Application Note

Agilent
Trusted Answers

Utilization of Agilent SureSelect Target Enrichment for Whole Genome Sequencing of Viruses and Bacteria

Authors

Rachel J. Williams, Helena Tutill, Sunando Roy, Erika Yara Romero, Charlotte A. Williams and Judith Breuer

Division of Infection and Immunity, University College London, Gower Street, London WC1E 6BT, UK

Daniel P. Depledge

Department of Medicine, New York University School of Medicine, New York, NY 10016, USA

Abstract

In this Application Note, we present modifications that should be considered when using the Agilent SureSelect XT, SureSelect XT HS, or SureSelect XT Low Input library preparation protocols for targeted enrichment of viral and bacterial whole genomes from various sample types.

Introduction

The ability to perform whole genome sequencing of viruses and bacteria directly from clinical research samples is important for understanding the genetics of host-virus interactions. Billions of pathogen genome copies may be found in 1 mL of a clinical research sample. However, the proportion of host nucleic acid in an extract massively outweighs that of pathogen nucleic acid, as is reflected by directly sequencing such extracts[1,2]. To overcome this issue, we have applied several modifications to the standard SureSelect protocols for the enrichment and sequencing of pathogen genomes. These genomes are taken from various clinical research samples including blood, sera, plasma, stool, urine, sputum, and nasopharyngeal aspirates. Probe libraries have been designed to capture whole genomes of a selection of RNA viruses, DNA viruses and bacteria (Table 1). These libraries are based on comprehensive sequence databases for each organism to allow for the deep sequencing of whole pathogen genomes. Capture probes have also been designed for specific regions of the *Mycobacterium tuberculosis* genome, targeting all known regions involved in drug resistance.

Table 1. Pathogen probe designs.

Pathogen	Genome/ Target Size (Approx. kb)	Pathogen	Genome/ Target Size (Approx. kb)
RNA viruses			
Enterovirus	7.5	Influenza A and B	13.5 to 14.5
Enterovirus (A-D)/rhinovirus	7.2 to 8.5	Norovirus	7.5
Hepatitis C virus	9.6	Parainfluenza 1–3	15
HIV-1	9.7	Respiratory viruses	13 to 15
HIV-2	10	Respiratory syncytial virus	15
DNA viruses			
Adenovirus	26 to 46	Herpes simplex virus 1 and 2	152 to 155
Cytomegalovirus	236	HPV and polyomaviruses (SV40, JC, and BK)	5.1 to 8
Epstein-Barr virus	170	Human herpes virus 6 and 7	145 to 160
Hepatitis B virus	3.2	Varicella zoster virus	125
Bacteria			
Chlamydia trachomatis	1000	*Neisseria meningitidis*	2200
Mycobacterium tuberculosis	4000	*Legionella pneumophila*	4000
Mycobacterium tuberculosis (selected gene targets)	160	N/A	N/A

This Application Note focuses on the optimization of the SureSelect protocols for whole genome sequencing of pathogens using Illumina sequencing platforms. Modifications to the standard SureSelect protocols include changes to the shearing conditions, increases in the number of pre- and post-capture PCR cycles, and a reduction in the amount of capture probe library added to the hybridization. The combination of pathogen, pathogen load, and sample type has a bearing on sequencing success. This Application Note is intended as a general guide and starting point, and users should further adapt the protocol for their samples and sequencing requirements.

Experimental guidelines, results and discussion

Recommended methods for quality control and processing of RNA and DNA samples with SureSelect library preparation kits and pathogen baits

Reagent kits

Agilent SureSelect XT, SureSelect XT HS, and SureSelect XT Low Input reagent kits are used for targeted enrichment of whole genomes from various sample types.

Guidelines for RNA and DNA extractions

As high-quality intact DNA or RNA is required, we recommend using a column or bead-based extraction method. Where recommended by the manufacturer and to improve yields, particularly for low biomass samples, synthetic carrier RNA should be used in RNA extractions as this will not interfere with the library preparation.

DNA/RNA purity should be measured using spectrophotometric analysis to obtain the 260/280 and 260/230 ratios. For DNA, these ratios should be as close as possible to 1.8 and 2.0, respectively. For RNA, both ratios should be close to 2.0.

DNA and RNA concentrations should be measured with a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) or similar system. RNA concentrations may not be quantifiable prior to conversion to cDNA as from some sample types, e.g. stool, RNA yields can be below the limits of the quantification.

cDNA synthesis from RNA samples

For RNA viruses, cDNA should be synthesized from RNA extracts using Invitrogen random primers (Thermo Fisher; part number 48190011). We regularly use Invitrogen SuperScript IV reverse transcriptase (Thermo Fisher; part number 18090050) followed by the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module (New England Biolabs, Ipswich, MA, USA; part number E6111) using all RNA obtained from one extraction. Extract volumes should be reduced to the required volume by vacuum centrifugation at 65 °C. Avoid drying down the RNA completely as cDNA yields may be reduced. Column or bead-based (1:1.8 sample:beads ratio) methods can be used to purify the cDNA, with elution in nuclease-free H_2O (non-DEPC treated). The cDNA concentration should be measured by Qubit or a similar system.

Guidelines for starting input amount

Pathogen genome copies

The percentage of sequencing reads mapping to the pathogen genome (on-target reads, OTRs) can vary significantly depending on the quality of and the number of genome copies in the starting material. For examples of the relationship between genome copies and OTRs for an RNA (norovirus, n=414) and a DNA (human cytomegalovirus n=119) virus, see Figure 1. The pathogen genome copy input also informs some of the modifications to the library preparation protocol. Thus, where possible, the pathogen load of the DNA or RNA extract should be measured by qPCR or qRT-PCR, or otherwise estimated from those measured in previous extractions from the same sample. Particularly for RNA viruses, any estimation from previous extractions should be

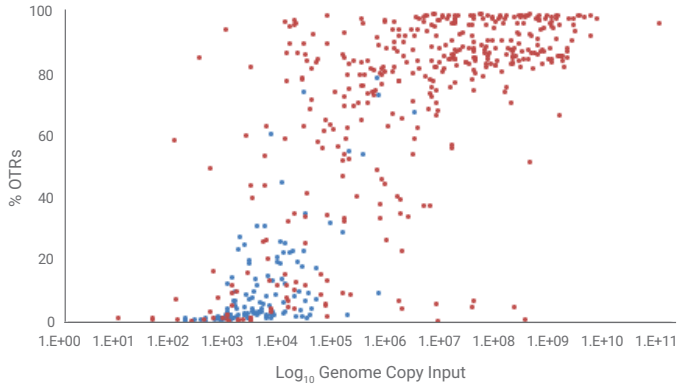

Figure 1. Estimated $\log_{10}$ genome copy input versus percentage on-target reads (%OTRs) for norovirus (red) and HCMV (blue).

Table 2. Summary of modifications to existing SureSelect protocols.

		XT 1 to 3 µg	XT 200 ng	XT HS/XT Low Input
Shearing parameters (Covaris E-series ultrasonicators)	Duty factor	5%	5%	
	Peak incident power (PIP)	175	175	
	Cycles per burst	200	200	
	Treatment time (seconds)	240	150	
	Bath temperature	4 to 6 °C	4 to 6 °C	
Precapture PCR	No. of cycles	8	12	See Table 3
Hybridization	Input amount	2 µg	2 µg	1 µg
	Pathogen capture library (probes)	1:10 dilution		
Postcapture PCR	No. of cycles	18 to 22*	18 to 22*	18 to 22*

**See comments in Postcapture PCR.*

used as a rough guide, as we are aware of both increases and decreases in pathogen-specific Ct values between extractions from the same sample.

To achieve ≥ 90% genome coverage and mean read depths of 20X (bacteria, larger DNA viruses) and 100X (small RNA/DNA viruses) for at least 90% of samples, as a starting point we recommend a minimum total genome copy input of 10^3 to 10^4 for the smaller RNA and DNA viruses, >10^4 for the larger DNA viruses, and >10^5 for bacteria. For some good quality DNA/RNA samples, these sequencing parameters can be obtained with lower genome copy input.

SureSelect XT 200 ng input protocol

Typical samples (i.e., those with more than the minimum genome copy input in ≤200 ng) should be processed using the SureSelect XT protocol for 200 ng input DNA. Samples with <150 ng DNA/cDNA input should be bulked with human genomic DNA (Promega, cat no. G1471) to bring the total input to between 150 and 200 ng. These samples should be processed using the same conditions as samples with 200 ng starting material.

In general, the more pathogen genome copies entered into the library preparation, the greater the OTRs, genome coverage and read depth achieved. Thus, for samples with low pathogen loads (i.e., close to or less than the suggested minimum genome copy input given above) and enough material, the amount added can be increased up to 500 ng with no additional protocol modifications required. Conversely, samples with high volumes but low DNA/cDNA concentrations should be concentrated by vacuum centrifugation to achieve the maximum number of pathogen genomes added to the library preparation.

Some samples will not achieve the minimum genome copy input in 500 ng material: this could be due to high DNA/cDNA concentrations or very low pathogen loads. In this case, if additional extract is available, consider using the 3 µg SureSelect protocol. Users should aim to enter all available material into the library preparation to increase the genome copy input for samples with low pathogen loads. Bulking human gDNA should be added to samples with <1 µg total input nucleic acid to bring the total input nucleic acid to a minimum of 1 µg. Samples with between 1 to 3 µg starting material should be processed using the same conditions.

SureSelect XT HS and SureSelect XT Low Input protocol

If feasible, 200 ng starting material should be used. For samples with <150 ng material, follow existing Agilent recommendations for reduced input amounts. Alternatively, for ease or when using the automation protocol for processing up to 96 samples simultaneously, all samples can be bulked with human gDNA to equivalent input amounts. When taking this approach, we recommend that a minimal amount of bulking gDNA is used. For samples with low pathogen loads and high volumes, increasing to 500 ng input and concentrating the sample respectively can be used as with the 200 ng input protocol.

Modifications to the existing SureSelect protocols

The adjustments to the SureSelect XT/SureSelect XT HS/SureSelect XT Low Input protocols are summarized in Table 2 and discussed below.

Sample shearing conditions

For subsequent sequencing using the longer read MiSeq kits, we recommend shearing to larger fragment sizes in the range of 250 to 300 bp (Figure 2). The shearing parameters provided in Table 2 can be used to fragment nucleic acids extracted

from all sample types. However, further optimization may be required to shear the reduced input amounts used for the SureSelect XT HS and SureSelect XT Low Input protocol.

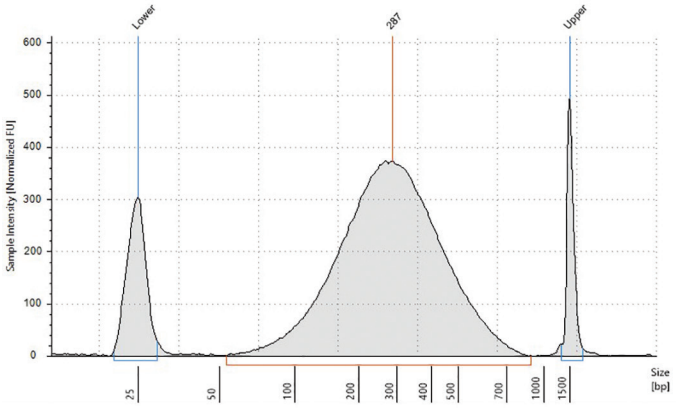

Figure 2. An example post-shearing Agilent High Sensitivity D1000 ScreenTape trace of an HIV-1 sample.

Pre-capture PCR

To ensure sufficient amplification of adapter-ligated libraries, the number of pre-capture PCR cycles used for each library prep method has been adjusted compared to the standard protocols (Table 2). For the SureSelect XT HS and SureSelect XT Low Input protocols, the number of pre-capture PCR cycles used is modified depending on sample input amount (Table 3). For the SureSelect XT 200 ng/3 µg protocols, the PCR cycles given in Table 2 are used irrespective of sample input amount.

Table 3. Number of pre-capture PCR cycles used during library preparation with the SureSelect XT HS and Low Input kits.

Sample Input Amount (ng)	No. of Pre-capture PCR Cycles	
	DNA Pathogens	RNA Viruses
251 to 500	5 to 6	8
100 to 250	7 to 8	10

Hybridization: Input amount (ng) and preparation of pathogen probe capture libraries

The recommended amount of material to add to the hybridization reaction is shown in Table 2. When this amount is not available, the entire prehybridization library should be concentrated using a vacuum concentrator to the volume stated in the protocol (3.4 µL for standard 200 ng/3 µg SureSelect XT and 12 µL for SureSelect XT HS/SureSelect XT Low Input). If concentrating the sample, avoid total dehydration of prehybridization libraries where possible, although this can become impractical when dealing with large sample numbers. Please note that the ScreenTape traces may show unmeasurable prehybridization DNA concentrations. However, these traces do not necessarily indicate library preparation failure (Figure 3); for these samples, all material should be added to the hybridization.

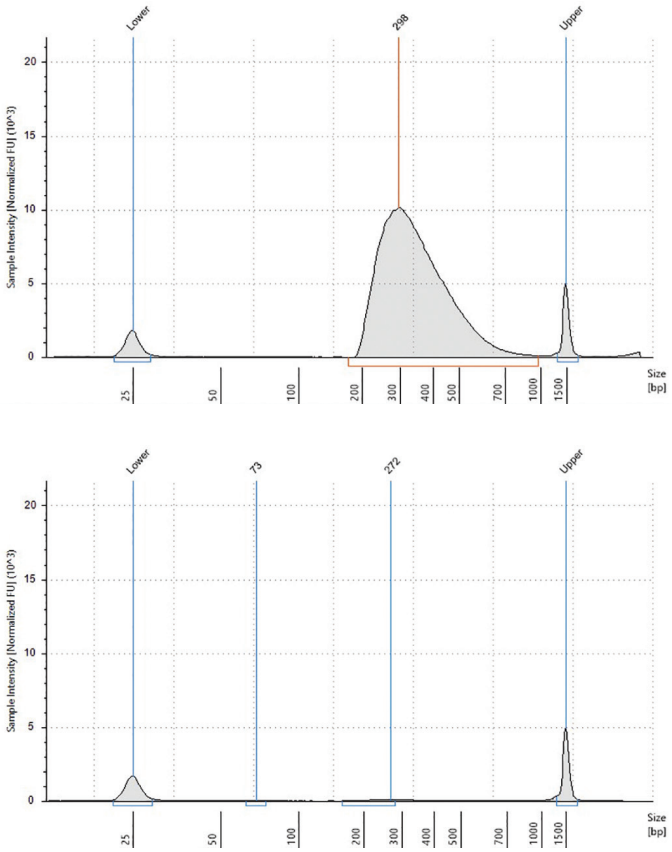

Figure 3. Examples of two prehybridization libraries that produced good final libraries that were successfully sequenced.

The SureSelect XT capture probe library should be prepared just before use. The probe library should be diluted (Table 2) in nuclease-free H_2O (non-DEPC treated) and immediately added to the capture library master mix. The volumes of diluted probes used are as in the standard protocols.

Postcapture PCR

Further optimization may be required to ensure that sufficient input material for Illumina sequencing is generated during the amplification of the postcapture libraries dependent on your specific pathogen of interest. Irrespective of the pathogen capture library used, we recommend using 18 cycles in the postcapture PCR as a starting point (Table 2). The number of PCR cycles can be increased with decreasing pathogen loads or decreased with increasing pathogen loads as necessary.

Sequencing strategy

The final libraries are sequenced using Illumina short-read sequencing platforms. The Illumina platform and kit used will depend on the number of libraries, the genome size, and the sequencing criteria required for the downstream analysis. For low frequency variant calling, one might need a higher minimum coverage at each base (>100X) to confidently identify variants. This should also feed into the multiplexing strategy. As mentioned previously, both the pathogen load and the quality of the starting material are critical to sequencing success. Determine the correct multiplexing strategy for your samples using Table 4. This provides a general guide to the number of libraries that can be pooled on a MiSeq or NextSeq run to achieve ≥90% genome coverage and mean read depths of 20X (for bacteria and the larger DNA viruses, i.e., CMV, EBV) and 100X (for the small RNA/DNA viruses). Figure 4 shows the percentage genome coverage and mean read depths achieved for the same norovirus (panel A) and HCMV (panel B) libraries analyzed in Figure 1.

Table 4. Pooling strategies for sequencing on Illumina MiSeq and NextSeq platforms.

Genome Size	MiSeq (v2, 500 cycles)	NextSeq (v2.5, mid output, 300 cycles)
Smaller viruses <15 kb	48 to 96 samples	NA
Larger viruses 150 to 200 kb	24 to 48 samples	96 to 144 samples
Bacteria 1 to 4 Mb	8 to 12 samples	48 to 96

Data analysis

The fastq files can be generated directly during the run in a MiSeq sequencer, or in the case of the NextSeq can be generated using Illumina bcl2fastq software. The fastq files are then checked for quality and adapters trimmed using a combination of tools like Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed fastq files can be mapped to a suitable reference using tools like BWA (http://bio-bwa.sourceforge.net/), Bowtie (http://bowtie-bio.sourceforge.net/index.shtml), or BBMap (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/) to generate a mapped BAM file. Trimmed reads can also be de novo assembled using SPAdes (http://cab.spbu.ru/software/spades/) into contigs or in case of bacterial genomes Unicycler (https://github.com/rrwick/Unicycler). The contigs can be used to then remap the reads. The BAM files produced either by mapping to a reference or mapping to contigs are further processed using SAMtools (http://samtools.sourceforge.net/) and Picard (https://broadinstitute.github.io/picard/) to remove duplicates as these would lead to incorrect identification of variants and variant frequencies in the population. These files are then ready for consensus extraction and variant calling using tools like VarScan 2.0 (http://dkoboldt.github.io/varscan/) or GATK (https://software.broadinstitute.org/gatk/) SNP caller.

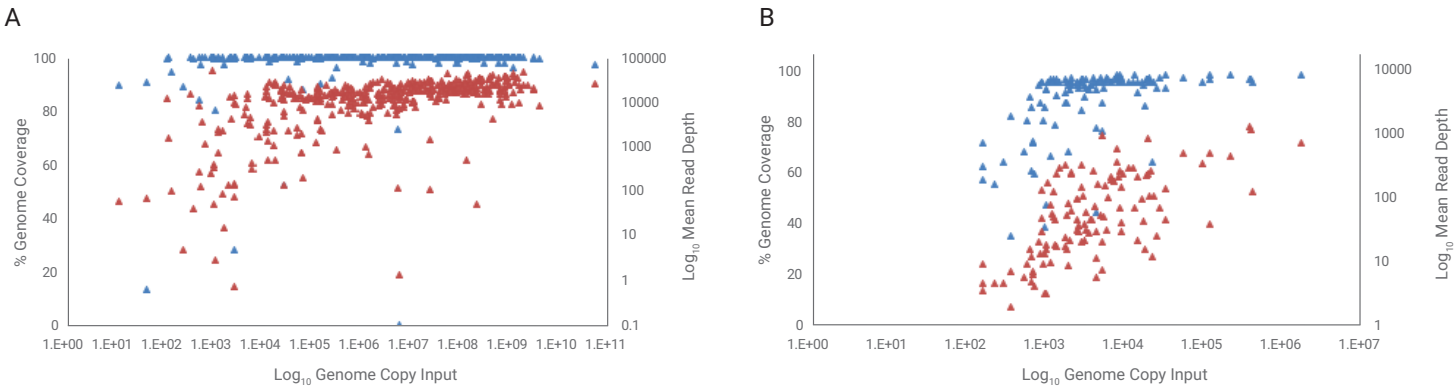

Figure 4. Percentage genome coverage (blue) and mean read depth (red) for norovirus (A) and HCMV (B). All libraries were sequenced on a MiSeq v2 500 cycle kit. The norovirus libraries were multiplexed at 48 samples and HCMV at 24 samples per run.

Conclusion

In this Application Note, we have described modifications to the SureSelect targeted enrichment protocols for deep sequencing of complete pathogen genomes from clinical research samples. Since our first publication in 2011[1], we have demonstrated the utility of this approach for identifying transmission events[3–5]; detection of minority variants including low frequency multidrug resistance[6,7]; and viral evolution[7–9]. Whole genome sequencing of pathogens directly from clinical research samples can also inform clinical research management of patients by rapidly identifying a full genetic resistance profile in one assay[10]. The methods are sensitive, generating whole genomes from samples containing a few hundred sequences[11], reproducible, and representative of the original sequence[12].

We have successfully sequenced pathogen genomes from many different types of clinical research samples including blood, stool, urine, CSF, sputum, amniotic fluid, formalin-fixed, and fresh tissue using our custom designed pathogen panels that are now available from Agilent. The methodology has the advantage of working well with fragmented nucleic acid, making it ideal for recovering genomes from archival and suboptimal samples. This recovery has been demonstrated in stored HIV-1 samples (Breuer and Leigh-Brown, unpublished) and ancient Mtb samples (Pallen and Breuer, unpublished)[13].

References

1. Depledge, D. P.; Palser, A. L.; Watson, S. J.*et al.* Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. *PLoS One* **2011,** *6*(11):e27805. doi:10.1371/journal.pone.0027805.
2. Thomson, E.; Ip, C. L. C.; Badhan, A. *et al.* Comparison of Next-Generation Sequencing Technologies for Comprehensive Assessment of Full-Length Hepatitis C Viral Genomes. *J. Clin. Microbiol.* **2016,** *54*(10), 2470–2484. doi:10.1128/JCM.00330-16.
3. Brown, J. R.; Roy, S.; Shah, D. *et al.* Transmission Dynamics in a Pediatric Hospital Using Full Genome Sequences. *Clin. Infect. Dis.* **2019,** *68*(2), 222–228. doi:10.1093/cid/ciy438.
4. Houldcroft, C. J.; Roy, S.; Morfopoulou, S. *et al.* Use of Whole-Genome Sequencing of Adenovirus in Immunocompromised Pediatric Patients to Identify Nosocomial Transmission and Mixed-Genotype Infection. *J. Infect. Dis.* **2018,** *218*(8), 1261–1271. doi:10.1093/infdis/jiy323.
5. Depledge, D. P.; Brown, J.; Macanovic, J. *et al.* Viral Genome Sequencing Proves Nosocomial Transmission of Fatal Varicella. *J. Infect. Dis.* **2016,** *214*(9), 1399–1402. doi:10.1093/infdis/jiw398.
6. Houldcroft, C. J.; Bryant, J. M.; Depledge, D. P. *et al.* Detection of Low Frequency Multi-Drug Resistance and Novel Putative Maribavir Resistance in Immunocompromised Pediatric Patients with Cytomegalovirus. *Front. Microbiol.* **2016,** *7,* 1317. doi:10.3389/fmicb.2016.01317.
7. Cudini, J.; Roy, S.; Houldcroft, C. J. *et al.* Human Cytomegalovirus Haplotype Reconstruction Reveals High Diversity due to Superinfection and Evidence of Within-Host Recombination. *Proc. Natl. Acad. Sci.* **2019,** *116*(12), 5693–5698. doi:10.1073/pnas.1818130116.
8. Depledge, D. P.; Kundu, S.; Jensen, N. J. *et al.* Deep Sequencing of Viral Genomes Provides Insight into the Evolution and Pathogenesis of Varicella Zoster Virus and Its Vaccine in Humans. *Mol. Biol. Evol.* **2014,** *31*(2), 397–409. doi:10.1093/molbev/mst210.
9. Weinert, L. A.; Depledge, D. P.; Kundu, S. *et al.* Rates of Vaccine Evolution Show Strong Effects of Latency: Implications for Varicella Zoster Virus Epidemiology. *Mol. Biol. Evol.* **2015,** *32*(4), 1020–1028. doi:10.1093/molbev/msu406.
10. Nimmo, C.; Doyle, R.; Burgess, C. *et al.* Rapid Identification of a Mycobacterium tuberculosis Full Genetic Drug Resistance Profile Through Whole Genome Sequencing Directly from Sputum. *Int. J. Infect. Dis.* **2017,** *62,* 44–46. doi:10.1016/j.ijid.2017.07.007.
11. Brown, J. R.; Roy, S.; Ruis, C. *et al.* Norovirus Whole-Genome Sequencing by SureSelect Target Enrichment: a Robust and Sensitive Method. *J. Clin. Microbiol.* **2016,** *54*(10), 2530–2537. doi:10.1128/JCM.01052-16.
12. Christiansen, M. T.; Brown, A. C.; Kundu, S. *et al.* Whole-Genome Enrichment and Sequencing of Chlamydia trachomatis Directly from Clinical Samples. *BMC Infect. Dis.* **2014,** *14*(1), 591. doi:10.1186/s12879-014-0591-3.
13. Kay, G. L.; Sergeant, M. J.; Zhou, Z. *et al.* Eighteenth-Century Genomes Show That Mixed Infections Were Common at Time of Peak Tuberculosis in Europe. *Nat. Commun.* **2015,** *6*(1), 6717. doi:10.1038/ncomms7717.

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

PR7000-2035

Printed in the USA, July 17, 2019
5994-0909EN

Agilent

Trusted Answers