OneSeq Target Enrichment

Simultaneous detection of genome-wide copy number changes, cnLOH, indels, and gene mutations

Application Note

Authors

Anniek De Witte
Kyeong-Soo Jeong
Arjun Vadapalli

Abstract

Array comparative genomic hybridization (aCGH) profiling is currently the gold standard for measuring constitutional chromosomal copy number changes. Although this technology has assisted in the identification of multiple novel microdeletion syndromes, the ultimate resolution for genomic interrogation is at the level of the base pair. Initial work based on whole genome sequencing (WGS) showed that a high number of reads with deep coverage across the genome is required to enable SNP calling alongside high resolution copy number analysis. This amount of sequencing is not compatible with the high-throughput, cost-sensitive requirements of most clinical research laboratories. This Application Note introduces OneSeq, a revolutionary all-in-one SureSelect target enrichment assay and accompanying SureCall analysis software that detects genome-wide copy number changes, copy neutral loss of heterozygosity (cnLOH), indels, and gene mutations in one comprehensive assay. We analyzed samples with known chromosomal aberrations and show that copy number changes from as small as 150 kb to whole chromosome triploidy, stretches of cnLOH, indels, and single base pair mutations, can be detected. We conclude that OneSeq target enrichment offers a practical solution for measuring genome-wide copy number changes and gene mutations simultaneously.

Agilent Technologies

Introduction

Congenital structural malformation and developmental disorders, including intellectual disability, autism, and attention deficit hyperactivity disorder (ADHD), are neuropsychiatric disorders that manifest in early childhood as deviations from the normal development. In the past 5 years, major advances have been made in the identification of specific genetic causes of these disorders. The methods used in previous studies include mainly karyotyping and fluorescence *in situ* hybridization (FISH) for fusion genes, aCGH for copy number changes, and direct sequencing and PCR for gene mutations. Although WGS has the potential to offer a single platform solution for determining the full range of abnormalities from single gene mutations to aneuploidy, the current cost and turnaround time of deep coverage WGS prevent it from being implemented in high-throughput clinical research laboratories. With targeted sequencing, only a subset of genes or defined genomic regions are sequenced, allowing time, expenses, and data storage to be focused on the regions of the genome of interest. However, it has not been possible to perform a genome-wide survey of copy number changes with targeted sequencing. The OneSeq target enrichment kits are designed so that genome-wide copy number changes, cnLOH, indels, and targeted mutations can be simultaneously determined. New algorithms have been implemented in Agilent SureCall Software v3.0 to allow for the streamlined analysis of OneSeq data.

Methods

Target enrichment panel design

The OneSeq target enrichment kits are based on the Agilent SureSelect technology and consist of a first set of backbone baits for genome-wide copy number change detection by comparing an experimental sample to a known reference sample. A second set of baits that target genomic regions with high minor allele frequency SNPs allows for the detection of cnLOH. A third set of baits that target specific regions of interest allows for the detection of mutations and indels. The catalog OneSeq Constitutional Research Panel (Figure 1) is a 28 Mb design. It includes baits (12 Mb) for a functional copy number resolution of 300 kb and cnLOH resolution of 5 Mb in the genome-wide backbone, and a higher 25–50 kb resolution in disease-associated ClinGen regions. It also includes all content (16 Mb) from the Agilent SureSelect Focused Exome Panel targeting disease-associated genes. The OneSeq CNV Backbone + Custom Panel allows for the addition of the genome-wide backbone to any custom target gene panel, up to 12 Mb, using Agilent SureDesign, a free web-based design application.

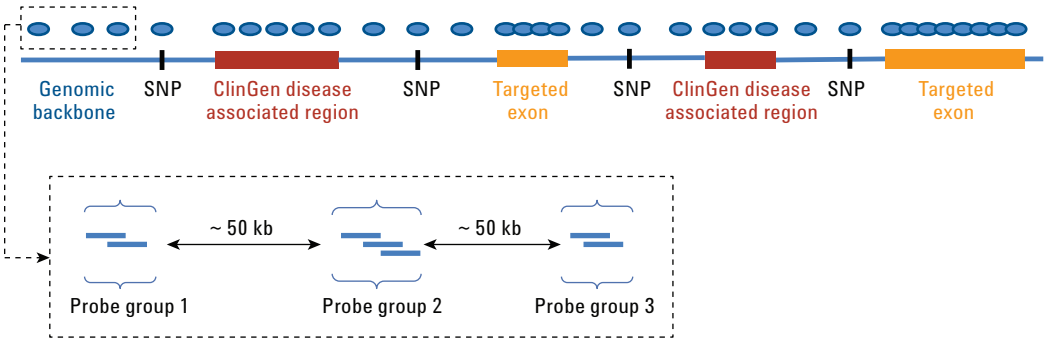

Figure 1. Bait design schema used for OneSeq target enrichment.

Sample preparation

Six DNA samples obtained from the Coriell Cell Repository (http://www.coriell.org/), NA03997, NA11419, NA08254, NA04592, NA02948, and NA20409, were processed by following the Agilent SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library Version B.1 using 200 ng DNA per sample and 10 post-capture PCR cycles. Agilent reference Male and Female samples were processed in parallel, and were used as controls in the data analysis. Each captured library was loaded on the Illumina HiSeq 2500 2×100 bp platform for sequencing. Adequate depth and coverage was achieved with 7 Gb of sequencing per sample.

Data analysis

The raw image files were processed by the Illumina base calling software with default parameters and FASTQ files were generated. The FASTQ files were then imported in the Agilent SureCall Software v3.0 (Figure 2). After removal of the adaptor sequences, the reads were aligned to the genome using the BWA alignment algorithm incorporated in SureCall.

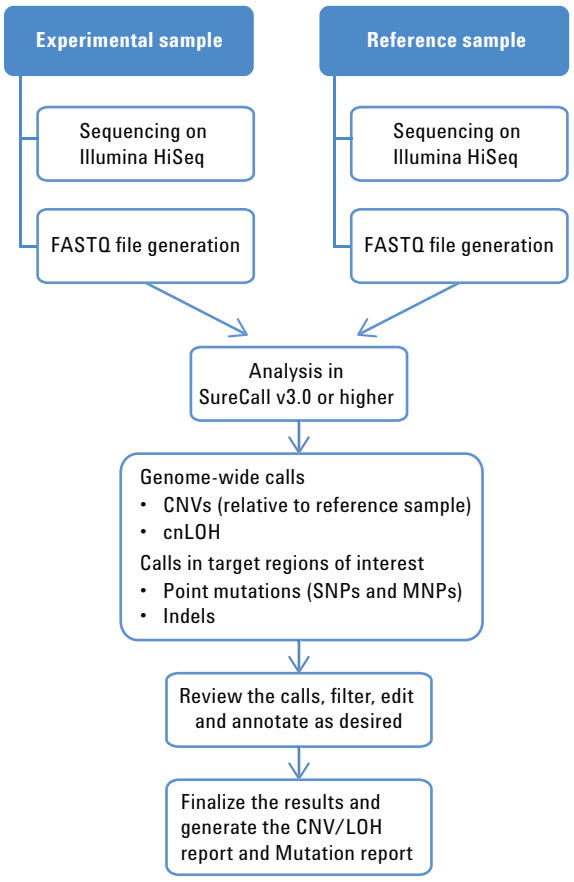

Figure 2. Steps for running a OneSeq analysis in Agilent SureCall.

Copy number changes are detected by comparing an experimental sample to a known reference sample (Figure 3). First, a summarization method is applied to capture the central tendency of read distributions over genomic regions covered by baits with the goal of minimizing the noise introduced due to outliers. Aberrations are called on log ratios that are generated by dividing the read depth of the sample over a reference at each bait interval. The log ratios are then subjected to an undecimated wavelet transform to detect abrupt changes or break points. The transformed log ratios are then analyzed at various genomic length scales. After combining and ranking the resulting breakpoints, a false discovery rate step is used to only select those that pass a certain threshold for statistical significance. The significant intervals are then further considered as candidates for amplifications and deletions by examining them at a finer resolution.

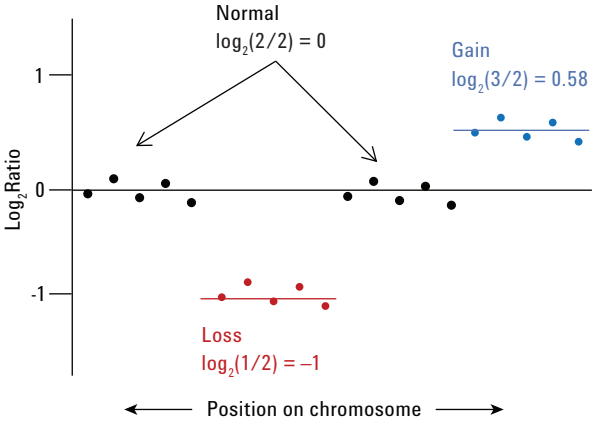

Figure 3. Determination of copy number changes in Agilent SureCall Software. Log_2ratios of the sequencing read-depth of the sample versus the sequencing read-depth of the control are plotted along the chromosome. No copy number change corresponds to a log_2ratio of 0 (black dots), a one copy loss corresponds to a log_2ratio of –1 (red dots), a one copy gain corresponds to a log_2ratio of 0.58 (blue dots).

The in-house developed SNP calling algorithm SNPPET was used to call point mutations and indels. The SNPPET algorithm has two basic steps. The first step is a quick search for variants where each locus is evaluated under two models. One model assumes that at the base under consideration, all non-reference alleles are due to sequencing error, and the next model considers each non-reference allele to be a true variant. The second step is a careful local search in the neighborhood of the variant. All potential variant combinations are evaluated as haplotypes, and adjusted for additional nearby variant sites.

The high minor allele frequency SNPs covered in the OneSeq backbone are used to determine cnLOH. Regions of cnLOH or UPD are located by identifying genomic regions with a statistically significant scarcity of heterozygous SNP calls. The LOH algorithm first attempts to assign the sample to a known population with 99% confidence using the allele frequencies determined at the available SNP locations. In cases where a sample cannot be assigned to a known population, average heterozygosity rates available from UCSC are used instead. Then, a sequential Fisher's test is used to score genomic regions that are enriched in SNPs that have lost heterozygosity. The final LOH score takes into account the presence of indels and multiple alleles that might be present at the candidate SNP locations. A more detailed description of the algorithms can be found in the SureCall help system.

Results and Discussion

QC data

For all eight samples, the percentage of reads on target ±100 bp was higher than 75%, and the number of duplicate reads was very low (Figure 4). This is similar to other SureSelect target enrichment kits, such as the Human All Exon V5 kit. The coverage was high, with more than 95% of the bases having at least 20 reads. As expected, the number of bases with at least 50 or 100 reads was significantly lower. When increasing the amount of sequencing from 7 Gb to 10 Gb, the number of bases with at least 50 reads was higher than 90% (data not shown).

Detection of large CNVs

We were able to detect the expected whole chromosome copy number changes in several samples with known aberrations. Figure 5 shows the detection of trisomy 13 in Coriell sample NA02948 with karyotype 47,XY,+13. The measured average $\log_2$ ratio for the entire chromosome was close to the expected $\log_2$ ratio value of 0.58 for a single copy gain.

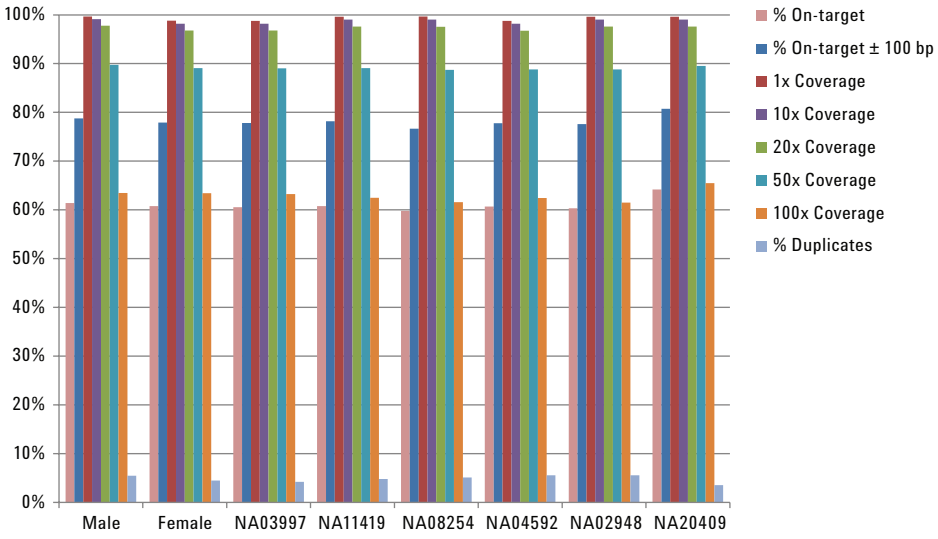

Figure 4. On-target coverage of the OneSeq Constitutional Research Panel.

Figure 5. OneSeq copy number data analysis in Agilent SureCall Software showing trisomy 13 in Coriell sample NA02948 (47,XY,+13). Each red plus sign represents a raw data point. Note the higher data point density in specific regions of interest. The blue shading and the blue line show the aberration call, indicating an entire chromosome 13 gain.

Comparison with aCGH

We compared the copy number profiles generated by OneSeq with those obtained by aCGH. The CGH data was generated on the Agilent CGH+SNP 4×180K microarrays. Table 1 shows a comparison of the CNV calls larger than 150 kb obtained by both methods for Coriell sample NA08254. The same eight CNVs could be detected with both methods. The genomic coordinates of the start and stop of the aberrations were not identical due to differences in aCGH probe and OneSeq bait placement, resulting in minor differences in aberration sizes. Figures 6 and 7 show the comparison data for a 13 Mb deletion on chromosome 13 and a 370 kb deletion on chromosome 6.

Table 1. CNVs larger than 150 kb detected in Coriell sample NA08254 by aCGH and OneSeq sorted from largest to smallest.

Chromosome	Aberration type	aCGH aberration size (kb)	OneSeq aberration size (kb)	OneSeq average log_2ratio
chr13	del	12427	13335	-0.89
chr15	del	2240	1667	-0.37
chr16	del	772	863	-0.43
chr14	amp	987	544	0.54
chr6	amp	370	372	0.61
chr2	del	828	307	-0.46
chr17	amp	163	201	0.49
chr22	amp	172	191	3.00

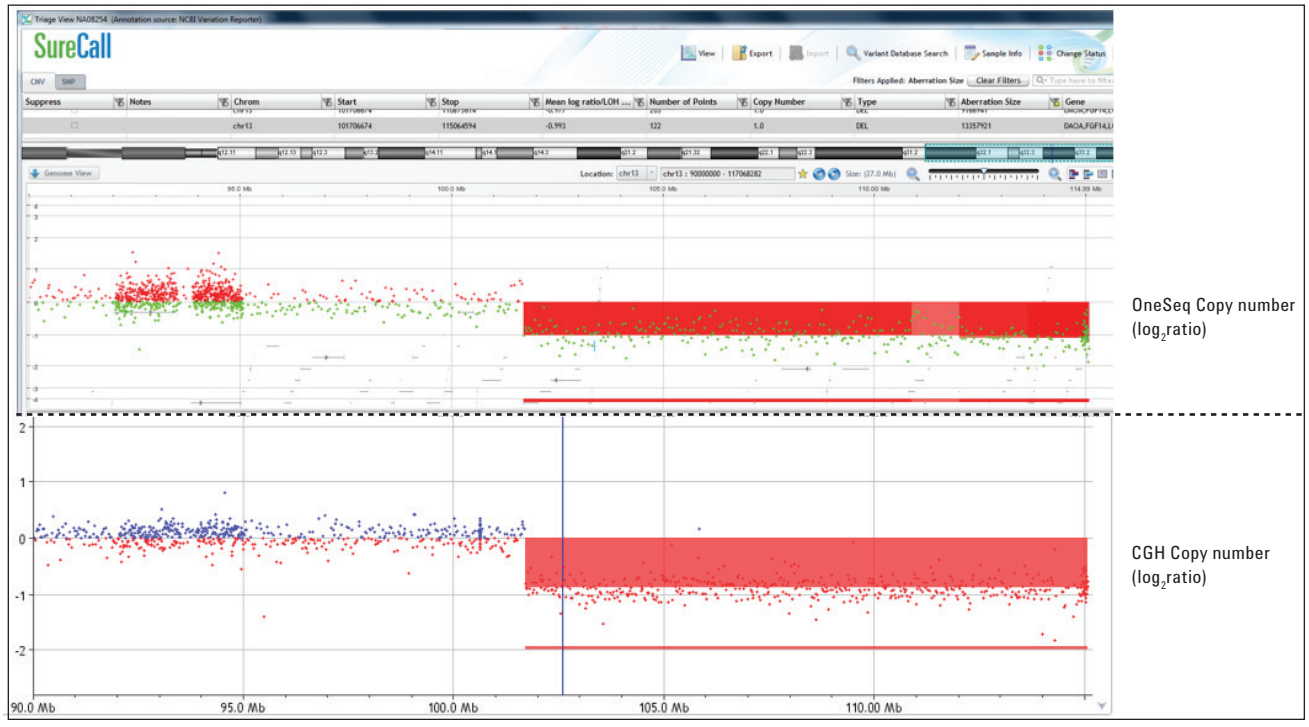

Figure 6. OneSeq copy number data analysis in Agilent SureCall Software (top panel) and aCGH copy number data analysis in Agilent CytoGenomics Software v3.0 (bottom panel) showing a 13 Mb deletion on chromosome 13 in Coriell sample NA08254.

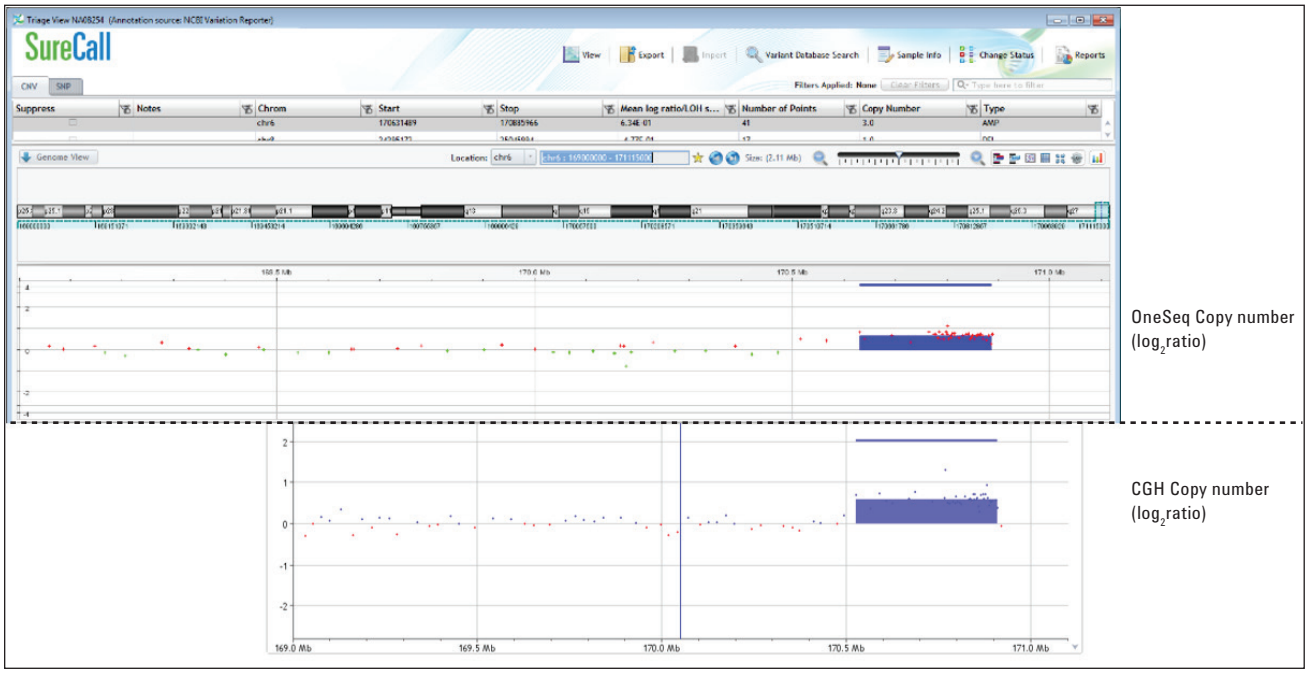

Figure 7. OneSeq copy number data analysis in Agilent SureCall Software (top panel) and aCGH copy number data analysis in Agilent CytoGenomics Software v3.0 (bottom panel) showing a 370 kb amplification on chromosome 6 in Coriell sample NA08254.

Detection of cnLOH

The detection of UPD 15 (uniparental disomy) in Coriell sample NA20409 with complete paternal UPD is shown in Figure 8. This UPD call was made with high confidence because the B-allele frequency of almost all SNPs was 0% or 100%, and virtually no heterozygous SNPs could be detected across the entire chromosome.

Figure 8. Copy number and LOH data analysis in Agilent SureCall Software data analysis showing UPD 15 in Coriell sample NA20409. The top panel shows the copy number data. Each red and green cross represents a raw data point. The entire chromosome, except for a known common CNV close to the centromere, is diploid. The bottom panel shows the LOH data. Each aqua dot represents the B allele frequency (BAF) of a SNP. The aqua shading indicates UPD of chromosome 15.

Detection of mutations and indels

The high read depth allowed for the detection of single point mutations and indels in the targeted regions of interest in all Coriell samples. Figure 9 shows an example of 26-bp indel in Coriell sample NA16382 known to carry a 26 base pair deletion beginning at nucleotide 1160 of the gene encoding methyl-CpG binding protein 2 (MECP2).

Conclusion

OneSeq target enrichment in combination with Agilent SureCall Software provides a streamlined method for detecting high resolution copy number changes by comparing the number of sequence reads in nonoverlapping windows between an unknown sample and a control sample. OneSeq enables cnLOH, SNP, and indel calling alongside copy number analysis. In contrast to WGS, OneSeq does not require a large amount of sequencing. OneSeq offers the convergence of existing genetic technologies while maintaining cost-effectiveness and throughput and can become the single platform solution for clinical research of a broad range of abnormalities from single gene mutations to aneuploidy.

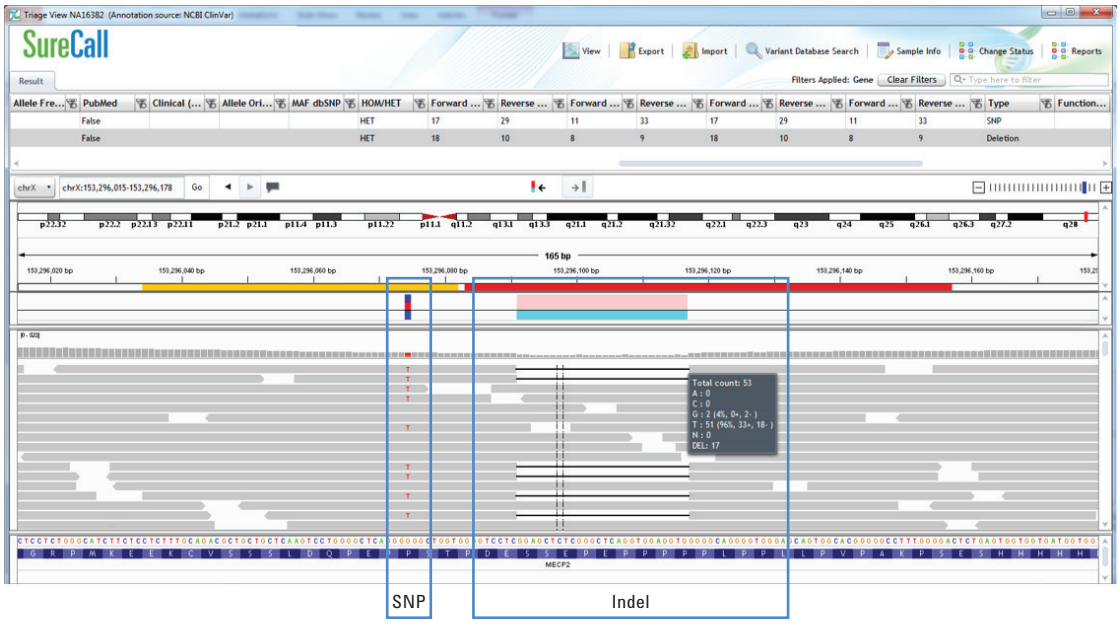

Figure 9. Identification of mutations and indels in Agilent SureCall Software showing a 26-bp indel (on the right) and a heterozygous SNP (on the left) in the MECP2 gene of Coriell sample NA16382.

FOR MORE INFORMATION:

www.agilent.com/genomics/oneseq

NGS resource page:
www.agilent.com/genomics/NGSResource

U.S. and Canada, call **800-227-9770** or for other regions, consult **www.agilent.com/genomics/contactus**

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

PR7000-00528

Published in the USA, June 24, 2016
5991-5631EN

Agilent Technologies