NEOPLASIA
www.neoplasia.com

Volume 13 Number 11 November 2011 pp. 1019–1025 1019

Detection of Somatic Copy Number Alterations in Cancer Using Targeted Exome Capture Sequencing[1,2]

Robert J. Lonigro*,†, Catherine S. Grasso*,‡, Dan R. Robinson*,‡, Xiaojun Jing*,‡, Yi-Mi Wu*,‡, Xuhong Cao*,‡,§, Michael J. Quist*,‡, Scott A. Tomlins*,‡, Kenneth J. Pienta*,†,¶,# and Arul M. Chinnaiyan*,†,‡,§,#,3

*Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, MI, USA; †Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, USA; ‡Department of Pathology, University of Michigan Medical School, Ann Arbor, MI, USA; §Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI, USA; ¶Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA; #Department of Urology, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract

The research community at large is expending considerable resources to sequence the coding region of the genomes of tumors and other human diseases using targeted exome capture (i.e., "whole exome sequencing"). The primary goal of targeted exome sequencing is to identify nonsynonymous mutations that potentially have functional consequences. Here, we demonstrate that whole-exome sequencing data can also be analyzed for comprehensively monitoring somatic copy number alterations (CNAs) by benchmarking the technique against conventional array CGH. A series of 17 matched tumor and normal tissues from patients with metastatic castrate-resistant prostate cancer was used for this assessment. We show that targeted exome sequencing reliably identifies CNAs that are common in advanced prostate cancer, such as androgen receptor (*AR*) gain and *PTEN* loss. Taken together, these data suggest that targeted exome sequencing data can be effectively leveraged for the detection of somatic CNAs in cancer.

Neoplasia (2011) 13, 1019–1025

Introduction

Recognition that copy number alterations (CNAs) in tumor genomes, which can result in the amplification of oncogenes or the deletion of tumor suppressors, contribute significantly to cancer etiology has led to the development of multiple techniques for their comprehensive identification. Initial global approaches for CNA detection relied primarily on array based technologies: whole-genome array comparative genomic hybridization (aCGH) tests the relative frequency of probe

Address all correspondence to: Arul M. Chinnaiyan, MD, PhD, Comprehensive Cancer Center, University of Michigan Medical School, 1400 E Medical Center Dr, 5316 CCGC, Ann Arbor, MI 48109-0602. E-mail: arul@umich.edu

[1]This work was supported in part by the National Institutes of Health (NIH) Specialized Program of Research Excellence (P50 CA69568) and the Early Detection Research Network (U01 CA111275). A.M.C. is supported by the Howard Hughes Medical Institute, the Prostate Cancer Foundation, the Taubman Research Institute, the Doris Duke Foundation, and the American Cancer Society as a clinical research professor. K.J.P. is supported by the Prostate Cancer Foundation, the Taubman Research Institute, and the American Cancer Society as a clinical research professor (NIH 1 PO1 CA093900 and 1 U01CA143055).
[2]This article refers to supplementary materials, which are designated by Figures W1 to W3 and Tables W1 to W4 and are available online at www.neoplasia.com.
[3]Dr Chinnaiyan is an investigator for Howard Hughes Medical Institute, an American Cancer Society professor, an S. P. Hicks Endowed Professor of Pathology, and a professor of pathology and urology.
Received 2 September 2011; Revised 2 September 2011; Accepted 28 September 2011

DOI 10.1593/neo.111252

DNA segments between two genomes [1–4], whereas single-nucleotide polymorphism (SNP) arrays measure the probe intensities at known SNP loci to identify shifts in zygosity relative to another genome [5–9]. The recent advent of high-throughput sequencing has made the sequencing of whole human genomes feasible and has made possible the development of sequencing-based approaches to CNA identification [10–17].

The prohibitive cost and time constraints of whole-genome sequencing has necessitated further innovation, and hybridization-based approaches to high-throughput sequencing that focus on the human exome have been recently applied to detect novel somatic point mutations in tumor genomes [18–22]. Targeted exome sequencing allows one to achieve very high depths of coverage (100× coverage or greater) of regions of interest and thus provides advantages over whole-genome sequencing for mutation detection especially in the context of the highly deranged genomes of many tumors. Because targeted exome sequencing yields depth of coverage data, it is reasonable to ask whether exome sequencing data can also be used to detect CNAs, especially because a recent application to unmatched cancer cell lines indicated the potential of this approach [23]. In addition, the recent development of third-generation sequencing approaches has made sequencing of a tumor exome achievable within a week, making its application to detect somatic mutations in a clinical setting imminent. As a result of its wide applicability, there would be a clear and demonstrable advantage to applying exome sequencing to generate CNA data because it would obviate the need for performing aCGH or whole-genome sequencing to detect CNAs in patients awaiting treatment.

Here we propose a method for the detection of somatic CNAs from exome sequencing of a matched tumor/normal pair. By comparing depth of coverage across the exome between the tumor and normal samples, we detect regions with predicted copy gain or loss in the tumor sample. A comparison of these data to aCGH copy number data for the same samples demonstrates a high level of agreement between the two platforms. We apply our method to identify copy number aberrations from exome data generated for 17 prostate tumor-normal pairs, showing that our method identifies aberrations in multiple genes known to have copy number gains and losses in prostate cancer including *AR*, *NCOA2*, *PTEN*, *RB1*, and *TP53* [24]. Together, these analyses show that exome sequencing data, in addition to being useful for detecting point mutations and indels, can be used in place of aCGH and whole-genome sequencing for the generation of CNA data.

Materials and Methods

Tissue Samples

Prostate tissues were from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program [25], both of which are part of the University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. All samples were collected with informed consent of the patients and previous institutional review board approval.

High–Molecular Weight Genomic DNA Isolation

Frozen tissue samples were taken as chunks or sections from OCT-embedded, flash-frozen tissue blocks. Genomic DNA (gDNA) was isolated using the Qiagen DNeasy Blood and Tissue Kit (Valencia, CA) according to the manufacturer's instructions. Briefly, cell or tissue lysates were incubated at 65°C in the presence of proteinase K and SDS, purified on silica membrane-based mini columns, and eluted in buffer AE (10 mM Tris-HCl and 0.5 mM EDTA pH 9.0).

Exome Capture Sequencing

Exome libraries of matched pairs of tumor/normal gDNAs were generated using the Agilent SureSelect Human All Exon Kit (Agilent, Santa Clara, CA; the 38-Mb kit, including 165,637 exon targets, was used on three tumor/normal matched pairs and the 50-Mb kit, including 213,050 exon targets, was used on the remaining 14; Table W2) and the Illumina Paired-End Genomic DNA Sample Prep Kit (Illumina, San Diego, CA) following the manufacturers' instructions. Three micrograms of each gDNA was sheared using a Covaris (Woburn, MA) S2 to a target peak size of 250 bp. Fragmented DNA was concentrated using AMPure XP beads (Beckman Coulter, Indianapolis, IN), and DNA ends were repaired using T4 DNA polymerase, Klenow polymerase, and T4 polynucleotide kinase. 3′ A-tailing with exo-minus Klenow polymerase was followed by ligation of Illumina paired-end adapters to the gDNA fragments. The adapter-ligated libraries were electrophoresed on 3% Nusieve 3:1 (Lonza, Walkersville, MD) agarose gels, and fragments between 250 and 350 bp were recovered using QIAEX II gel extraction reagents (Qiagen). Recovered DNA was then amplified using Illumina PE1.0 and PE2.0 primers for nine cycles. The amplified libraries were purified using AMPure XP beads, and the DNA concentration was determined using a Nanodrop spectrophotometer (NanoDrop 8000; Thermo Scientific, Wilmington, DE). Five hundred nanograms of the libraries was hybridized to the Agilent biotinylated SureSelect Capture Library at 65°C for 65 hours. The targeted exon fragments were captured on Dynal M-280 streptavidin beads (Invitrogen, Carlsbad, CA), washed, eluted, and enriched by amplification with the Illumina PE1.0/PE2.0 primers for eight additional cycles. After purification of the polymerase chain reaction products with AMPure XP beads, the quality and quantity of the resulting exome libraries were analyzed using an Agilent Bioanalyzer (Agilent). All captured DNA libraries were sequenced with the Illumina HiSeq 2000 (Illumina) in paired-end mode trimmed to yield 78-bp reads. The reads that passed the chastity filter of Illumina BaseCall software were used for subsequent analysis. Next, mate pairs were pooled and then mapped as single reads to the reference human genome (NCBI build 36.1, hg18), excluding unordered sequence and alternate haplotypes, using Bowtie [26], keeping unique best hits, and allowing up to two mismatched bases per read.

Array Comparative Genomic Hybridization

aCGH of six samples (matched tumor and normal) from three metastatic prostate cancer patients was performed using gDNA on Agilent's 244K aCGH microarrays (Human Genome CGH 244K Oligo Microarray) using Agilent's Standard Direct Method protocol and Wash Procedure B. Briefly, 1.5 to 3 μg of gDNA from prostate specimens (isolated as above) was restriction digested with *Alu*I and *Rsa*I, labeled with Cy-5 (test channel), purified using Microcon YM-30 columns (Millipore, Hayward, CA), and hybridized with an equal amount of Cy-3 (reference channel)-labeled human male gDNA (Promega, Madison, WI) for 40 hours at 65°C. Posthybridization wash was performed with acetonitrile wash and Agilent Stabilization and Drying Solution wash. Scanning was performed on an Agilent scanner (Model G2505B; 5-μm scan with software v7.0), and data were extracted using Agilent Feature Extraction software v9.5 using protocol CGH-v4_95_Feb07.

For data analysis, quantifications for each probe were determined as rProcessedSignal/gProcessedSignal and analyzed on the log2 scale. To

focus on somatic copy number changes, log2 ratios in the matched normal sample were subtracted from the log2 ratios in each tumor sample, and the resulting differences were used for analysis. Replicate probes on the array were summarized by computing the median value across replicates for each sample and using this median for analysis. The resulting log2 ratios were median centered for each tumor/normal matched pair.

Segmentation Analysis

Segmentation analysis for both aCGH and exome log2 copy number ratios was performed through the use of the Circular Binary Segmentation Algorithm [27], as implemented in the DNAcopy package in R version 2.11.1. Default values for all parameters were used, except that consecutive segments were merged using the undo.splits = "sdundo" option with the undo.SD parameter set to 0.3/DLRS, where DLRS (derivative log ratio spread) represents the local SD in log ratio units, a well-known measure of local variability for aCGH microarrays. In this way, the segmentation algorithm was tuned to detect copy number changes of at least 0.3 in magnitude on the log2 scale. Segments were reported as amplified or deleted if the corresponding estimated copy number ratio was greater than 1.25 or less than 0.75, respectively; high-level amplifications and homozygous losses were called whenever the estimated copy number ratio was greater than 2.0 or less than 0.5. ROC analysis was performed using the ROC package in Bioconductor.

Informative Genes

The list of 2016 "informative" cancer genes used to quantify performance of copy number calls was generated by combining the list of 457 genes comprising the Sanger Institute's Cancer Gene Census together with a list of 1933 Protein Kinases, Tumor Suppressors, Tyrosine Kinases, and Oncogenes downloaded from Memorial Sloan-Kettering Cancer Center's Cancer Genes resource. This resulted in a list of 2217 unique genes, of which 2016 were targeted by the Agilent SureSelect Human All Exon Kit and did not map to the Y chromosome.

Results

Algorithm: Detecting Copy Number Alterations from Whole Exome Sequencing Data

Figure 1 illustrates our approach to generating copy number data from whole exome sequencing data. Libraries are prepared from tumor and matched normal DNA, targeted exons are sequenced, and per-exon coverage is computed for both tumor and normal samples. In contrast to genomic sequencing, in which depth of coverage is approximately proportional to genomic copy number [11], exome sequencing involves varying capture efficiencies across the human exome, making the relationship between coverage and copy number less apparent. We accounted for variation in coverage across exons by performing per-exon comparisons between each tumor sample and its matched normal sample. This approach also corrects for variation in observed coverage across different regions of the human genome due to the presence of repetitive sequences and variation in GC content.

More precisely, we generated copy number ratios for each tumor/normal matched pair through the following algorithm. First, exons containing fewer than 10 reads (i.e., 780 bp) worth of coverage in the matched normal sample were excluded from analysis. Second, coverage values were perturbed slightly by adding 780 bp of coverage to each exon's coverage quantification in both samples. Third, per-exon coverage in the tumor sample was divided by per-exon coverage in the matched normal sample, resulting in coverage ratios for each exon. These modified coverage ratios were then globally normalized by dividing each of them by the ratio of human mappings between the two samples (tumor/normal). After log2-transforming these normalized coverage ratios, the overall median value was subtracted, resulting in a set of log2-transformed coverage ratios with median zero for each tumor/normal matched pair. Ratios and logarithms were well defined throughout this process owing to the filtering of low-coverage exons in the benign sample, which ensured that division by zero did not occur, and the small perturbation of coverage values, which ensured that coverage ratios were always nonzero. The normalized log2-transformed coverage ratios were used for downstream segmentation analysis. The resulting data structure is analogous to that of a two-channel microarray with "probes" at each targeted exon and sequencing coverage replacing signal intensity.

Benchmarking: Concordance with Copy Number Alterations Generated Using aCGH

We investigated the ability of these normalized exome capture coverage ratios to yield accurate copy number assessments by comparing them against the corresponding copy number ratios from Agilent 244K CGH microarray data. We used three matched metastatic castration-resistant prostate tumor/normal pairs for this comparison, using the 38-Mb

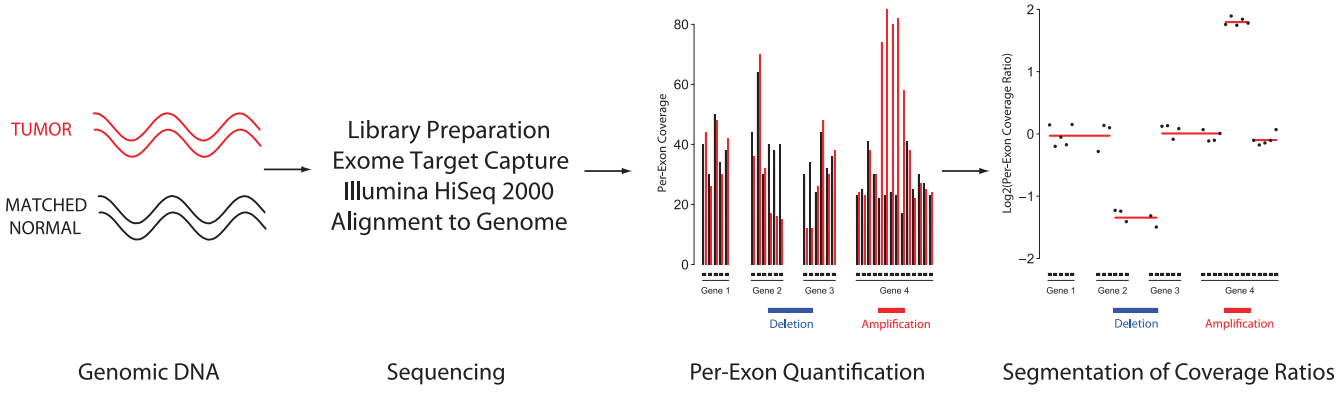

Figure 1. Overview of copy number analysis by whole exome sequencing. Vertical bars represent per-exon coverage in the tumor (red) and matched normal (black) tissue. Log-transformed coverage ratios between tumor and normal tissues are computed for each exon (black dots) and altered regions are identified through segmentation analysis (red line segments).

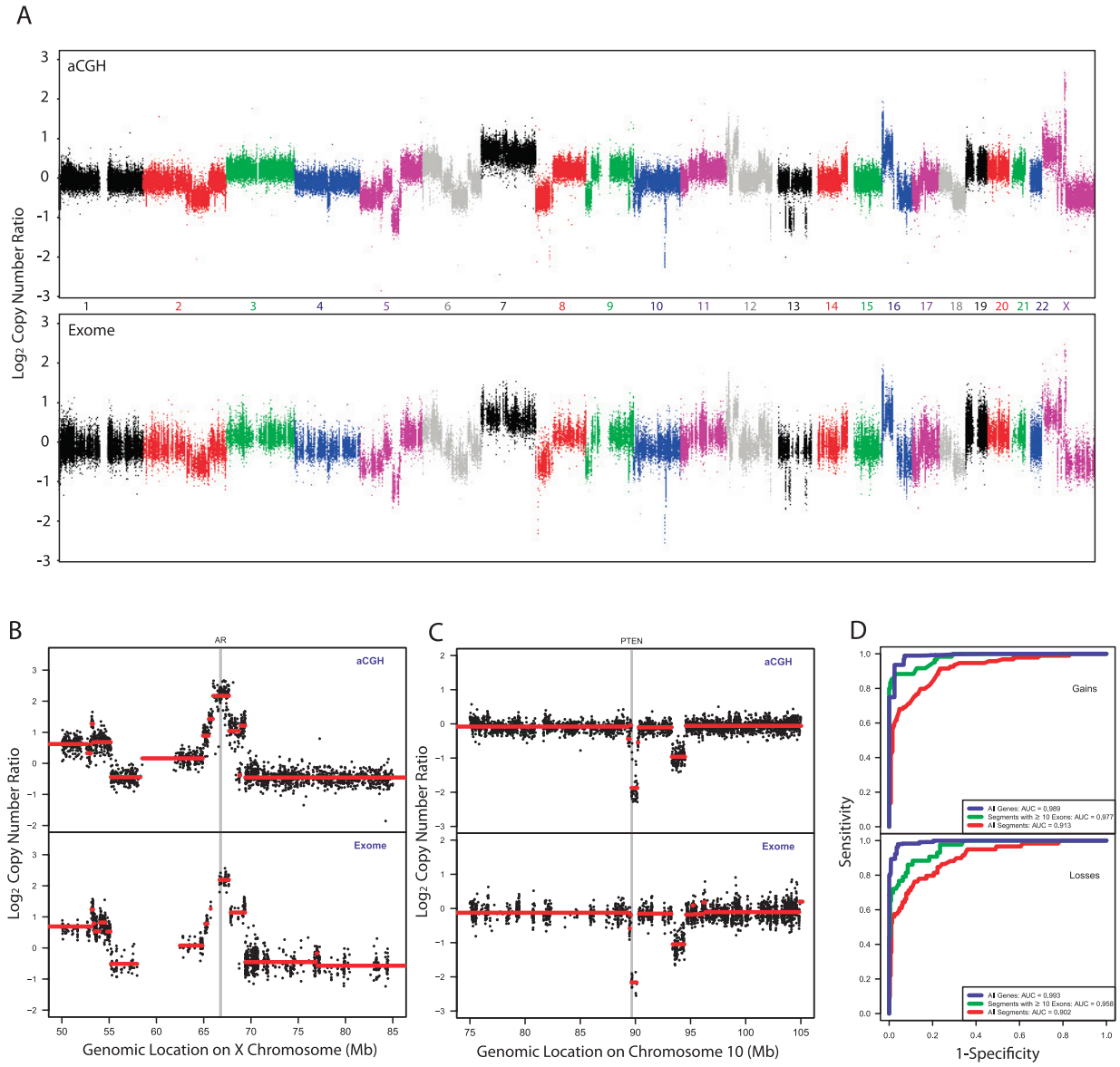

Figure 2. Comparison of exome sequencing to array CGH in detecting CNAs. (A) Overall copy number across the genome for metastatic prostate tumor sample WA54 by aCGH (upper panel) and exome sequencing (lower panel). Log2(copy number ratio) between tumor and matched normal is shown on the vertical axis; each point represents the log-transformed ratio for each aCGH probe or targeted exon, ordered by genomic coordinates. (B) Copy number assessment for sample WA54 by aCGH and exome sequencing in a 35-Mb region containing the *AR* gene. Red line segments represent segmented copy number data. (C) Copy number assessment for sample WA54 by aCGH and exome sequencing in a 30-Mb region containing the *PTEN* gene. Red line segments represent segmented copy number data. (D) Classification performance of exome capture sequencing relative to aCGH for sample WA54. ROC curves are shown, using aCGH copy number assessments as a criterion standard. ROC curves are presented for classifying all aCGH segments (red), segments containing at least 10 targeted exons (green), and all targeted genes (blue).

Agilent SureSelect Human All Exon Kit to target the human exome and performing next-generation sequencing on the Illumina HiSeq 2000 platform. Representative assessments for one sample WA54 are shown in Figure 2 with assessments for the other two samples in Figures W1 and W2. Genome-wide copy number ratios are highly concordant between the two technologies (Figure 2*A*) with large-scale amplifications agreeing in magnitude. Large regions of gain and loss spanning whole chromosomes and chromosomal arms, such as chromosome 7 gain, 8p loss, 8q gain, 16p gain, 16q loss, 18q loss, and Xp gain, are easily visible by both technologies. We verified this concordance more formally by comparing copy number ratios on disjoint windows covering the genome. Specifically, we partitioned the genome into windows containing at least five targeted exons and five aCGH probes and computed mean log ratios by each technology on each window. The resulting quantifications exhibit strong correlations for each of the three samples (minimum Pearson correlation coefficient = 0.92, $P < .001$; Figure W3). This genome-wide comparison illustrates that copy number ratios from exome capture sequencing exhibit strong concordance with and are on the same scale as those from aCGH microarrays.

In addition, we examined copy number at genes well known to be gained or lost in prostate cancer and found that copy number assessments for these genes were concordant as well. Both technologies

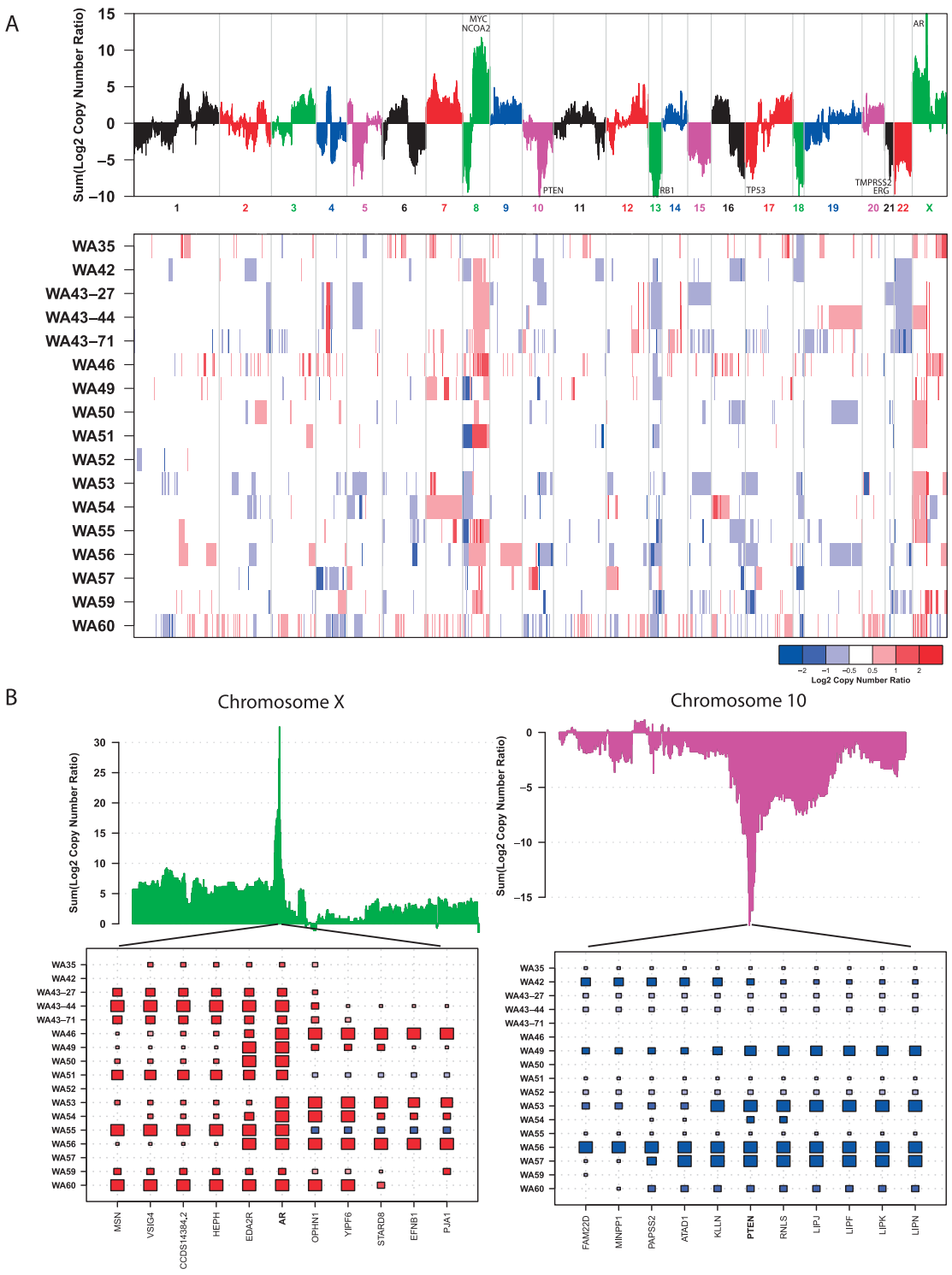

Figure 3. Qualitative comparison to prior CNAs observed in prostate cancer. (A) Overall summary of copy number across 17 lethal metastatic castration-resistant prostate cancers. Summed segmented log2 copy number ratios (top panel) for all targeted genes across the 17 samples are shown. Genes exhibiting recurrent amplifications or deletions across the cohort will have large positive or negative values, respectively. Regions of copy number gain and loss for all 17 samples are shown in a heat map (bottom panel). Red represents amplification; white, copy number neutral; blue, deletion. Three samples are derived from three different metastatic foci from a man with lethal castrate-resistant prostate cancer: celiac lymph node metastatic site (WA43-27), lung metastatic site (WA43-71), and bladder metastatic site (WA43-44). (B) Focal amplifications of the *AR* gene and deletions of the *PTEN* gene in this cohort. *AR* has the largest positive summed log copy number ratio across the 17 samples, with a total sum of 32.6, whereas *PTEN* has the largest negative summed log copy number ratio, with a total sum of −17.5. A plot of this sum over the entire chromosome (top) is shown; a large positive peak is present at *AR* and a large negative peak is present at *PTEN*. Segmented copy number ratios are represented by boxes, with the area (absolute log2 ratio) and color intensity (log2 ratio; copy number gain in red; loss in blue) of each box proportional to mean copy number across that gene. Missing boxes indicate that the gene is neither amplified nor deleted in that sample.

reveal a focal amplification of the *AR* gene in sample WA54 (Figure 2*B*), and the patterns of copy number changes in that region are strikingly similar. The estimated number of copies in the segment overlapping the *AR* gene was similar by each approach (4.50 copies by aCGH, 4.57 copies by exome capture), revealing that exome capture coverage ratios exhibit sufficient dynamic range to capture high-level amplifications. Similarly, both aCGH and exome sequencing reveal a focal region of two-copy loss at the *PTEN* gene in sample WA54 (Figure 2*C*), and the two technologies agree on the approximate number of copies: 0.27 copies of *PTEN* by aCGH and 0.22 copies of *PTEN* by exome capture.

To quantify the ability of exome capture sequencing to identify regions of gain and loss, we performed ROC analysis of exome capture quantifications, using the matched aCGH data as a criterion standard (Figure 2*D*). First, we performed segmentation analysis (Materials and Methods) on both aCGH and exome capture log-transformed copy number ratios, and using copy number ratio cutoffs of 1.25 and 0.75 to define regions of gain and loss, respectively, we identified a set of altered regions by aCGH. The segmented exome capture copy number ratios were computed on these altered segments; performance of this classifier relative to the aCGH calls was quantified using the area under the ROC curve (AUC). For each of the three samples, exome copy number analysis performed well in classifying these aCGH segments, with AUCs of at least 0.89 across the three samples (Table W1). As expected, restricting attention to segments that contain targeted exons improved classification performance; for segments containing at least two targeted exons, the minimum AUC across the three samples was 0.94. Finally, we did a gene-centric analysis, comparing copy number calls for each of the 18,090 genes targeted by the exome capture kit, and performance was even better, with a minimum AUC of 0.989 across the three samples. We repeated this analysis for a smaller set of 2016 "informative" genes (Materials and Methods) that have been implicated in cancer (e.g., kinases, oncogenes, and tumor suppressors) and verified that the strong performance persists when restricted to genes that are likely to be relevant to cancer progression.

Application: Copy Number Alterations in Prostate Cancer Tissues

Next, we sought to demonstrate that our CNA detection method can be used to generate results qualitatively equivalent to aCGH-based methods by applying it to the exomes of 17 lethal metastatic castration-resistant prostate cancers (Table W2), including the three samples used for benchmarking, and matched normal tissues from the same patients. In total we generated 395,489,506,152 bases, with 105.27 average fold coverage of each targeted base per tissue sample (Table W3). Using copy number ratio cutoffs of 1.25 to define regions of gain and 0.75 to define regions of loss, we identified a median of 93 gained regions and 79 lost regions across the 17 samples (Table W4). The median total length of these altered regions across samples was 407.5 Mb (gains) and 406.2 Mb (losses). Using more stringent copy number ratio cutoffs of 2.0 to define high-level gain and 0.5 to define homozygous loss identified a median of 19 high-level gains and 17 homozygous losses covering 23.3 and 25 Mb, respectively. Three of the cancer samples were derived from different metastatic sites from the same patient; these three samples had multiple amplifications and deletions in common, including focal amplifications on chromosomes 4 and 14, the broad 8q amplification, loss of chromosome 22, and a focal loss on the end of 2q, reflecting the likely clonal origin of the tumor (Figure 3*A*).

Global analysis of copy number profiles of all 17 prostate cancers (Figure 3*A*) identified recurrent aberrations previously associated with prostate cancer development and progression, including broad losses of 1p, 8p, 6q, and losses of large regions of chromosome 13, 15, 18 and 22, as well as gains of 1q, 3q, 7q, and 8q, containing two prostate cancer oncogenes, *MYC* and *NCOA2* [24,28–30]. We also identified previously reported deletions between *TMPRSS2* and *ERG* in cases with *TMPRSS2:ERG* gene fusions [24,28–30]. In addition, we identified recurrent focal amplifications of *AR* (Figure 3*B*) and recurrent homozygous focal deletions of *PTEN* (Figure 3*C*), consistent with prior observations [24,28–30]. Examination of each sample's copy number ratio in these regions shows that the pattern of nearby amplifications around *AR* can be different for each sample; however, the region of gain always includes *AR* itself. The same is true with respect to the region of loss including *PTEN*. We also detected specific disruptions of *RB1* and *TP53* (Figure 3*A*), two genes previously associated with focal losses in prostate cancer [24,28,30].

Discussion

There are a number of benefits of using targeted exome sequencing for assessing CNAs. One benefit of the exome-based approach to identifying CNAs, over using evenly spaced genomic hybridization probes in aCGH, is the possibility of using the data to gain exon-level resolution of the genomic rearrangements associated with copy number changes. Another benefit is that one has the potential to leverage a vast amount of data that is already being generated as part of large genome sequencing projects, such as The Cancer Genome Atlas (TCGA). For example, at the time of writing, the TCGA project had just released 316 whole exomes from ovarian cancer [22], more than 500 whole exomes had just been published, and sequencing for thousands more was underway. In essence, point mutations (such as *BRAF* V600E) and amplifications/deletions (such as amplification of *ERBB2* or loss of *PTEN*) can be monitored from the same whole exome sequencing data set. This type of assessment will be powerful for integrative mutational studies in the context of cancer and toward personalized medicine.

Sequencing-based approaches to copy number detection have the advantage of being able to not only assess CNAs using depth of coverage, like aCGH, but also using SNPs or somatic point mutation to assess for shifts in zygosity indicative of copy number changes, as in SNP array approaches. In this study, we have presented a depth of coverage approach to detect CNAs using exome data, but an approach using SNPs or somatic point mutation is equally feasible in theory. Moreover, a combined approach using both depth of coverage and SNPs has the potential to be even more effective, especially compared with both aCGH and SNP arrays.

As intimated, the exome approach is limited by mapping issues, making genes containing highly repetitive sequence difficult to target for exome sequencing and therefore difficult to assess for CNAs using this method. For example, the second exon of *FOXA1*, a two-exon gene, has two large gaps in coverage in both the 38- and the 50-Mb Agilent SureSelect All Human Exon design, resulting from repetitive sequence, so that the computed coverage of the exon is always a gross underestimate of the actual coverage. These sorts of coverage issues make detection of focal CNAs of certain genes difficult. A second limitation is that exome capture copy number analysis will clearly fail to detect genomic aberrations that occur in regions containing no nearby genes. If exome capture was deliberately used for assessing CNAs, both limitations could be overcome by optimizing the exome capture design to improve detection of these alterations using the excess

sequencing capacity afforded by deep sequencing of only the coding regions of the genome (which represent <1% of the genome). The results of this analysis suggest that additional effort should also be put toward designing exon capture platforms that add additional targets to improve detection of CNAs. This could be done by placing additional targets throughout the genome and near genes with highly repetitive regions, even if they are not directly sequencing a region of interest. Personalized medicine approaches that emphasize somatic mutations in informative coding genes would clearly benefit from an exon capture platform and could efficiently assess genes of interest for both somatic point mutations and for somatic CNAs.

Acknowledgments

The authors thank Javed Siddiqui and Rohit Mehra for assisting with sample acquisition, Terrence Barrette for assisting with sequence data generation using the Illumina pipeline, and Jyoti Athanikar for assistance with manuscript preparation.

References

[1] Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207–211.

[2] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528.

[3] Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Seagraves R, et al. (2005). Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78–88.

[4] Carter NP (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16–S21.

[5] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, deBakker PI, Maller JB, Kirby A, et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166–1174.

[6] Cooper GM, Zerr T, Kidd JM, Eichler EE, and Nickerson DA (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* **40**, 1199–1203.

[7] Conrad DF, Andrews TD, Carter NP, Hurles ME, and Pritchard JK (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75–81.

[8] Hinds DA, Kloek AP, Jen M, Chen X, and Frazer KA (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38**, 82–85.

[9] McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. (2006). Common deletion polymorphisms in the human genome. *Nat Genet* **38**, 86–92.

[10] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722–729.

[11] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, and Landers ES (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99–103.

[12] Nord AS, Lee M, King MC, and Walsh T (2011). Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics* **12**, 184.

[13] Magi A, Benelli M, Yoon S, Roviello F, and Torricelli F (2011). Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res* **39**, e65.

[14] Kim TM, Luquette LJ, Xi R, and Park PJ (2010). rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* **11**, 432.

[15] Medvedev P, Fiume M, Dzamba M, Smith T, and Brudno M (2010). Detecting copy number variation with mated short reads. *Genome Res* **20**, 1613–1622.

[16] Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, and Barillot E (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**, 268–269.

[17] Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061–1067.

[18] Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, Shi JY, Zhu YM, Tang L, Zhang XW, et al. (2011). Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat Genet* **43**, 309–315.

[19] Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma. *Nature* **469**, 539–542.

[20] Harbour JW, Onken MD, Roberson ED, Duan S, Cao L, Worley LA, Council ML, Matatall KA, Helms C, and Bowcock AM (2010). Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science* **330**, 1410–1413.

[21] Jones S, Wang TL, Shih leM, Mao TL, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA Jr, Vogelstein B, et al. (2010). Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231.

[22] The Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615.

[23] Chang H, Jackson DG, Kayne PS, Ross-Macdonald PB, Byseck R, and Siemers NO (2011). Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS One* **6**, e21097.

[24] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, et al. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22.

[25] Rubin MA, Putzi M, Mucci N, Smith DC, Wojno K, Korenchuk S, and Pienta KJ (2000). Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* **6**, 1038–1045.

[26] Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

[27] Olshen AB, Venkatraman ES, Lucito R, and Wigler M (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.

[28] Liu W, Laitinen S, Khan S, Vihinen M, Kowalski J, Yu G, Chen L, Ewing CM, Eisenberg MA, Carducci MA, et al. (2009). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**, 559–565.

[29] Holcomb IN, Young JM, Coleman IM, Salari K, Grove DI, Hsu L, True LD, Roudier MP, Morrissey CM, Higano CS, et al. (2009). Comparative analyses of chromosome alterations in soft-tissue metastases within and across patients with castration-resistant prostate cancer. *Cancer Res* **69**, 7793–7802.

[30] Demichelis F, Setlur SR, Beroukhim R, Perner S, Korbel JO, Lafargue CJ, Pflueger D, Pina C, Hofer MD, Sboner A, et al. (2009). Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer* **48**, 366–380.

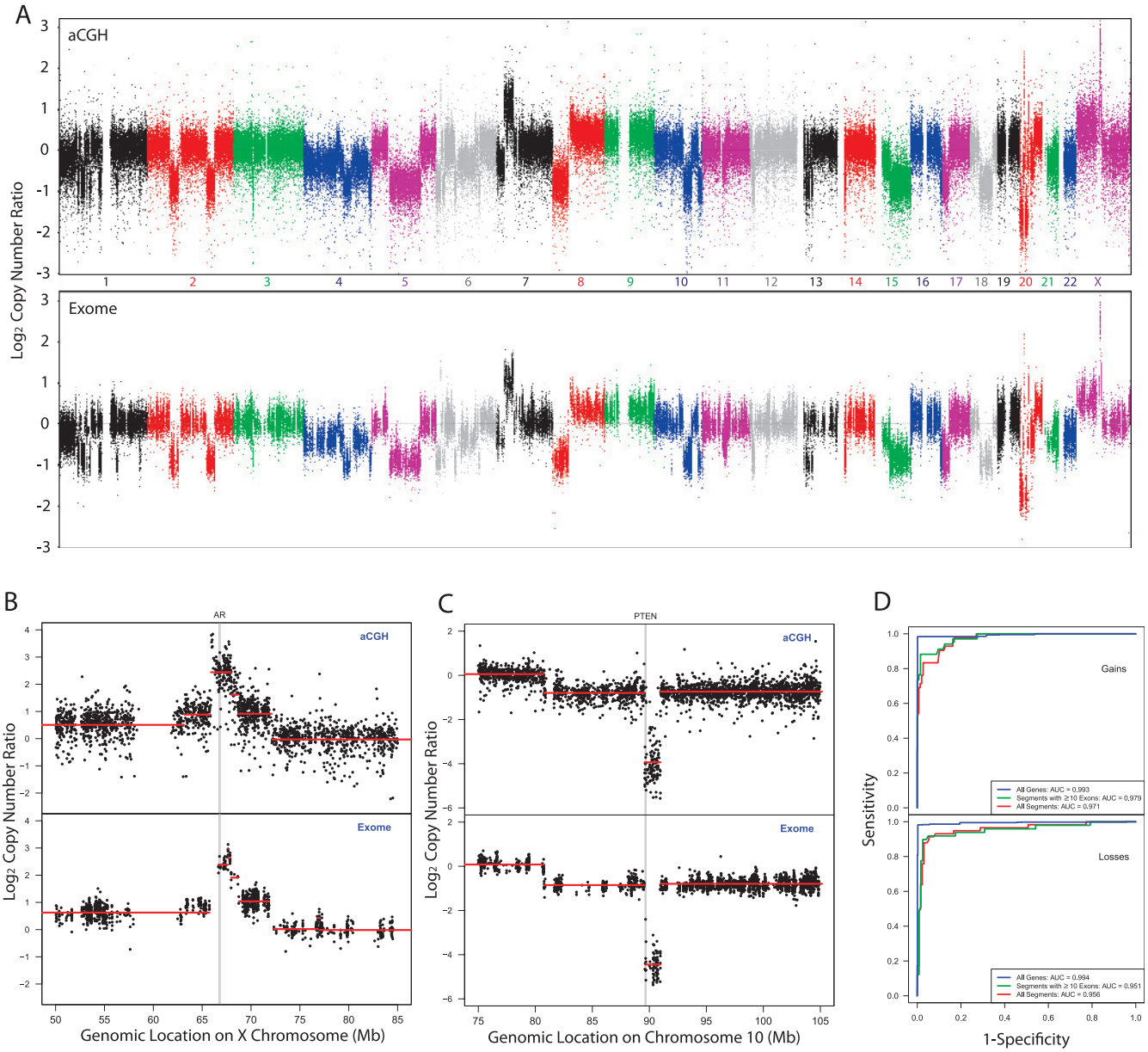

Figure W1. Concordance of aCGH and exome capture copy number assessments in sample WA53. (A) Overall copy number across the genome for sample WA53 by aCGH and exome sequencing. Log2(copy number ratio) between tumor and matched normal is shown on the vertical axis; each point represents the log-transformed ratio for each aCGH probe or targeted exon, ordered by genomic coordinates. Large-scale amplifications and deletions are visible and agree in magnitude across the two technologies. (B) Copy number for sample WA53 by aCGH and exome sequencing in a 35-Mb region containing the AR gene. Both technologies reveal the same focal pattern of amplification and give similar estimates of the number of copies of the AR gene. Red line segments represent segmented copy number data. (C) Copy number for sample WA53 by aCGH and exome sequencing in a 30-Mb region containing the PTEN gene. Both technologies reveal focal 2-copy loss of PTEN in this sample. Red line segments represent segmented copy number data. (D) Classification performance of exome capture sequencing relative to aCGH for sample WA53. ROC curves are shown, using aCGH copy number assessments as a criterion standard. ROC curves are presented for classifying all aCGH segments (red), segments containing at least ten targeted exons (green), and all targeted genes (blue).

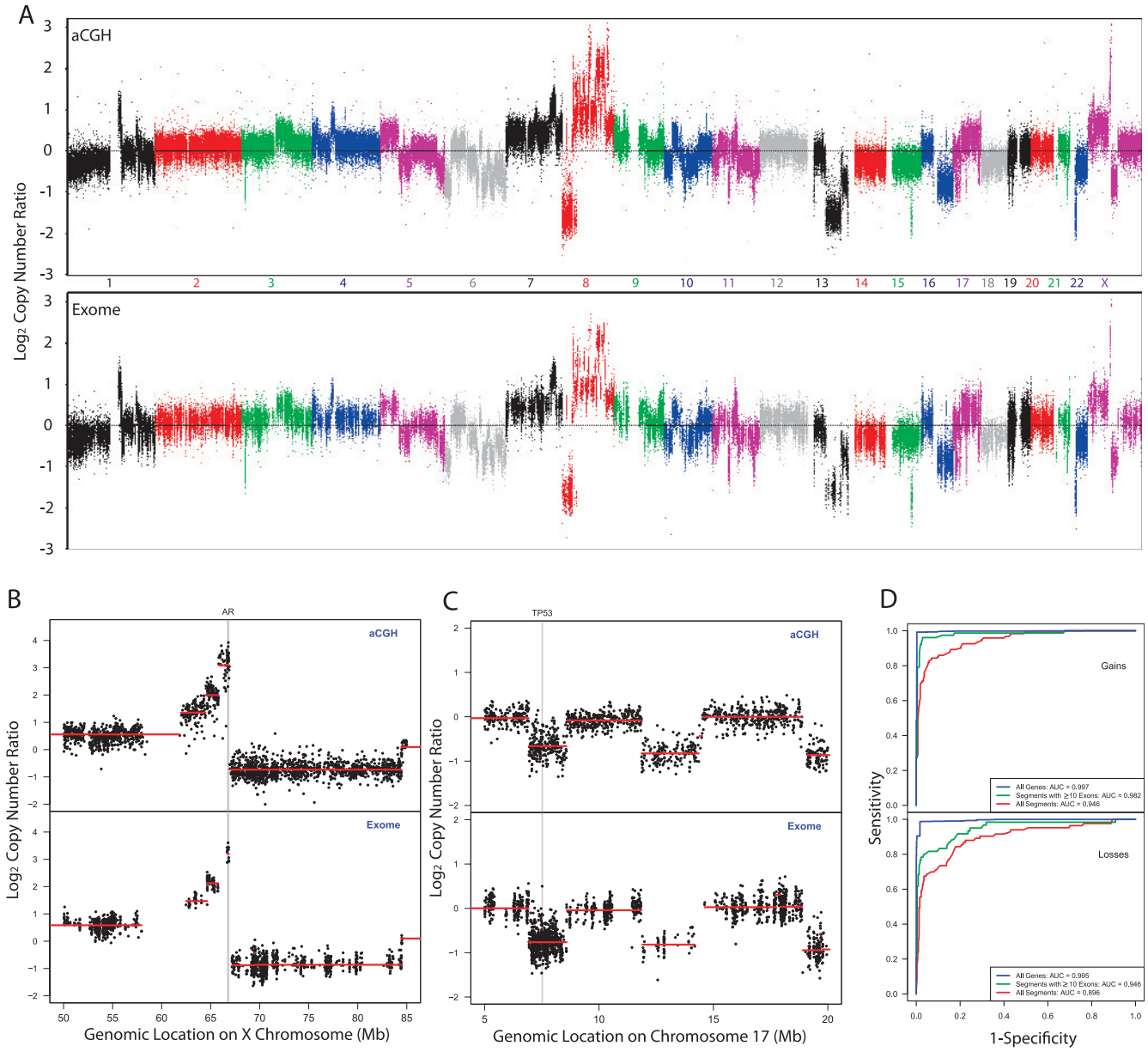

Figure W2. Concordance of aCGH and exome capture copy number assessments in sample WA55. (A) Overall copy number across the genome for sample WA55. Log2(copy number ratio) between tumor and matched normal is shown on the vertical axis; each point represents the log-transformed ratio for each targeted exon or aCGH probe, ordered by genomic coordinates. Large-scale amplifications and deletions are visible and agree in magnitude across the two technologies. (B) Copy number for sample WA55 by aCGH and exome sequencing in a 35-Mb region containing the AR gene. Both technologies reveal the same focal pattern of amplification and give similar estimates of the number of copies of the AR gene. Red line segments represent segmented copy number data. (C) Copy number for sample WA55 by aCGH and exome sequencing in a 15-Mb region containing the TP53 gene. Both technologies reveal focal one-copy loss of TP53 in this sample. Red line segments represent segmented copy number data. (D) Classification performance of exome capture sequencing relative to aCGH for sample WA55. ROC curves are shown, using aCGH copy number assessments as a criterion standard. ROC curves are presented for classifying all aCGH segments (red), segments containing at least ten targeted exons (green), and all targeted genes (blue).

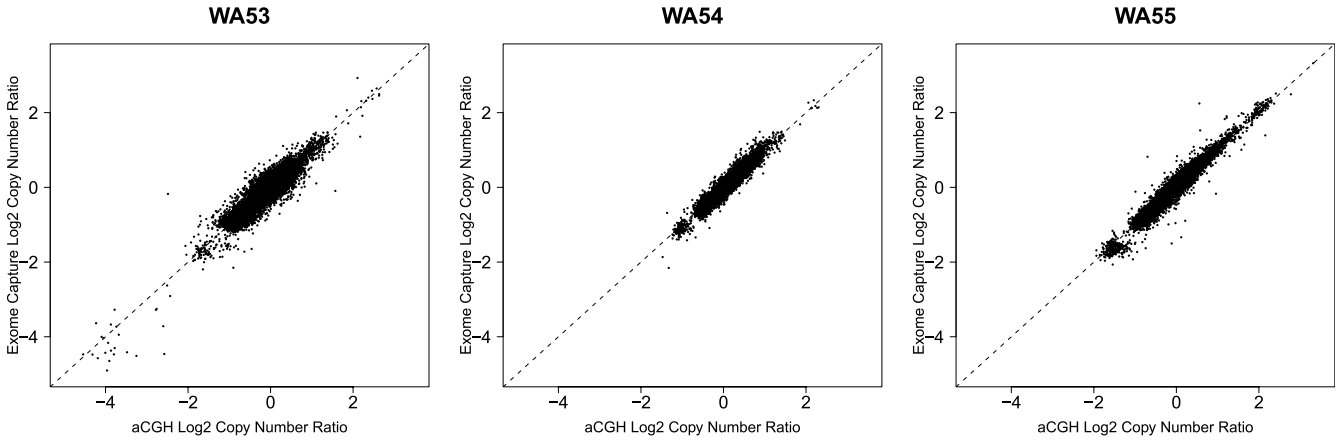

Figure W3. Concordance of copy number by aCGH and targeted exome sequencing through genomic windows analysis. Scatterplots show log2 copy number ratios by each of aCGH and exome sequencing, computed as mean log2 ratios over windows covering the genome. Windows were chosen to include at least five aCGH probes and at least five targeted exons to enable reliable comparison. For each sample, the mean log2 copy number ratios were very highly correlated (WA53; 0.92, WA54; 0.96, WA55; 0.97, $P < .001$ for each sample).

Table W1. AUC Results for Exome Sequencing Compared with aCGH.

	Gain			Loss		
	WA53	WA54	WA55	WA53	WA54	WA55
All segments*	0.971	0.913	0.946	0.956	0.902	0.896
Segments containing ≥2 exons†	0.979	0.945	0.970	0.957	0.956	0.942
Segments containing ≥10 exons†	0.979	0.977	0.982	0.951	0.958	0.946
All genes‡	0.993	0.989	0.997	0.994	0.993	0.995
Informative genes§	0.984	0.991	0.998	0.997	0.991	0.994

*ROC curves for predicting aberration status (gain *vs* no gain or loss *vs* no loss) of each segment identified from segmentation analysis of aCGH data.
†ROC curves for predicting aberration of each segment identified from segmentation analysis of aCGH data, excluding segments overlapping fewer than 2 (or 10) targeted exons.
‡ROC curves for predicting aberration status of each gene targeted by exome sequencing.
§ROC curves for predicting aberration status of each informative gene targeted by exome sequencing.

Table W2. Prostate Cancer Tissue Specimens Used for Exome Sequencing.

Sample Name	Disease State*	Age†	Gleason Score‡	Prior Treatment§	ETS/RAF Status¶
WA53	CRPC	68	NA	H, C, X	ERG+
WA54	CRPC	73	NA	P, R, H, C, X	ERG+
WA55	CRPC	72	NA	H, C, X	ERG+
WA35	CRPC	71	NA	R, H, C, X	Negative
WA42	CRPC	61	NA	H, C	Negative
WA43-27 (celiac LN)	CRPC	52	NA	P, R, H, C	Negative
WA43-44 (bladder)	CRPC				
WA43-71 (right lung)	CRPC				
WA46	CRPC	71	NA	P, R, H, C, X	Negative
WA49	CRPC	68	NA	P, R, H, C	ERG+
WA50	CRPC	78	NA	P, R, H, C	ERG+
WA51	CRPC	65	NA	P, H, C, X	Negative
WA52	CRPC	80	NA	P, H, C	ERG+
WA56	CRPC	79	NA	P, R, H, C, X	ERG+
WA57	CRPC	73	NA (NE diff)	R, H, C,	ERG+
WA59	CRPC	59	NA	H, C, X	Negative
WA60	CRPC	62	NA	H, C	ERG+

*Localized prostate cancer (PC) or castrate-resistant metastatic PC (CRPC).
†Age at diagnosis (PC) or death (CRPC).
‡Gleason score of profiled prostatectomy specimen for PC. CRPCs with neuroendocrine (NE) differentiation are noted.
§C indicates chemotherapy; H, hormone therapy; P, prostatectomy; R, radiation; X, palliative radiation.
¶Rearrangements in ETS or RAF family genes.

Table W3. Prostate Cancer Exome Sequencing Statistics.

Sample	Status	Agilent SureSelect Human All Exon Kit (Mb)	Bases in Target Region	Reads Sequenced (after Quality Filtering)	Bases Sequenced (after Quality Filtering)	Bases Mapped to Genome	Bases Mapped to Target Region	Mean No. Reads per Targeted Base
WA35	Tumor	50	51,712,500	170,916,320	13,331,472,960	10,558,678,026	5,238,994,222	101.31
	Normal	50	51,712,500	157,827,174	12,310,519,572	9,722,413,350	4,443,761,328	85.93
WA42	Tumor	50	51,712,500	169,227,800	13,199,768,400	11,152,269,258	5,328,029,502	103.03
	Normal	50	51,712,500	160,543,858	12,522,420,924	10,584,966,288	5,238,213,050	101.29
WA43	Tumor 43-27	50	51,712,500	106,207,750	8,284,204,500	7,078,252,896	3,474,444,817	67.19
	Tumor 43-44	50	51,712,500	119,846,711	9,348,043,458	8,062,722,408	4,302,838,713	83.21
	Tumor 43-71	50	51,712,500	115,921,897	9,041,907,966	7,819,368,804	3,702,454,417	71.60
	Normal	50	51,712,500	109,694,911	8,556,203,058	7,301,239,530	3,538,639,763	68.43
WA46	Tumor	50	51,712,500	174,132,908	13,582,366,824	11,159,069,376	5,653,532,988	109.33
	Normal	50	51,712,500	174,753,667	13,630,786,026	11,418,269,070	5,428,312,362	104.97
WA49	Tumor	50	51,712,500	155,083,960	12,096,548,880	10,251,439,692	5,091,480,149	98.46
	Normal	50	51,712,500	150,941,382	11,773,427,796	9,990,123,234	5,142,727,707	199.45
WA50	Tumor	50	51,712,500	151,428,570	11,811,428,460	10,163,517,936	4,825,565,562	93.32
	Normal	50	51,712,500	146,431,894	11,421,687,732	9,712,458,132	4,918,010,589	95.10
WA51	Tumor	50	51,712,500	165,709,454	12,925,337,412	10,739,425,398	4,824,644,281	93.30
	Normal	50	51,712,500	170,866,524	13,327,588,872	11,006,378,916	5,103,100,934	98.68
WA52	Tumor	50	51,712,500	196,334,388	15,314,082,264	12,710,410,284	6,105,297,939	118.06
	Normal	50	51,712,500	182,664,677	14,247,844,806	11,851,474,518	5,394,831,142	104.32
WA53	Tumor	38	37,806,033	170,043,479	13,263,391,362	11,275,509,726	6,813,696,982	180.23
	Normal	38	37,806,033	160,836,761	12,545,267,358	10,592,709,894	6,593,161,035	174.39
WA54	Tumor	38	37,806,033	109,465,569	8,538,314,382	7,274,785,830	4,409,679,103	116.64
	Normal	38	37,806,033	168,886,512	13,173,147,936	11,227,983,078	7,016,349,732	185.59
WA55	Tumor	38	37,806,033	169,683,500	13,235,313,000	11,190,529,662	6,730,794,029	178.03
	Normal	38	37,806,033	168,001,511	13,104,117,858	11,095,714,656	6,872,481,717	181.78
WA56	Tumor	50	51,712,500	171,138,470	13,348,800,660	10,979,986,056	5,177,318,788	100.12
	Normal	50	51,712,500	173,359,773	13,522,062,294	11,245,543,686	5,497,614,730	106.31
WA57	Tumor	50	51,712,500	172,761,810	13,475,421,180	10,834,401,240	4,778,197,473	92.40
	Normal	50	51,712,500	169,816,928	13,245,720,384	10,482,745,260	5,096,773,379	98.56
WA59	Tumor	50	51,712,500	159,528,926	12,443,256,228	10,058,917,362	4,475,487,277	86.55
	Normal	50	51,712,500	167,669,729	13,078,238,862	10,418,808,036	4,900,312,121	94.76
WA60	Tumor	50	51,712,500	166,908,169	13,018,837,182	10,453,483,014	4,569,838,785	88.37
	Normal	50	51,712,500	163,743,302	12,771,977,556	10,272,887,664	4,726,398,012	91.40
Mean across samples			49,105,037	158,449,321	12,359,047,067	10,271,452,571	5,169,155,707	105.27

Table W4. Summary of Aberrations across 17 Lethal Metastatic Prostate Samples.

Sample	All Gains (Copy Number Ratio $\geq$ 1.25)		All Losses (Copy Number Ratio $\leq$ 0.75)		High-level Gains (Copy Number Ratio $\geq$ 2.0)		Homozygous Losses (Copy Number Ratio $\leq$ 0.50)	
	No. Aberrations*	Length of Altered Genome (Mb)†	No. Aberrations*	Length of Altered Genome (Mb)†	No. Aberrations*	Length of Altered Genome (Mb)†	No. Aberrations*	Length of Altered Genome (Mb)†
WA35	233	407.5	70	280.2	35	26.5	9	1.2
WA42	12	286.0	81	428.1	0	0.0	19	19.7
WA43-27	53	334.6	56	538.0	18	23.1	9	0.7
WA43-44	51	368.5	61	492.3	23	30.1	9	0.7
WA43-71	490	849.5	609	379.0	25	23.3	89	25.9
WA46	596	720.5	79	71.1	112	108.2	7	0.5
WA49	173	404.7	43	264.8	29	26.1	9	37.5
WA50	88	429.2	96	183.1	3	1.2	17	3.4
WA51	30	273.4	81	538.3	3	101.5	22	56.4
WA52	33	240.8	55	79.8	0	0.0	3	3.5
WA53	93	332.9	117	513.3	18	22.5	17	18.8
WA54	100	353.1	65	475.7	19	7.7	9	30.7
WA55	104	506.3	83	282.6	29	58.7	18	75.5
WA56	51	421.9	71	406.2	2	3.5	18	70.2
WA57	57	421.3	78	639.6	11	26.2	34	144.6
WA59	306	562.5	282	277.2	37	31.7	28	28.0
WA60	743	433.6	746	912.0	19	21.4	59	67.9
Median	93	407.5	79	406.2	19	23.3	17	25.9

*Number of aberrations refers to the number of segments identified from segmentation analysis exceeding the given threshold.
†Length of altered genome refers to the combined length of these aberrant segments per sample, expressed in megabases.