Published in final edited form as:

Infect Genet Evol. 2015 June ; 32: 191–198. doi:10.1016/j.meegid.2015.03.018.

SISPA-Seq for rapid whole genome surveys of bacterial isolates

Meredith S. Wright¹, Timothy B. Stockwell², Erin Beck², Dana A. Busam², Saralee Bajaksouzian³, Michael R. Jacobs^{3,4}, Robert A. Bonomo^{4,5}, and Mark D. Adams^{1,*}

¹J. Craig Venter Institute, La Jolla, CA

²J. Craig Venter Institute, Rockville, MD

³University Hospitals Case Medical Center

⁴Case Western Reserve University

⁵Louis Stokes Cleveland Veterans Affairs Medical Center

Abstract

Whole genome sequencing (WGS) of large isolate collections has many applications, yet sequencing costs are still significant. We sought to develop a rapid and cost efficient WGS method to address fundamental questions in clinical microbiology. We evaluated the performance of SISPA (Sequence-Independent, Single-Primer Amplification) combined with next-generation sequencing (SISPA-Seq) of 75 clinical isolates of *Acinetobacter baumannii* to establish whether SISPA-Seq resulted in sufficient coverage and quality to 1) determine strain phylogenetic placement and 2) and carriage of known antibiotic resistance (AbR) genes. Strains for which whole genome sequences were available were included for validation. Two libraries for each strain were constructed from separate SISPA reactions with different barcoded primers, using genomic DNA prepared from either high quality or rapid heat-lysis preparations. SISPA-Seq resulted in a median of 65x genome coverage when reads from both primer sets were combined. Coverage and quality were sufficient for detection of AbR genes by comparison of reads to the ARG-ANNOT database and were often sufficient to distinguish between different allelic variants of the same gene. kSNP and RAxML were used to construct a robust phylogeny based on single-nucleotide variants (SNVs) that showed that the SISPA-seq data was sufficient for sensitive and accurate phylogenetic placement. Advantages of the SISPA-Seq method include inexpensive and rapid DNA preparation and a typical total cost less than one-half that of standard genome sequencing. In summary, SISPA-Seq can be used to survey whole genomes of a large strain collection and identify strains that should be targeted for additional sequencing.

Keywords

whole genome survey; SISPA; SISPA-Seq; *Acinetobacter baumannii*

*Corresponding author: Mark D. Adams, Ph.D., J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA 92037, 858-200-1864, madams@jvvi.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Whole genome analyses of large strain collections have yielded important insights into the evolutionary processes contributing to the emergence of multidrug-resistant bacteria and the epidemiology of pathogens (Grad & Lipsitch, 2014; Koser, Ellington, & Peacock, 2014; L. M. Li, Grassly, & Fraser, 2014). These analyses typically use single nucleotide variant (SNV) patterns to assess population structure or transmission routes (Chewapreecha et al., 2014; Eyre et al., 2013; Long, Beres, Olsen, & Musser, 2014; Snitkin et al., 2012), while other studies have also investigated genome content to add further resolution to understanding evolutionary dynamics (Holt et al., 2008; Katz et al., 2013; Nasser et al., 2014). Though sequencing costs have decreased in recent years, whole genome sequencing of single isolates still represents a significant investment, in large part due to costs for DNA isolation and purification, and for library construction, that can limit the number of samples selected for whole genome sequencing. Multi-locus sequencing typing (MLST), pulsed-field gel electrophoresis (PFGE), repetitive element sequence based-polymerase chain reaction (rep-PCR), and other molecular epidemiological typing methods are useful for identifying coarse population changes (Decker et al., 2012; Diancourt, Passet, Nemec, Dijkshoorn, & Brisse, 2010; Hamouda, Evans, Towner, & Amyes, 2010; Wisplinghoff et al., 2008), but address only a small part of the genome, or do not provide detail on genome content variation. Therefore, a more cost-efficient method for surveying whole genome content and assessing population structure of large contemporary and historical strains collections would yield additional information about the emergence of antibiotic resistance, dissemination of different multidrug resistant (MDR) lineages, and pathogen surveillance within health-care facilities. We sought to develop a more rapid, cost-effective read-based method to survey whole genome content of isolate collections.

Sequence-Independent Single-Primer Amplification (SISPA) has been used in viral genomics and metagenomics to obtain whole genome sequences (Castrignano et al., 2013; Depew et al., 2013; Djikeng et al., 2008; Rosseel et al., 2013), but its use in bacterial genomics has not been explored. The method uses a barcoded amplification primer with 3'-end random nucleotides that anneal to template DNA and a second round of single primer PCR amplification. Amplified products can be pooled and used directly in next generation sequencing library construction. Complete or near-complete viral genomes can readily be obtained using this method, but potential coverage biases have been observed that are believed to primarily be due to variation in priming efficiency (Neill, Bayles, & Ridpath, 2014; Rosseel et al., 2013) and can result in some genomic regions being under-represented.

To determine its applicability for surveying whole bacterial genomes, we used SISPA-Seq to assess the population structure and antibiotic resistance gene carriage of a health-care associated pathogen *Acinetobacter baumannii* collection from one US hospital system in Northeast Ohio, USA. Genome sequences of a subset of these isolates have been previously reported (Wright et al., 2014), and these were included for validation.

Methods

Strains

Clinical isolates of the *A. baumannii* study population have been described previously (Decker et al., 2012; Perez et al., 2010; Wright et al., 2014). A list of the 75 strains and their relevant genomic features and antimicrobial susceptibility patterns is provided in Table S1. Strains were cultured from a variety of clinical sources between 2007 and 2008.

DNA Preparation

SISPA-Seq was attempted on genomic DNA isolated from all 75 strains. Two methods of DNA preparation were performed to compare the performance of the method on DNA from 1) a rapid, low-cost heat-lysis preparation with 2) higher quality DNA obtained from a commercial DNA isolation and purification kit. DNA was prepared from 62 strains using a rapid, inexpensive heat-lysis protocol (“boil prep”). Heat-lysed samples were prepared by diluting 10 µl of an overnight culture into 90 µl Tris EDTA, followed by incubation at 95 °C for 10 minutes. After heat treatment, cellular debris was pelleted by centrifugation for 5 min at 10000 × g and the supernatant was stored at -20 °C. High quality (“HQ”) DNA from the other 13 strains was isolated and purified using an alkaline lysis and ethanol precipitation approach with the MasterPure Gram Positive kit (Epicentre Biotechnologies). Twelve of the strains with HQ DNA have genome sequences available (whole genome shotgun or “WGS” contigs) and one of the boil prep strains, UH7707, has also been sequenced (Wright et al., 2014). Two SISPA amplifications were performed on each genomic DNA, using unique barcoded adapters (Figure 1; Supplemental Table 2).

SISPA reaction

To anneal random hexamers to template DNA, we combined 8.0 µL of 25 ng/µL template DNA, 0.8 µL of DMSO, 4.2 µL of water, and 2.0 µL of 10 µM Primer 1 (base primer + 3′ random hexamer). Each sample had a distinct base primer with a unique 20-base barcode to allow for pooling during sequencing (Table S2). This mixture was then incubated at 95°C for 5 minutes, and immediately cooled on ice for 5 minutes. This was followed by a Klenow reaction to fill in gaps and generate dsDNA, which consisted of adding 5.0 µL of master mix containing 2.0 µL 10X NEBuffer 2, 2.0 µL 10mM dNTP Mix, and 1.0 µL Klenow fragment (3′-5′ exo-) (New England Biolabs) to each sample from the previous step. This reaction mix was then incubated at 37°C for 60 minutes followed by 75°C for 10 minutes.

For the single primer amplification PCR reaction, Primer 2, the same base primer as above containing only the fixed 20-base barcode portion, not the random hexamer, was used. The PCR master mix consisted of the following: 10.0 µL 5X Colorless GoTaq Flexi Buffer, 0.2 µL GoTaq Hot Start Polymerase (Promega), 27.8 µL water plus 10.0 µL Klenow reaction product and 2.0 µL of 12.5 µM PCR primer. PCR conditions for amplification were as follows: 94°C for 2 minutes; 35 cycles of: 94°C for 30 seconds, 55°C for 30 seconds, 68°C for 48 seconds; 65°C for 10 minutes and a final hold at 4°C. PCR products were gel-verified and treated with exonuclease I. All 150 SISPA reactions were pooled, purified using a SPRI bead protocol (DeAngelis, Wang, & Hawkins, 1995) and quantified and sized by the Agilent 2100 Bioanalyzer using the High Sensitivity DNA Kit. The pool was then used for

standard adapter-ligation paired end Illumina library construction using the NEBNext DNA Library kit (New England Biolabs). This library, comprising all 150 SISPA reactions, was run on a single HiSeq 2000 2×100 base sequencing lane.

Raw sequence data files were processed to remove reads from duplicate molecules using custom Java software *removeRedundantMatePairs* which is available as part of the Executive for Large-scale VIRal Assembly (Elvira) package (<http://elvira.sourceforge.net/>). The resulting deduplicated read dataset and the mapping of DNA samples and associated SISPA barcodes used were then processed with DNA Barcode Deconvolution software (<http://deconvolver.sourceforge.net/>) to generate SISPA barcode trimmed read files for each DNA sample-SISPA barcode combination.

Phylogenetic analysis

The read-based, reference-free method kSNP (Gardner & Hall, 2013), a k-mer based approach for detecting SNVs, was used for phylogenetic analyses. kSNP is able to detect SNVs in a mix of both genomes with unassembled reads, and assembled genomes. kSNP k-mer size was 19 (i.e., k=19), and a variant position had to be present in 90% of genomes for inclusion in the dataset (i.e., m=0.9). Additionally, kSNP requires a minimum of 10 reads to support a base call. Publically-available reference genomes (Table S1) and genomes from Wright et al. (Wright et al., 2014) were also included to improve phylogenetic resolution. Prior SNV analysis of *A. baumannii* genomes (Wright et al., 2014) identified SNV positions that are in recombinant regions, so SNVs were filtered to include only core, non-recombinant locations. These SNVs were then used as input into RAxML for maximum likelihood tree construction (Stamatakis, Ludwig, & Meier, 2005). Positions without reads for a given genome were treated as missing data.

For scenarios in which either no *a priori* information is available about population structure of strain sequence types, or for guidance as to which reference organisms to include for phylogenetic context, a read-based *in silico* multi-locus sequence typing (MLST) analysis was assessed. All allelic variants of the seven loci used in the *A. baumannii* MLST scheme were downloaded from the Acinetobacter MLST database (<http://www.pasteur.fr/recherche/genopole/PF8/Abaumannii.html>). Reads were aligned using BLAST, and parsed to retain those aligning over greater than 20bp at 100% identity to generate aligned read counts. The top allele match was then identified for each locus using a RPKM calculation (number of mapped Reads Per Kilobase of gene length per Million total mapped reads).

To assess the potential for a barcoding bias to skew read distribution, reads were mapped to a reference genome, *A. baumannii* ACICU using bwa, a software package for mapping low-divergent sequences against a reference genome (H. Li & Durbin, 2009). Read density along the *A. baumannii* ACICU chromosome was then quantified in non-overlapping 10-kb segments.

Genome content analysis

The ability to use SISPA-Seq reads to survey genome content was assessed using a publically available database of antibiotic resistance genes, ARG_ANNOT, a comprehensive collection of antibiotic resistance genes and variants (Gupta et al., 2014). Reads were

aligned to the database using BLAST and were then parsed to retain those aligning at a minimum identity (e.g., 99%) and minimum length (e.g., 20 bases). Parsed aligned reads were then counted for each gene as input into a RPKM calculation as above.

Data and software availability

The SISPA-Seq reads generated in this study have been made available in the Short Read Archive at NCBI under BioProject PRJNA268655 (accession numbers for individual genomes are listed in Table S1). An analysis pipeline that performs SNV analysis and target gene mapping is available for download here: <https://github.com/JCVI-Cloud/SISPA-related>.

Results

This investigation evaluated the ability of SISPA-Seq data to 1) determine the phylogenetic placement of isolates using a read-based single nucleotide variant (SNV) detection method; and to 2) assess antibiotic resistance gene presence using a BLAST-based assessment of genome content. Both analysis approaches rely on primary sequence reads rather than assembled and annotated genomes. Nearly all (74 of the 75 strains) yielded at least 30 Mbp of sequence data (~8x genome coverage) and were analyzed further (Table S1). A large majority, 84% of the strains, had good representation of reads from both barcoded SISPA amplifications, with no more than 75% of reads originating from one barcode. The number of reads generated between the two DNA preparation methods did not vary. Additionally, there was no difference in the average fraction of reads that aligned to the *A. baumannii* ACICU reference genome (average 69.8% for boil prep vs 68.1% for HQ prep). Median genome-wide coverage was 65x when reads from both primer reactions were combined. Only six of 75 strains had lower than 20x coverage. The read density varied considerably across the genome when considered in 10 kbp regions (Figure 2). Read coverage across the genome varied much more widely in the SISPA-Seq data compared with reads from a standard library method for the same genome (Figure 2). Assembly of the SISPA-Seq reads was attempted using velvet (Zerbino, 2010), but the best assemblies had hundreds of contigs and low N50 values, so our analysis focused on methods that relied only on the primary reads.

Phylogeny

The kSNP program allows for reference-free SNV detection using a k-mer based approach. We combined SISPA-Seq read sets with draft and finished genomes in the kSNP analysis. kSNP detected 62,520 informative SNV positions in areas of the genome outside of known recombination regions (Wright et al., 2014). Of these SNVs, 18,593 were high-quality calls occurring in at least 90% of the genomes. The loss of SNVs using these criteria was primarily driven by a few low-coverage genomes and divergent strains (Figure 3A). For example, two strains, UH1107 and UH1507, were initially identified as *A. baumannii*, but SISPA-Seq data indicates that they are actually *A. calcoaceticus*. These two isolates each had 11,822 and 11,313 missing SNV data points respectively. The isolate with the lowest coverage, UH10407 (7.4x coverage), had 15,338 missing SNVs, due to the constraint of 10x coverage for kSNP to report a SNV.

These 18,593 high quality SNV positions were then used to infer the phylogeny using RAxML. Using reads from both SISPA primers yielded the most accurate and sensitive strain phylogenetic placement (Figure 4, Figure S1). An alternative phylogeny built on using 14,592 SNVs from just one primer resulted in strains being assigned to the same sub-clade as when data from both primers were used, but there was generally lower bootstrap support for branches (Figure S1). Genomes with a significant number of missing SNV values were still assigned to the correct clade. For example, UH0407 and UH0407B represent different DNA preparations of the same isolate, and though UH0407B only had 15x coverage, it was still placed in the same sub-clade as UH0407. WGS contigs and SISPA samples from all validation strains were located together on the same terminal branch, or correctly assigned to the correct subclade in the case of the Global Clone II (GC2, i.e., MLST ST2) isolates. Strains within each GC2 subclade are generally distinguished by fewer than 15 SNVs, so we did not expect that contigs and SISPA samples would always be adjacent in terminal branches.

Some clustering of SISPA-Seq isolates was observed. A count of the number of missing bases in each SISPA-Seq sample (i.e., N's) indicates that this clustering effect was largely due to missing data. Missing data was primarily related to coverage, with SISPA-Seq genomes with highest coverage having the fewest missing values (Figure 3B). When comparing the positive control isolates for which both WGS and SISPA sequence data was available, the original shotgun reads did align at those positions, while SISPA-Seq reads did not, indicating a systematic SISPA artifact, a limitation of the analysis. However, positions with missing SISPA-Seq data were generally unique to each genome, possibly due to barcode-specific biases. For example of the 62,520 SNV positions initially identified by kSNP, only 871 of these positions were missing from all SISPA genomes.

The *in silico* MLST analysis was able to identify the top allelic match for six out of seven loci for all validation strains (Table S3) indicating that this approach can be used for an initial assessment of population structure and strain diversity. The exception was the *rpoB* locus for MLST ST79 strains (i.e., clade E in Figure 4), where four allelic variants each scored the top RPKM value. These four variants (*rpoB*-5, -30, -48, and -54) differ by less than two base pairs over the 456 bases of the gene fragment used in the MLST scheme.

Gene content survey

SISPA-Seq reads were aligned to ARG_ANNOT, a database of known antibiotic resistance genes (Gupta et al., 2014), to evaluate the presence or absence of known resistance determinants in the 13 genomes that had been sequenced previously (Table 1). Using a “zero-tolerance” threshold in which a single SISPA-Seq read was considered a positive result, only one strain yielded a false positive. The *armA* RPKM was 4.5 for UH10007, an *armA*-negative strain based on WGS contigs, compared to a mean of RPKM 107.9 and a minimum value of 36.6 for true positive strains.

Correspondence between antibiotic resistance genotyping and antimicrobial susceptibility testing was investigated for amikacin, a drug for which there was a variable phenotype in this strain collection, and for which a resistance phenotype can be attributed to a specific gene. There was a strong association between presence of the amikacin resistance genes,

aphA6 and *armA*, and amikacin resistance (Table S5). However, there were two false positives (gene observed in SISPA-Seq data, but amikacin susceptible phenotype measured) and two false negatives (*aphA6* and *armA* not detected, but strain tested amikacin resistant). False positives were likely due to trace contamination in both UH2207 and UH1507, as the *armA* RPKM values were much lower compared to the MLST loci RPKM values (Table S5). The false negatives were likely due to low genome coverage for UH14008, and either a low coverage artifact or inaccurate susceptibility test for UH1207. Therefore, low RPKM values and data from low-coverage genomes should be interpreted with caution.

A. baumannii carry chromosomally-encoded β -lactamases, *bla*_{OXA-51}-like, of which many allelic variants have been described (Evans & Amyes, 2014), with some variants potentially conferring higher levels of resistance (Mitchell & Leonard, 2014). SISPA read analysis was able to unambiguously determine which variant was present, and whether the variant was linked to an upstream *ISAbal*, based on RPKM values in all validation strains (Table 1, Table S4, Figure S2).

A. *baumannii* population analysis

SISPA-Seq data indicate that the previously described sub-clade population structure of GC2 isolates (clades A-D in Figure 4) identified in Wright, et al. (Wright et al., 2014) is maintained across hospitals, and that resistance gene carriage is relatively consistent within each clade (Figure 4). The exceptions to this are *bla*_{OXA-23} and *aphA6* in clade B, which are variably present in strains in this sub-clade. These two genes are located on a pACICU2-like plasmid in this population, which was previously shown to be transferred from a clade A donor (i.e., UH6107) into a clade B recipient (i.e., UH6207) during a mixed infection (Wright et al., 2014). Additional non-GC2 lineages are also present in the study population, typically with lower resistance gene carriage than the GC2 isolates. However, all strains in clade E (MLST ST79) carry *bla*_{OXA-24/40}, for example, which only appears in another non-GC2 strain in this population. The presence of *bla*_{OXA-24/40} in UH8307 suggests that this strain would make it a good candidate for additional sequencing to investigate the genomic context of the resistance gene in a new clade.

Discussion

Our analysis showed that SISPA-Seq data from a large bacterial strain collection was able to accurately characterize isolates phylogenetically using SNVs and on the basis of resistance gene presence or absence, even for genomes with only about 8x coverage (UH14008 and UH10407). Over 18,000 SNV marker loci contributed to the phylogeny, considerable more than are typically used for MLST analysis.

SISPA-seq should be useful for quantifying changes in the relative abundance of dominant lineages over time, leading to more rigorous models of population dynamics. This approach may also facilitate characterization of gene gain and loss events, including those associated with plasmids, transposons, and other mobile elements, and reveal how these are disseminated among lineages or populations. Furthermore, data can be used to survey isolates to identify those with novel genome content or the appearance of a new sequence type to target for additional more complete genome sequencing. A limitation of this method

is that it may not provide the high-resolution SNV data needed for identifying transmission routes in genomic epidemiology applications where only a few key genomic positions must be defined at high confidence. For example, we found that only about 30% of genome positions identified by kSNP were covered adequately (10x coverage) in all 74 strains for high-confidence determination of the genotype. 87% of positions had high-confidence SNV calls for 75% of the genomes. A limitation of SISPA is that read coverage across the target genome is less random than observed using standard library preparation methods, leading to low-coverage or missing data as well as some over-represented regions, presumably due to the contribution of barcode bases, in addition to random bases, in primer binding efficiency across the genome.

The cost of genome sequencing has declined dramatically in recent years, and it could be argued that it is inexpensive enough that survey approaches like SISPA-Seq are unnecessary. When considering the cost of sample preparation (DNA isolation and library construction), sequencing, and analysis (assembly, annotation, and comparative analysis) for large strain collections, the cost can rapidly escalate. SISPA-Seq takes advantage of a very simple sample preparation and a straightforward analysis pipeline that uses the raw reads. Sequencing costs vary by institution, but we estimated the cost of performing the SISPA-Seq reactions at \$32.20 per isolate, including labor and materials for the rapid boil preparation of DNA, DNA quality control, two SISPA amplification reactions per sample, and Illumina library construction and QC, based on pooling 48 samples (96 SISPA reactions) per sequencer run. This is approximately half of the cost of \$66.50 for labor and materials for NexteraXT (Illumina) reactions and quality control. In addition, the efficacy of NexteraXT on DNA from the rapid boil prep has not been investigated; it is possible that a more time-consuming and expensive DNA prep would be necessary to supply DNA of sufficient quality for NexteraXT libraries, further contributing to the cost difference with the SISPA approach.

SISPA-Seq analysis of 63 previously unsequenced isolates confirmed our previous observation that isolates and resistance genes in this integrated health care system from one Midwestern US city are not segregated by location (Wright et al., 2014). This suggests that these strains and resistance genes are moving among hospital locations, and that the majority of MDR strains causing infections in this health care system are GC2 sequence types. The presence of reference genomes from other hospitals and countries at nodes within the GC2 sub-clades (e.g., MDR-ZJ06 and MDR-TJ and TYT-1 from East Asia) supports the conclusion that the sub-clades diverged before strains entered the hospital system, but resistance gene patterns indicate that local evolutionary processes such as plasmid acquisitions have contributed to strain differentiation in this hospital system.

The SISPA technique has been used successfully on many different types of viruses, indicating that the method is robust across genome architecture (Rosseel et al., 2013), and therefore should be generalizable to other clinically relevant pathogens. However, *A. baumannii* genome size is typically around 4.0 Mbp, and larger genomes may need higher genome coverage to provide sufficient sensitivity. While attempts to assemble SISPA-Seq reads were not successful, assembly results could potentially be improved by use of better designed primers that incorporate genome-specific base composition information (Djikeng & Spiro, 2009) or by using three or more barcodes per sample.

The results of this study support the application of SISPA-Seq to high resolution analyses of bacterial population structure and resistance gene mapping. Genome coverage and sequence quality were sufficient for detection of clinically significant antibiotic resistance genes by comparison of reads a database known resistance genes, were able to distinguish between allelic variants of the same gene, and importantly resistance genotyping was consistent with susceptibility testing. This read-based approach could be further extended to determining the presence of other significant traits such as those associated with virulence or metabolic capabilities. The clear advantages of the SISPA-Seq method include inexpensive and rapid DNA preparation and a typical total cost less than one-half that of standard genome sequencing. This point is particularly relevant when characterizing large strain collections where the cost between standard WGS and SISPA-Seq is appreciable. SISPA-Seq could be applied to survey whole genome content of large strain collections to identify strains that are maximally informative for more complete sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Numbers R01AI072219 and R01AI063517 to RAB and R01GM094403 to MDA. This study was supported in part by funds and/or facilities provided by the Cleveland Department of Veterans Affairs, the Veterans Affairs Merit Review Program Award 1101BX001974 and the Geriatric Research Education and Clinical Center VISN 10 to RAB. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Veterans Administration. We thank the JCVI sequencing group for producing Illumina sequence data.

References

- Castrignano SB, Nagasse-Sugahara TK, Kisielius JJ, Ueda-Ito M, Brandao PE, Curti SP. Two novel circo-like viruses detected in human feces: complete genome sequencing and electron microscopy analysis. *Virus Res.* 2013; 178(2):364–373. DOI: 10.1016/j.virusres.2013.09.018 [PubMed: 24055464]
- Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, ... Bentley SD. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet.* 2014; 46(3): 305–309. DOI: 10.1038/ng.2895 [PubMed: 24509479]
- DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* 1995; 23(22):4742–4743. [PubMed: 8524672]
- Decker BK, Perez F, Hujer AM, Hujer KM, Hall GS, Jacobs MR, ... Bonomo RA. Longitudinal analysis of the temporal evolution of *Acinetobacter baumannii* strains in Ohio, USA, by using rapid automated typing methods. *PLoS One.* 2012; 7(4):e33443.doi: 10.1371/journal.pone.0033443 [PubMed: 22511922]
- Depew J, Zhou B, McCorrison JM, Wentworth DE, Purushe J, Koroleva G, Fouts DE. Sequencing viral genomes from a single isolated plaque. *Virol J.* 2013; 10:181.doi: 10.1186/1743-422X-10-181 [PubMed: 23742765]
- Diancourt L, Passet V, Nemec A, Dijkshoorn L, Brisse S. The population structure of *Acinetobacter baumannii*: expanding multiresistant clones from an ancestral susceptible genetic pool. *PLoS One.* 2010; 5(4):e10034.doi: 10.1371/journal.pone.0010034 [PubMed: 20383326]
- Djikeng A, Halpin R, Kuzmickas R, Depasse J, Feldblyum J, Sengamalay N, ... Spiro DJ. Viral genome sequencing by random priming methods. *BMC Genomics.* 2008; 9:5.doi: 10.1186/1471-2164-9-5 [PubMed: 18179705]

- Djikeng A, Spiro D. Advancing full length genome sequencing for human RNA viral pathogens. *Future Virol.* 2009; 4(1):47–53. DOI: 10.2217/17460794.4.1.47 [PubMed: 19884976]
- Evans BA, Amyes SG. OXA beta-lactamases. *Clin Microbiol Rev.* 2014; 27(2):241–263. DOI: 10.1128/CMR.00117-13 [PubMed: 24696435]
- Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, ... Walker AS. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med.* 2013; 369(13): 1195–1205. DOI: 10.1056/NEJMoa1216064 [PubMed: 24066741]
- Gardner SN, Hall BG. 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.* 2013; 8(12):e81760.doi: 10.1371/journal.pone.0081760 [PubMed: 24349125]
- Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol.* 2014; 15(11):538.doi: 10.1186/s13059-014-0538-4 [PubMed: 25418119]
- Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother.* 2014; 58(1):212–220. DOI: 10.1128/AAC.01310-13 [PubMed: 24145532]
- Hamouda A, Evans BA, Towner KJ, Amyes SG. Characterization of epidemiologically unrelated *Acinetobacter baumannii* isolates from four continents by use of multilocus sequence typing, pulsed-field gel electrophoresis, and sequence-based typing of bla(OXA-51-like) genes. *J Clin Microbiol.* 2010; 48(7):2476–2483. DOI: 10.1128/JCM.02431-09 [PubMed: 20421437]
- Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, ... Dougan G. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet.* 2008; 40(8):987–993. DOI: 10.1038/ng.195 [PubMed: 18660809]
- Katz LS, Petkau A, Beaulaurier J, Tyler S, Antonova ES, Turnsek MA, ... Tarr CL. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio.* 2013; 4(4):doi: 10.1128/mBio.00398-13
- Koser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.* 2014; 30(9):401–407. DOI: 10.1016/j.tig.2014.07.003 [PubMed: 25096945]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25(14):1754–1760. DOI: 10.1093/bioinformatics/btp324 [PubMed: 19451168]
- Li LM, Grassly NC, Fraser C. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol.* 2014; 15(11):541.doi: 10.1186/s13059-014-0541-9 [PubMed: 25418281]
- Long SW, Beres SB, Olsen RJ, Musser JM. Absence of Patient-to-Patient Intrahospital Transmission of *Staphylococcus aureus* as Determined by Whole-Genome Sequencing. *MBio.* 2014; 5(5):doi: 10.1128/mBio.01692-14
- Mitchell JM, Leonard DA. Common Clinical Substitutions Enhance the Carbapenemase Activity of OXA-51-Like Class D beta-Lactamases from *Acinetobacter* spp. *Antimicrob Agents Chemother.* 2014; 58(11):7015–7016. DOI: 10.1128/AAC.03651-14 [PubMed: 25155607]
- Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, ... Musser JM. Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. *Proc Natl Acad Sci U S A.* 2014; 111(17):E1768–1776. DOI: 10.1073/pnas.1403138111 [PubMed: 24733896]
- Neill JD, Bayles DO, Ridpath JF. Simultaneous rapid sequencing of multiple RNA virus genomes. *J Virol Methods.* 2014; 201:68–72. DOI: 10.1016/j.jviromet.2014.02.016 [PubMed: 24589514]
- Perez F, Endimiani A, Ray AJ, Decker BK, Wallace CJ, Hujer KM, ... Bonomo RA. Carbapenem-resistant *Acinetobacter baumannii* and *Klebsiella pneumoniae* across a hospital system: impact of post-acute care facilities on dissemination. *J Antimicrob Chemother.* 2010; 65(8):1807–1818. DOI: 10.1093/jac/dkq191 [PubMed: 20513702]
- Rosseel T, Van Borm S, Vandenbussche F, Hoffmann B, van den Berg T, Beer M, Hoper D. The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for

- efficient massive parallel sequencing. PLoS One. 2013; 8(9):e76144.doi: 10.1371/journal.pone.0076144 [PubMed: 24086702]
- Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, ... Segre JA. Group, Nisc Comparative Sequencing Program. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med. 2012; 4(148):148ra116.doi: 10.1126/scitranslmed.3004129
- Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005; 21(4):456–463. DOI: 10.1093/bioinformatics/bti191 [PubMed: 15608047]
- Wisplinghoff H, Hippler C, Bartual SG, Haefs C, Stefanik D, Higgins PG, Seifert H. Molecular epidemiology of clinical *Acinetobacter baumannii* and *Acinetobacter* genomic species 13TU isolates using a multilocus sequencing typing scheme. Clin Microbiol Infect. 2008; 14(7):708–715. DOI: 10.1111/j.1469-0691.2008.02010.x [PubMed: 18558944]
- Wright MS, Haft DH, Harkins DM, Perez F, Hujer KM, Bajaksouzian S, ... Adams MD. New insights into dissemination and variation of the health care-associated pathogen *Acinetobacter baumannii* from genomic analysis. MBio. 2014; 5(1):e00963–00913. DOI: 10.1128/mBio.00963-13 [PubMed: 24449752]
- Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics. 2010; Chapter 11(Unit 11):15.doi: 10.1002/0471250953.bi1105s31

Highlights

SISPA-Seq analysis generated a high resolution *Acinetobacter baumannii* phylogeny

SISPA-Seq analysis was able to accurately assess antibiotic resistance gene content

SISPA-Seq is a cost-efficient and rapid whole genome survey method for bacteria

Figure 1. The SISPA-Seq method

A. Overview of SISPA-Seq with phylogenetic and gene content analysis pipeline. B. SISPA reaction details. C. General design of SISPA primers.

Figure 2. Sequence read distribution

Example comparison of mapped read density for UH7007 A) SISPA-Seq preparation individual barcoding primer A; B) primer B; and c) a dataset of combined reads from both primers; or D) standard Illumina library preparation ('WGS'). Reads were mapped to *A. baumannii* ACICU (CP000863) using bwa. Mean read density (reads per 10kb) for UH7007 WGS was 9504 (standard deviation (SD): 2244), compared to 8106 (SD: 4000) for the SISPA-Seq combined primers.

Figure 3. Relationship between read coverage and the number of informative SNVs

A) Effect of varying the minimum percentage of genomes required to have a base call at each SNV position for that SNV locus to be included in the SNV dataset, i.e., varying the parameter in 'm' in kSNP. B) Relationship between genome coverage and number of missing SNV values (i.e., N's) for each genome.

Figure 4. kSNP-based phylogeny of SISPA-Seq and publicly available *A. baumannii* genomes
 The presence of selected antibiotic resistance genes is shown. Primary clades identified in (Wright et al., 2014) are indicated by colored branches, and node labels A–E. Strain names are color-coded based on the hospital from which the strain originated in the integrated hospital system. Bootstrap values for nodes with >50% support are indicated.

Table 1

Sensitivity and specificity of detection for selected AbR genes in validation strains.

	True +	True –	False +	False –
<i>bla</i> _{OXA-23}	4	9	0	0
<i>bla</i> _{OXA-40}	2	11	0	0
<i>bla</i> _{TEM-1}	6	7	0	0
<i>armA</i>	2	10	1 [*]	0
<i>aphA1</i>	7	6	0	0
<i>aphA6</i>	4	9	0	0
<i>aadA1</i>	3	10	0	0
<i>aac6-Ib</i>	2	11	0	0
<i>aac3-II</i>	3	10	0	0
correct <i>bla</i> _{OXA-51} variant	13	0	0	0
<i>bla</i> _{OXA-51} /IS <i>AbaI</i> association	4	9	0	0

* False positive is strain UH10007, where the *armA* RPKM value was 4.5 compared to a mean of RPKM 107.9 and a minimum value of 36.6 for true positive strains.