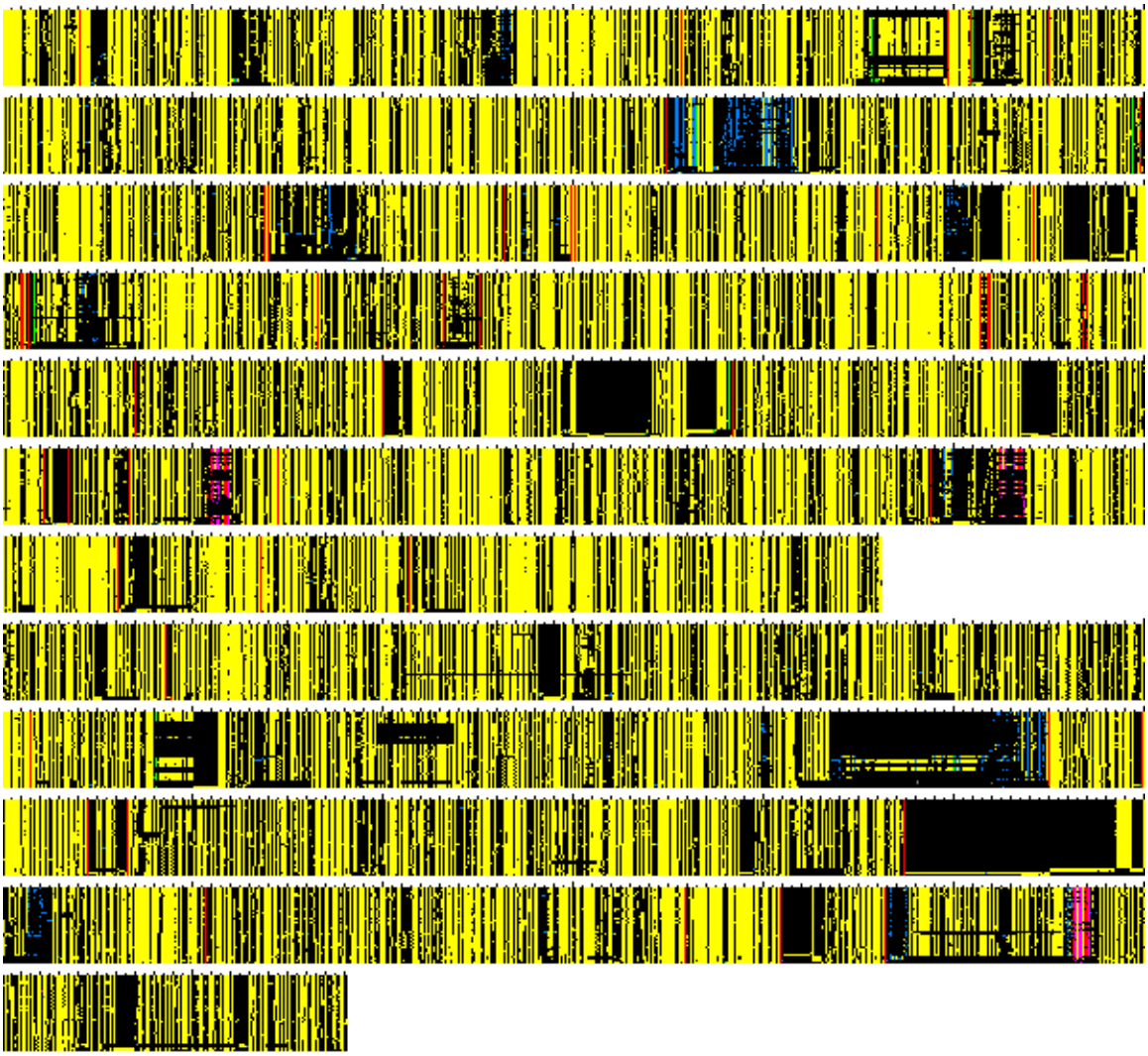

Figure 1. Chromosomally ordered pan-genome: total gene content of forty genomes. Each bar has 40 *Brucella* genomes in phylogenetic order in rows along the y-axis, and 600 genes in pan-genome order along the x-axis. The bars wrap after each 600-gene set to show the two complete pan-chromosomes. Color coding: black, gene absent; pink, rRNA genes, red, tRNA genes, green, integrase genes; blue, transposase genes; yellow, other genes.

Pan-Genome Protocol. The first challenge is to organize all genes into the chromosomally-ordered pan-genome. My currently preferred method is to align all genomes under comparison (typically a genus-level group) into sequence blocks using mugsy [3], to order the blocks with phylogenetic consideration using Gasts [4], and then annotate the pan-genome ordering of DNA to convert it to a pan-genome ordering of genes using mafAnnotate.pl. The output is converted to "xy-color" files using PanXyc.pl, that specify the color-coding at each gene, and these are converted into the final image and javascripted HTML pages using xyColorPan.pl.

3. PHYLOGENOMIC IDENTIFICATION OF GENOMIC ISLANDS

While it is valuable to enable users to explore phylogenomic island data visually, as above, it is also important to develop methods that call genomic islands automatically without requiring user intervention. Several aspects of genomic islands have been used as a basis for automated identification: i) their preference for integrating into tRNA/tmRNA genes, ii) their frequently observed bacteriophage nature, iii) their differing nucleotide and amino acid composition relative to the core genome, and iv) their sporadic occurrence among closely related strains, the latter aspect lending itself to phylogenomic approaches to island-finding. Because all these aspects have exceptions, none is a perfect approach for island-finding; combined approaches will likely perform best.

To understand the evolution of the multidrug-resistant *Klebsiella pneumoniae* ATCC BAA-2146 (Kpn2146), whose genome sequence we recently completed [5], we sought to characterize its genomic island content. We applied my Islander program for finding genomic islands in tRNA and tmRNA genes [6]. This program is distinguished by its precise specification of island termini, providing gold standard island sets; it found six islands in Kpn2146. We also applied the program Phast which identifies clusters of phage genes [7]. This program supported three of the Islander islands, and identified four additional islands, one precisely. The ten resulting islands accounted for 6.3% of the Kpn2146 chromosome, and were used as a training set to identify additional islands in a novel phylogenomic method termed "learned phyloblocks".

The chromosomes from Kpn2146 and the 11 other complete genomes within the *Klebsiella* clade were aligned using mugsy [3]. This alignment partitioned the Kpn2146 chromosome into intervals termed "phyloblocks" where all positions share the same phylotype, *i.e.*, presence/absence profile among the other genomes (Fig. 2). Phyloblocks largely respect gene and other feature boundaries (phyloblock junctions fall into inter-feature spaces 3.81-fold more frequently than expected for random distribution on the chromosome, P=2e-16), providing biological validation. All phylotypes were evaluated for phylogenetic complexity, by reconciling a robust genome tree with its subtree containing only the phylotype taxa. This classified each nonubiquitous phylotype as either simple (explainable by a single gain/loss event), or complex (requiring multiple gains/losses).

The nonubiquitous phylotypes accounted for much (47.5%) of the Kpn2146 chromosome, suggesting that gene flux is high in *Klebsiella*, and that nonubiquity is too broad a class to pinpoint additional integrative genomic islands. We reasoned that some nonubiquitous phylotypes might be more indicative than others of horizontally transferred islands, if there are particular "highways" of island transfer for *Klebsiella*. Phylotypes were ranked by the fraction of their nucleotides in the training islands. "Learned" phyloblocks, those for which this fraction was > 0.25, accounted for 7.6% of the chromosome, a reasonably-sized genome component within which to find additional islands. The phylogenetically complex types were significantly overrepresented among the learned phylotypes (36 of 38) relative to the non-learned phylotypes (183 of 246) (one-sided χ test of proportions: $P < 0.005$). However the learned group did contain two important phylotypes that had been classified as phylogenetically simple, Kpn2146-only and Kpn2146 with only its closest related genome.

Learned phyloblocks indicated the island Kpn23SapB, with an integrase gene and *att* site pair, that was missed by Islander and Phast. Kpn23SapB is the only known Kpn2146 genomic island bearing an antibiotic resistance gene (the AadA4 cassette of an integron-related region). Learned phyloblocks also indicated a large gene cluster for capsular polysaccharide synthesis

(*cps*). Although *Klebsiella* is generally described as an opportunistic pathogen, the capsule is considered one of its major pathogenicity determinants. One publication on *Klebsiella cps* regions has described them as highly "diverse" [8], while the phyloblocks result and further inspection moreover suggests that they may be mobile and a worthy object of future study. An overview of learned phyloblocks across the chromosome (Fig. 2) shows the tight mapping to *cps*, mobile islands, and transposing insertion sequences.

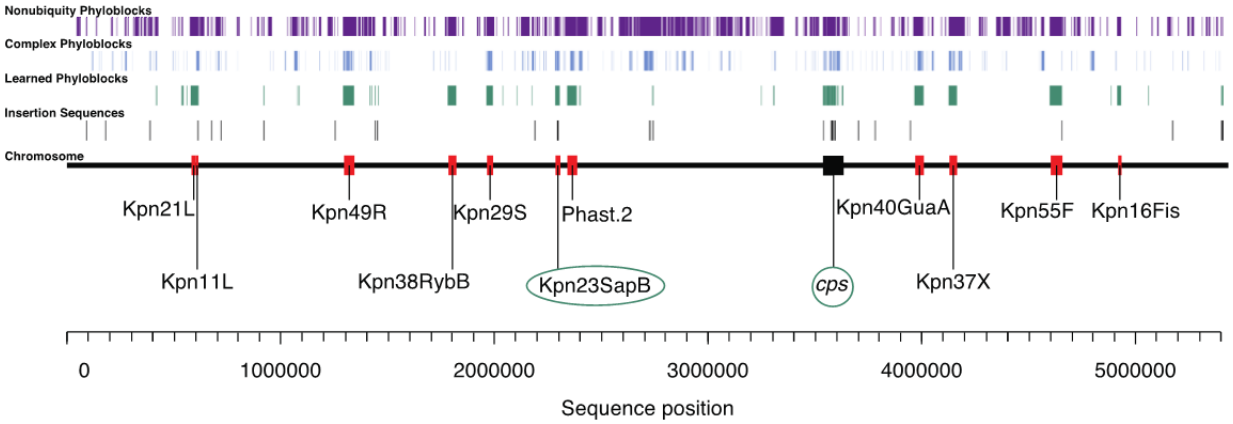

Figure 2. Learned phyloblocks indicate the only antibiotic-resistance island of Kpn2146 and the pathogenicity-determining capsule polysaccharide synthesis (*cps*) region. Nonubiquity phyloblocks: those missing in at least one of the 11 reference *Klebsiella* chromosomes. Complex phyloblocks: those requiring more than one gain/loss event to reconcile the phylotype with the genome tree of Fig. 1. Learned phyloblocks, those enriched in the ten uncircled training islands. As a percentage of their combined 411 kbp, the learned phyloblocks mapped either to the training islands (81.9%), the two circled newly indicated regions (12.0%), insertion sequences (2.1%), or to small scattered regions not showing island hallmarks (4.0%).

Learned phyloblocks also provided excellent definition of island termini (Fig. 3), confirming those from Islander, and helping identify *att* sites for two of the three coarsely determined Phast islands.

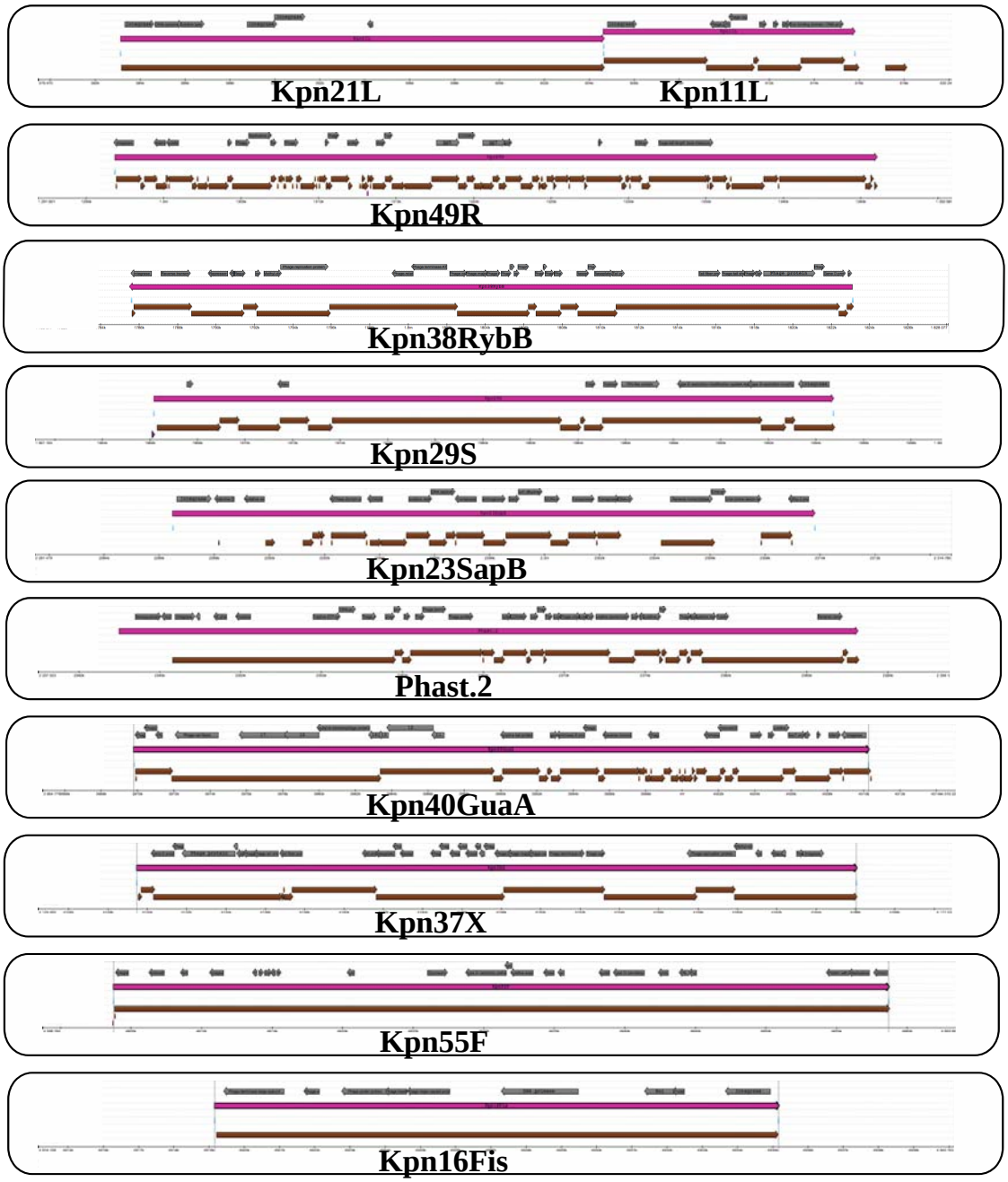

Figure 3. Learned phyloblocks define island endpoints. Each Kpn2146 genomic island is marked in pink. The lower brown segments are the learned phyloblocks, jittered for distinction. The upper grey arrows are protein-coding genes. Five kbp of each flank is shown for each island.

Island finding has advanced beyond the Kpn2146 analysis, to survey all available complete prokaryotic genomes. Islander and Phast programs have been applied, identifying over 4000 islands. Mugsy alignments have been produced for each multi-genome genus, and phylotypes have been taken, in preparation for applying learned phyloblocks.

4. CONCLUSIONS

This work advances Sandia's position in the area of emerging disease, by producing new tools for genomic analysis, databases of genomic islands (the building blocks of pathogens) and their gene content, and principles of how genomic islands move among bacteria to generate novel pathogens. It will allow bioinformatic prediction of novel pathogenicity genes. Such elucidation of the natural pathways of pathogen emergence provides background information that will enable determination of whether a threat organism's genome was produced by bioengineering.

With our software that finds islands based on target gene preference and on phylogenomic considerations, together with others' software based on nucleotide composition and phage-like gene clustering, we have an excellent arsenal of tools that I believe can solidify grant applications NIAID or DTRA. Our current aim is to focus these tools on *Burkholderia*, correlating islands with pathogenicity genes, and presenting a white paper to DTRA.

5. REFERENCES

1. Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. 2:414-424.
2. Jacob F. 1977. Evolution and tinkering. Science 196:1161-1166.
3. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334-342.
4. Xu AW, Moret BME. 2011. GASTS: parsimony scoring under rearrangements. WABI 2011: 351-363.
5. Hudson CM, Bent ZW, Meagher RJ, Williams KP. 2013. Resistance determinants and mobile genetic elements of an NDM-1-encoding *Klebsiella pneumoniae* strain. Antimicrob. Agents Chemother., submitted.
6. Mantri Y, Williams KP. 2004. Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. Nucleic Acids Res. 32:D55-58.
7. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Res. 39:W347-352.
8. Shu HY, Fung CP, Liu YM, Wu KM, Chen YT, Li LH, Liu TT, Kirby R, Tsai SF. 2009. Genetic diversity of capsular polysaccharide biosynthesis in *Klebsiella pneumoniae* clinical isolates. Microbiology 155:4170-4183.

DISTRIBUTION

1	MS9291	K. Williams	8623
1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911


Sandia National Laboratories