



## Comparative Genomic Hybridization (CGH) & Copy Number Variation (CNV)



## Application Note

# Detecting Copy Number Variants Associated with Basal Cell Carcinoma in an Arsenic-Exposed Population



**Rajini Haraksingh**<sup>1,2</sup>  
Graduate Student  
Yale University

### Contributing Authors:

**Alexander E. Urban**<sup>2,3</sup>, **Rajiv Kumar**<sup>4</sup>, **Eugene Gurzau**<sup>5</sup>, **Kathleen McCarty**<sup>6</sup>, **Michael Snyder**<sup>1</sup>

<sup>1</sup>Department of Genetics, Stanford School of Medicine, Stanford, California, USA

<sup>2</sup>Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA

<sup>3</sup>Department of Genetics, Yale University, New Haven, Connecticut, USA

<sup>4</sup>Division of Molecular Genetic Epidemiology, German Cancer Research Center, Heidelberg, Germany

<sup>5</sup>Environmental Health Center, Babes-Bolyai University, Cluj-Napoca, Romania

<sup>6</sup>Department of Epidemiology and Public Health, Division of Environmental Health Sciences, Yale University School of Medicine, New Haven, Connecticut, USA

## Introduction

Copy Number Variation (CNV) is a major source of human genomic variation comprising benign and pathological variants. These deletions and duplications of genomic regions have been mapped by dozens of studies employing diverse methods<sup>1,2,3</sup>. The thousands of reported CNVs in the human genome have been catalogued in the Database of Genomic Variants (<http://projects.tcag.ca/variation/>) and are now depicted in most genome browsing software. Known CNVs form a size continuum from small indels to whole chromosomal aneuploidies with current methods generally detecting events on the order of several kilobases. Despite the rapidly growing appreciation for the extent of CNV in the human genome, the frequency, biological relevance, and mechanisms by which they arise remain poorly understood<sup>4</sup>. It is clear, however, that CNVs are theoretically capable of reorganizing

functional elements of the genome by altering gene dosage, coding segments, and regulatory regions. Recently, several association studies have suggested that CNVs significantly impact certain disease phenotypes<sup>5,6,7</sup> though evidence for their functional consequences remains limited.

Performing CNV-phenotype association studies requires unbiased, genome-wide, high-resolution mapping of common and rare CNVs in a cost effective manner. Genome-wide CNV mapping technologies fall into three broad categories: array Comparative Genome Hybridization (aCGH) platforms, SNP genotyping platforms, and sequencing-based methods. Because the human genome is tiled at high resolution in a relatively unbiased fashion, our method of choice for large-scale association studies is aCGH. SNP genotyping platforms are somewhat biased in their genome-wide coverage as probes are necessarily limited to known SNP loci in the genome. However, SNP platforms can

in some cases provide absolute integer CNV genotypes while aCGH methods provide only relative CNV genotypes. Sequencing-based CNV detection methods are still prohibitively expensive for large population cohorts. Equally important for conducting CNV association studies is the availability of a comprehensive analysis platform for determining statistically and biologically significant events in the population. Here we demonstrate the use of NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D arrays (Roche NimbleGen) for CNV detection by aCGH coupled with Nexus Copy Number 4.1 (BioDiscovery), herein referred to as 'Nexus software,' for CNV analysis to identify CNVs associated with basal cell carcinoma (BCC) in a Romanian population exposed to arsenic in well water.

## Results

Genome-wide CNV detection in a study cohort is efficiently achieved by aCGH using NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D Arrays and analysis with Nexus Software

We performed high-resolution aCGH on three basal cell carcinoma research samples (herein referred to as 'cases') and one control research sample (herein referred to as 'control'). All four samples were from an ethnically homogenous Romanian population exposed to high levels of arsenic in well water. Genomic DNA was extracted from blood and analyzed versus reference DNA pooled from seven healthy females (Promega). The samples were labeled and co-hybridized to a NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D array. Arrays were washed and then scanned on a GenePix 4200A Scanner using GenePix 6.0 Software. Raw data were normalized using NimbleScan v2.4 Software (Roche NimbleGen). The normalized data were then processed using Nexus Software with default settings. The normalized data were also processed using NimbleScan Software using the segMNT algorithm with default settings (except minimum segment difference was set to 0.1). The data were loaded into Nexus Software as shown in Figure 1. In the bottom panel, the CNV profile of each sample is displayed as an individual track. Deletions are displayed in red below the track and duplications are displayed in green above the track. The CNV calls for each sample, detected by

Nexus software and NimbleScan software, are displayed as separate tracks. The samples can be sorted by a given factor. Here the samples are sorted by cancer phenotype (cases vs. control), as indicated. The top panel shows the chromosomal location (❶), a frequency plot showing the percentage of samples containing each CNV (❷), and annotation tracks displaying genes, exons, known CNVs, and miRNAs, as indicated. This display allows the user to visualize all the data in aggregate. Here, the frequency plot shows deletion of the entire X chromosome in all samples, which is expected because all research case samples are from males and aCGH was performed using female reference gDNA. In addition, it is clear that there is a common deletion in all samples on chromosome 1.

All CNV predictions are also displayed in an aggregate table as shown in Figure 2. Here, vital statistics about each predicted CNV, such as breakpoint coordinates, region length, cytoband location, type of event, number of genes, frequency in population, P-value, and % overlap with known CNVs are displayed. By specifying the aggregate percentage cutoff at 35%, we ensure that events in this report occur in at least two of the eight prediction tracks. Individual CNVs can be visualized by right clicking in the Nexus Software, ensembl, or UCSC



▲ **Figure 1:** Whole-Genome view of CNV calls displayed using Nexus software.

Nexus - BCC CNV Analysis (Human NCBI Build 36.1)

File Help

Data Set Comparisons External Data Results

Genome Chromosome Summary Table Aggregate Aggregate Participation

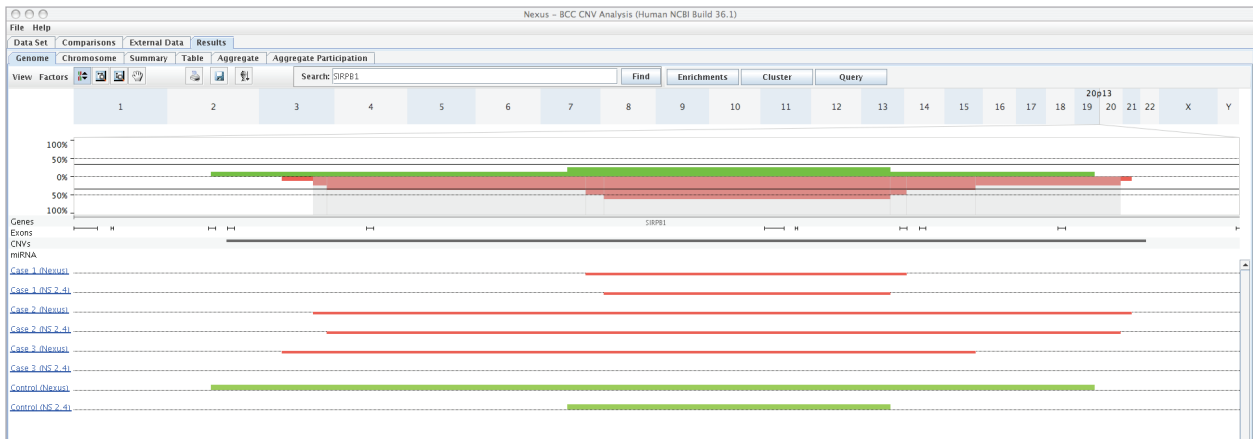
Export TXT View Annotations Enrichment Significant Peaks

Region	Region Length	Cytoband Location	Event	Genes	Frequency %	P-Value	% of CNV Overlap
chr8:39,352,437-39,453,190	100,753	p11.23	CN Gain	2	100	0	100
chr8:39,454,265-39,464,873	10,608	p11.23	CN Gain	1	100	0	100
chr8:39,466,412-39,500,625	34,213	p11.23 - p11.22	CN Gain	1	100	0	100
chr1:150,824,641-150,855,000	30,359	q21.3	CN Loss	2	75	0	100
chr3:164,009,807-164,021,177	11,370	q21.3	CN Loss	0	62.5	0	100
chr7:142,147,192-142,177,000	29,808	q11.23	CN Gain	1	62.5	0	100
chr20:1,518,332-1,532,000	13,668	q11.23	CN Loss	1	62.5	0	100
chr1:241,114,716-241,231,363	116,647	q43	CN Gain	0	50	0	100
chr2:146,579,930-146,591,952	12,022	q22.3	CN Gain	0	50	0	100
chr3:164,028,400-164,098,216	69,816	q26.1	CN Gain	0	50	0	100
chr4:9,822,140-9,841,281	19,141	p16.1	CN Loss	0	50	0	100
chr4:9,821,084-9,842,922	21,838	p16.1	CN Gain	0	50	0	100
chr4:69,055,072-69,117,175	62,103	q13.2	CN Loss	1	50	0	100
chr4:69,129,314-69,165,991	36,677	q13.2	CN Loss	0	50	0	100
chr11:5,743,021-5,765,082	22,061	p15.4	CN Loss	1	50	0	100
chr14:105,610,868-105,632,470	21,602	q32.33	CN Loss	0	50	0	100
chr15:21,015,080-21,143,973	128,893	q11.2	CN Gain	0	50	0.029	61.146
chr16:23,034-46,264	23,230	p13.3	CN Gain	2	50	0	6.324
chr18:61,869,221-61,882,768	13,547	q22.1	CN Gain	0	50	0	59.947
chr1:148,135,671-148,179,437	43,766	q21.2	CN Gain	5	37.5	0.04	0
chr1:200,676,258-200,705,322	29,064	q32.1	CN Gain	1	37.5	0.04	0
chr1:200,706,009-200,794,260	88,251	q32.1	CN Gain	1	37.5	0.04	0
chr2:87,896,955-88,051,094	154,139	p11.2	CN Loss	1	37.5	0	89.048
chr3:185,774,741-185,855,639	80,898	q27.1	CN Gain	1	37.5	0.013	0
chr4:69,524,428-69,568,514	44,086	q13.2	CN Loss	1	37.5	0.001	100
chr5:1,288,392-1,301,325	12,933	p15.33	CN Gain	1	37.5	0.001	0
chr5:21,328,131-21,387,332	59,201	p14.3	CN Loss	0	37.5	0.002	100
chr6:32,568,525-32,591,594	23,069	p21.32	CN Loss	0	37.5	0	100
chr8:47,843,894-47,957,880	113,986	q11.1	CN Gain	1	37.5	0.006	0
chr9:22,487,055-22,492,478	5,423	p21.3	CN Loss	0	37.5	0.011	100
chr9:40,467,126-40,573,868	106,742	p12	CN Loss	0	37.5	0.011	100
chr11:4,923,694-4,933,642	9,948	p15.4	CN Gain	2	37.5	0	100
chr12:28,832-225,731	196,899	p13.33	CN Gain	3	37.5	0.018	18.281
chr14:18,560,017-19,009,529	449,512	q11.1	CN Loss	1	37.5	0.01	100
chr14:19,252,578-19,493,998	241,420	q11.2	CN Loss	6	37.5	0.01	100
chr14:100,089,296-100,419,806	330,510	q32.2 - q32.31	CN Gain	5	37.5	0.041	0
chr15:19,875,551-19,883,740	8,189	q11.2	CN Loss	0	37.5	0	100
chr22:14,479,719-14,519,176	39,457	q11.1	CN Loss	0	37.5	0.028	100
chr22:14,781,238-14,808,733	27,495	q11.1	CN Loss	0	37.5	0.028	100
chr22:17,038,241-17,255,956	217,715	q11.21	CN Loss	2	37.5	0.028	100

P-Value cut-off 0.05 Aggregate % cut-off 35.0  Peaks only

Ready

▲ **Figure 2:** Aggregate table view of CNV calls displayed using Nexus software.

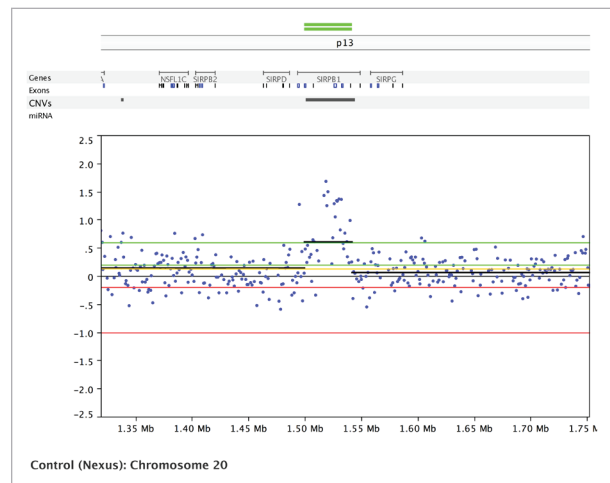


▲ **Figure 3:** Zoom-in view of a region of chromosome 20 harboring the SIRPB1 gene.

genome browsers.

SIRPB1 contains a deletion in basal cell carcinoma case samples and a duplication in the control sample

Figure 3 shows a zoom-in view of an interesting result from the table displayed using Nexus Software (Figure 2, 7<sup>th</sup> entry). The table indicates that a loss is present in 62.5% of the datasets (5/8 datasets in this case). Here we see that the SIRPB1 gene contains deletions in all case samples (not detected by NimbleScan Software in case 3). Interestingly, the control contains a duplication at this locus. We can easily visualize the extent of the CNV event in each individual sample and can access the coordinates of the individual breakpoints by moving the mouse over the CNV. We can determine exactly which exons of the gene are affected in each individual sample. Here we see that these CNVs lie in a known CNV region. While this is not a statistically significant finding due to the limited number of samples, this example illustrates how cancer-associated CNVs can be identified using these methods. Interestingly, this result may be biologically relevant as SIRPB1 belongs to the immunoglobulin superfamily and negatively regulates receptor tyrosine kinase-coupled signal transduction pathways that may affect cancer phenotypes. Additionally, SIRPB1 expression in mouse lung was



▲ **Figure 4:** Drill down of normalized data points on chromosome 20 for a control sample with CNV calls produced by Nexus Software. The horizontal axis shows the genome position along the chromosome. The vertical axis shows the  $\log_2$  ratio of test sample/reference sample.

shown to be affected by arsenic exposure<sup>8</sup>.

In addition to the summary view of the sample cohort, the raw data for an individual sample can be viewed using the drill down tool of the Nexus software. Figure 4 shows a drill down of the CNV displayed in Figure 3 for the control sample. Using this tool, the individual probes comprising the duplication in the region of the

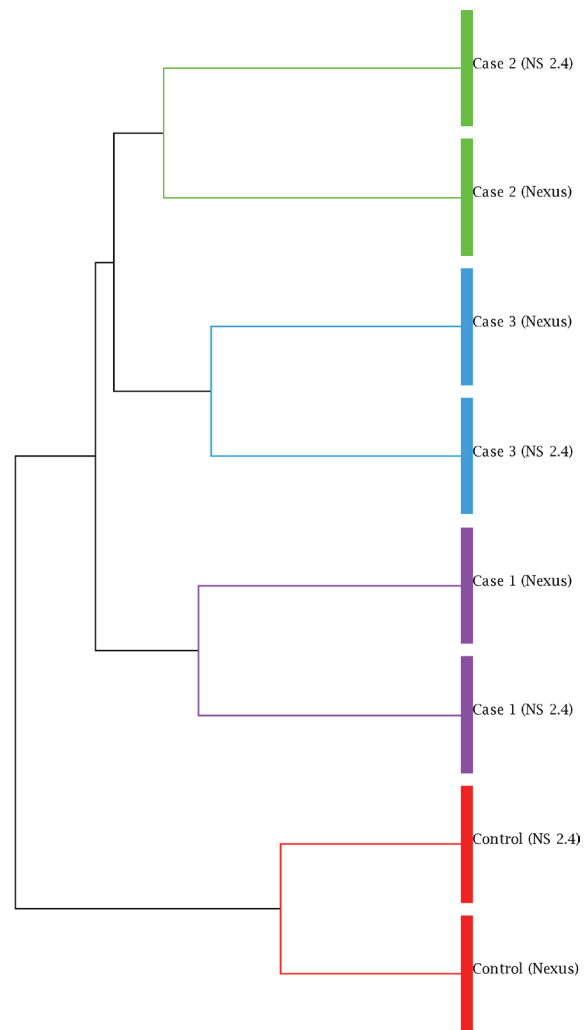
SIRPB1 gene are seen.

CNV profiles of Cases and Control cluster separately

Using the cluster tool, complete linkage hierarchical clustering based on the CNV profiles of all samples was performed. Figure 5 shows that the case samples cluster separately from the control sample.

Nexus Software and NimbleScan Software produce different but similar CNV panels for an individual sample

Figure 5 also shows that the CNV predictions of Nexus Software and NimbleScan Software tend to converge for each individual sample. However, in the example in Figure 3, we note that for a given sample, Nexus Software and NimbleScan Software report different breakpoints for the same event. These differences are due to the different algorithms and parameter settings used by both programs. NimbleScan Software uses the segMNT algorithm that requires at least two probes in a segment under the default settings. Nexus Software uses the Rank Segmentation algorithm that requires at least 5 probes in a segment under the default settings used for this analysis. Thus, the smallest CNVs called by Nexus Software are necessarily larger than those called by NimbleScan Software in this case. The median probe spacing on the NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D Array is approximately 1100 bp. Consequently the smallest events called by Nexus Software are at least 5 kb while the smallest events called by NimbleScan Software are at least 2 kb in this case. Increasing the minimum number of probes required per segment decreases the number of short CNVs called but increases the reliability of calls. Further differences in calls are due to other parameter discrepancies such as minimum  $\log_2$  ratio differences between adjacent probes and significance settings.



▲ **Figure 5:** Complete linkage hierarchical clustering of CNV profiles.

### CNVs in genes involved in arsenic metabolism

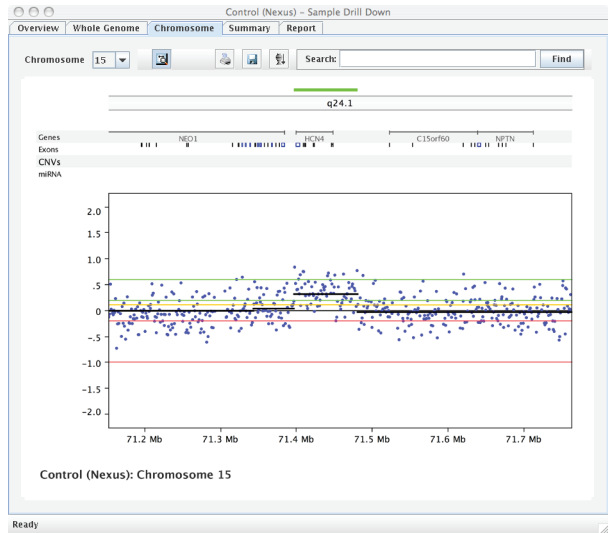
All cases and controls were exposed to arsenic at high levels in drinking water. Elevated inorganic arsenic exposure in drinking water is associated with basal cell carcinoma<sup>9</sup>. We hypothesize that CNVs affecting arsenic metabolism lead to increased arsenic toxicity and development of BCC in cases versus controls. Using the Nexus Software query tool, we directly assessed the CNV genotype at a targeted list of genes known to be involved in arsenic metabolism. The results are shown in Figure 6. We note a general trend of increased copy number of arsenic metabolism genes in the control sample and decreased copy number in the case samples. Although not statistically significant, these results are promising.

NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D Arrays can detect novel CNVs

In addition to revealing associations with known CNVs, the ultra-high resolution NimbleGen Human CGH Whole-Genome Tiling v2.0D Array enabled identification of novel CNVs including a 53 kb CNV on chromosome 15q24.1 overlapping the HCN4 gene (Figure 7).

Sample	GSTA1	GSTP1	GSTT1	UGT2B10	UGT2B15	UGT2B17	MGMT	MTHFR	MT1A	MT1X
Case 1 (Nexus)					CN Loss	CN Loss				
Case 1 (NS 2.4)					CN Loss	CN Loss				
Case 2 (Nexus)					CN Loss	CN Loss				
Case 2 (NS 2.4)					CN Loss	CN Loss				
Case 3 (Nexus)			CN Loss							
Case 3 (NS 2.4)			CN Loss							
Control (Nexus)	CN Gain	CN Gain					CN Gain	CN Gain	CN Gain	
Control (NS 2.4)	CN Gain	CN Gain					CN Gain			




▲ **Figure 6:** CNV genotypes of genes involved in arsenic metabolism.



▲ **Figure 7:** NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D Array can detect novel CNVs.

## References

1. Redon, R. et al., Global variation in copy number in the human genome. *Nature*. 2006 Nov 23;444(7118):444-54.
2. Korbelt, JO et al., Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007 Oct 19;318(5849):420-6.
3. Pinto et al., Copy-number variation in control population cohorts. *Human Molecular Genetics*. 2007 Oct 15;16 Spec No. 2:R168-73.
4. Hurles, ME et al., The functional impact of structural variation in humans. *Trends in Genetics*. 2008 May;24(5):238-45 Review.
5. Gonzalez, E. et al., The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005 307: 1434-1440.
6. Perry, GH et al., Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*. 2007 Oct;39(10):1256-60.
7. McCarroll, SA et al., Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genetics*. 2008 Sep;40(9):1107-12.
8. Kozul, CD et al., Chronic exposure to arsenic in the drinking water alters the expression of immune response genes in mouse lung. *Environmental Health Perspectives*. 2009 Jul;117(7):1108-15.
9. Yu, HS et al., arsenic carcinogenesis in the skin. *Journal of Biomedical Science*. 2006 Sep;13(5):657-66.

Array Specs	2.1M Array	3x720K Array	12x135K Array
			
Total features	2.1 million	3 x 720,00	12 x 135,000

Ordering Information		
Product	<b>D</b> Delivery Cat. No.	<b>S</b> Service Cat. No.
Human CGH 2.1M Whole-Genome Tiling v2.0D Array	05 541 921 001	05 543 991 001
NimbleGen Human CNV 2.1M v1.0 Array	05 913 152 001	05 913 195 001
Human CGH 3x720K Whole-Genome Tiling v3.0 Array	05 520 797 001	05 520 860 001
NimbleGen Human CNV 3x720K v1.0 Array	05 913 209 001	05 913 233 001
Human CGH 3x720K Whole-Genome Exon-Focused Array	05 542 073 001	05 544 122 001
Human CGH 12x135K Whole-Genome Tiling v3.0 Array	05 520 878 001	05 520 886 001

Microarray Processing Accessories	
Reagents	Cat. No.
NimbleGen Dual-Color DNA Labeling Kit	05 223 547 001
NimbleGen Hybridization Kit	05 583 683 001
NimbleGen Hybridization Kit, LS	05 583 934 001
NimbleGen Wash Buffer Kit	05 584 507 001
NimbleGen Array Processing Accessories	05 223 539 001
NimbleGen Sample Tracking Control Kit	05 223 512 001
Equipment	Cat. No.
NimbleGen Hybridization System 4 (110V)	05 223 652 001
NimbleGen Hybridization System 12 (110V)	05 223 679 001
NimbleGen Hybridization System 4 (220V)	05 223 687 001
NimbleGen Hybridization System 12 (220V)	05 223 695 001
NimbleGen Microarray Dryer (110V)	05 223 636 001
NimbleGen Microarray Dryer (220V)	05 223 644 001
NimbleGen MS 200 Microarray Scanner	05 394 341 001
Software	Cat. No.
NimbleScan Software – Individual License	05 225 035 001
NimbleScan Software – Site License	05 225 043 001

Roche Microarray Technical Support:  
[www.nimblegen.com/arraysupport](http://www.nimblegen.com/arraysupport)



## **Comparative Genomic Hybridization (CGH) & Copy Number Variation (CNV)**

### ***Application Note***

For life science research only. Not for use in diagnostic procedures.

NIMBLEGEN is a trademark of Roche.

Other brands or product names are trademarks of their respective holders.

*Published by:*  
Roche Diagnostics GmbH  
Roche Applied Science  
Werk Penzberg  
82372 Penzberg  
Germany

[www.roche-applied-science.com](http://www.roche-applied-science.com)

© 2010 Roche Diagnostics GmbH  
All rights reserved.