Published online 27 August 2010

Nucleic Acids Research, 2010, Vol. 38, No. 20 ***e190***
doi:10.1093/nar/gkq730

Reference-unbiased copy number variant analysis using CGH microarrays

Young Seok Ju[1,2,3], Dongwan Hong[1], Sheehyun Kim[1,3], Sung-Soo Park[4], Sujung Kim[4], Seungbok Lee[1,5], Hansoo Park[1,6], Jong-Il Kim[1,2,4,*] and Jeong-Sun Seo[1,2,3,4,5,*]

[1]Genomic Medicine Institute, Medical Research Center, Seoul National University, [2]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, [3]Macrogen Inc., Seoul 153-801, [4]Psoma Therapeutics Inc., Seoul, 153-801, [5]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, Korea and [6]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Received May 7, 2010; Revised July 7, 2010; Accepted August 3, 2010

ABSTRACT

Comparative genomic hybridization (CGH) microarrays have been used to determine copy number variations (CNVs) and their effects on complex diseases. Detection of absolute CNVs independent of genomic variants of an arbitrary reference sample has been a critical issue in CGH array experiments. Whole genome analysis using massively parallel sequencing with multiple ultra-high resolution CGH arrays provides an opportunity to catalog highly accurate genomic variants of the reference DNA (NA10851). Using information on variants, we developed a new method, the CGH array reference-free algorithm (CARA), which can determine reference-unbiased absolute CNVs from any CGH array platform. The algorithm enables the removal and rescue of false positive and false negative CNVs, respectively, which appear due to the effects of genomic variants of the reference sample in raw CGH array experiments. We found that the CARA remarkably enhanced the accuracy of CGH array in determining absolute CNVs. Our method thus provides a new approach to interpret CGH array data for personalized medicine.

INTRODUCTION

Genomic variants, including single nucleotide polymorphisms (SNPs) and structural variations (SVs), have been utilized to identify genetic factors underlying many complex diseases. Copy number variations (CNVs) are the most abundant form of SV (1). The effects of CNVs on human complex diseases have been widely assessed over the past few years, after the high-throughput detection of CNVs became technically feasible (2–7). The recently constructed common ultra-high resolution CNV maps in human populations are being widely used to study both the distribution and impact of CNVs on complex human traits, including complex diseases (8,9).

Comparative genomic hybridization (CGH) array technology has been widely used for genome-wide detection of CNVs (1,5,6,8–10), followed more recently by a massively parallel re-sequencing-based approach, including pair-end read mapping (PEM) and read-depth (RD) analysis (11–19). Although sequencing-based approaches offer some advantages for detecting *de novo* CNVs as well as accurately determining CNV breakpoints, there is a growing need for standardization of these methods. In particular, short reads from high or low GC genomic regions and repetitive genomic regions, such as segmental duplications and simple repeats, provide an unreliable RD of sequence coverage (14,17). In addition, the PEM method is inefficient for detecting large CN gains and small CN losses. To be more accurate, sequencing methods also require comparison of multiple individual genomes, which have yet to be standardized (18–20). Due to its accuracy and cost-effectiveness, CGH arrays have remained the most frequently used methods for genotyping personal CNVs, both for association studies and for developing personalized medicine.

The CGH array approach, however, depends heavily on an arbitrary reference sample, severely limiting the utility of this method. Since this method compares the amounts of DNA from samples of interest (test samples) and a reference sample hybridized to oligonucleotide probes, any aberrancy in DNA quantity of the reference sample

*To whom correspondence should be addressed. Tel: +82 2 740 8246; Fax: +82 2 741 5423; Email: jeongsun@snu.ac.kr
Correspondence may also be addressed to Jong-Il Kim. Tel: +82 2 740 8421; Fax: +82 2 741 5423; Email: jongil@snu.ac.kr

These authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

due to its CNVs introduces systematic inaccuracy in determining the absolute CNVs of the test samples. The reference sample is not an ideal DNA, as it may include thousands of genomic variants, resulting in misidentifications and misinterpretations of test sample CNVs during CGH array experiments (6). For example, a CN normal region (copy number 2) in a test sample may be reported as a gain or loss in CN, or as normal depending on the CN of the reference sample (Supplementary Figure S1). Likewise, CNVs of the test sample may be undetected due to CNV of the reference sample. This dependency on a reference sample may have been the main cause for the low level of concordance between CNVs from CGH arrays and massively parallel sequencing (11–13,17,18).

Although pooled DNA may resolve reference biases, it may also decrease the power of CNV detection in highly CN polymorphic regions (21). Alternatively, absolute CN can be assessed using a clustering algorithm (8); however, this method may not always be applicable or accurate when CGH array experiments do not include sufficient number of DNA samples. Moreover, CGH array data obtained by multiple platforms cannot be clustered. Thus, to include CNV information in personalized medicine, it is important to develop new methods for accurately determining absolute CNVs using CGH arrays independent of an arbitrary reference DNA.

DNA from a male of European ancestry, NA10851, has been used as a reference for CGH array in most studies of human CNVs (5,6,8–10,13,17,22), and many of the future CGH array experiments will undoubtedly use the identical reference sample. Hence, understanding CNV information specific to NA10851 is critical for interpretation of CGH array experiments. We recently reported absolute Asian CNVs using preliminary CNV information from NA10851 (9). Here we report the complete personal CNVs of NA10851 by both CGH array and massively parallel sequencing. In addition, we describe CARA algorithm, which enables simple derivation of accurate reference-independent CNV genotyping. This general algorithm is compatible with any CGH array platform. Finally, we report user-friendly software to implement this algorithm.

MATERIALS AND METHODS

Whole genome sequencing

The cell-line of NA10851 was acquired from the Coriell Cell Repository (http://ccr.coriell.org). Genomic DNA was extracted from the cell line using standard protocols. Libraries for massively parallel sequencing (Illumina Genome Analyzer IIx) were constructed according to the manufacturer's standard protocol (Illumina, Inc., USA). We constructed three different libraries, which were designed to have inserts of 500 bp between paired-end reads. Paired-end sequencing was performed using three read lengths, 2×36, 2×76 and 2×101 bp. All reads were aligned to the human reference genome (assembly build 36.3, hg18) using GSNAP alignment tool (17,23). A single position was randomly determined when short reads had multiple positions with identical highest alignment scores. To identify SNPs and short indels of NA10851 from sequencing data, we used our own scripts, which merge read alignment results and extract SNPs and short-indels, as described previously (17).

Using our own script, the RD of sequence coverage was calculated for each base of the human reference genome assembly build 36.3 (hg18). The effects of GC contents on the RD of sequence coverage were adjusted using a modification of the described method (18), in that we determined the relationship between GC contents and RD of coverage in 100 bp windows and adjusted the single-base RDs according to this relationship.

Ultra-high resolution CGH data included in this study

We analyzed ultra-high resolution CGH array data previously reported from 73 individuals, which used NA10851 as the reference sample (8,9). By pooling those CNV segments, we obtained putative NA10851 CNV information. Overall, the data included 43 911 CNV segments of 40 individuals (20 unrelated Europeans and 20 unrelated West Africans) using the NimbleGen 42 M-probe and 24 194 filtered CNV segments of 33 individuals (31 unrelated East Asians, 1 European and 1 West African) using the Agilent 24 M-probe CGH array.

Identifying NA10851 CNV regions

Genome-wide CNVs of NA10851 were screened using high-resolution CGH array data and confirmed by RD of massively parallel sequencing. We assumed that all 68 105 CNV segments were NA10851 candidate CNV regions. For each CNV segment, the deviation of RD from the average was calculated using the equation ($\text{RD ratio} = \overline{\text{RD}}_{\text{NA10851, CNV segment}} / \overline{\text{RD}}_{\text{NA10851, whole genome}}$, $\overline{\text{RD}}_{\text{NA10851, whole genome}} = 25.25\times$ and $12.67\times$ for autosomes and the chromosome X, respectively). Segments with RD ratio ≤ 1.15 and ≥ 0.85 were considered normal regions and removed ($n = 21\,479$). The remaining redundant CNV segments ($n = 46\,626$) were collapsed into minimally redundant CNV elements (CNVEs) using the criteria previously suggested [>50% reciprocal overlap (9)]. Of a total of 6499 CNVEs, 3002 with only one segment were considered false positives and removed.

To compare the RD of 3497 CNVEs with their flanking regions (defined as four times lengths of CNV candidates to upstream and downstream), we calculated the median value and standard deviation (SD) of RD for CNV candidates and their flanking regions. For CNVE size >1 Mb, >100 kb, >1 kb and >100 bp, we used window size of 1 kb, 100 bp, 50 bp and 30 bp, respectively. The deviation value (Z) was calculated using the formula:

$$Z_{5'(3')} = \left| \frac{\text{median RD}_{\text{CNVE}} - \text{median RD}_{5'(3')}}{\sqrt{\text{SD}_{\text{CNVE}}{}^2 / N_{\text{CNVE}} + \text{SD}_{5'(3')}{}^2 / N_{5'(3')}}} \right|$$

where N is the number of windows for the regions. A CNVE was selected when its $Z_{5'}$ and $Z_{3'}$ were ≥ 1.96 (level of significance ≤ 0.05). Subsequent visual inspection

of the 2621 remaining CNVEs showed that they could be categorized into four groups (see 'Results' section and Figure 1c for more details of visual inspection):

(i) Apparent CNV: RD was sufficiently stable for reliable CNV detection. In addition, RD of the candidate region was clearly higher or lower than those of its flanking regions.
(ii) Indistinct region: RD was stable, but the magnitude of the RD difference between the candidate region and its flanking regions was minimal.
(iii) Indeterminable region: RD was highly unstable. Most of the candidates within extreme segmental duplication regions (Supplementary Table S4) were categorized into this group.
(iv) Nested CNV: The candidate region was part of another apparent CNV, which better represents the CNV.

We finally selected 1309 CNVEs in Group 1 for NA10851 filtered CNV regions. CNVEs clearly shown to

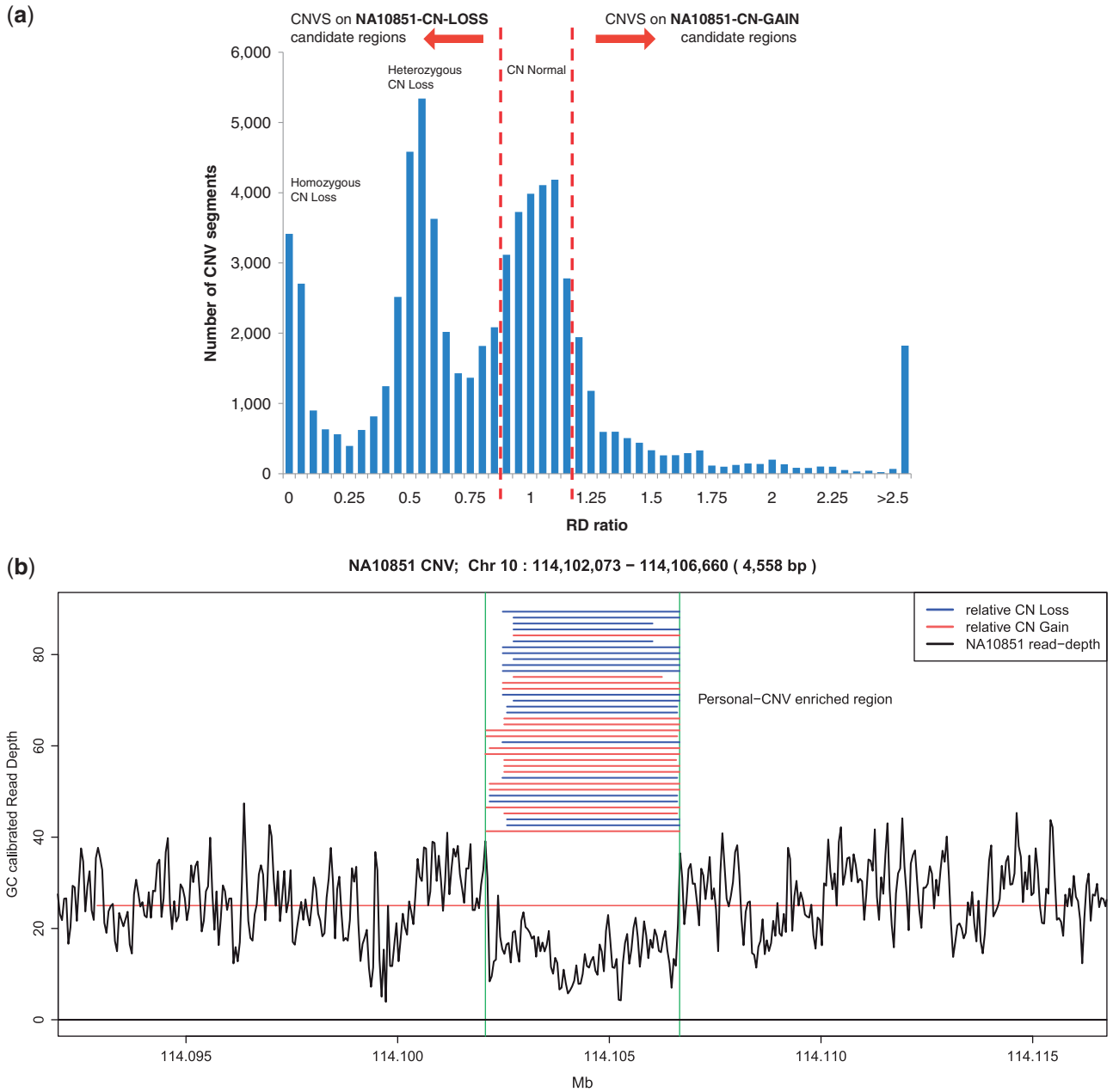

Figure 1. Detection of NA10851 CNVs using conjugative methods, massively parallel sequencing and ultra-high resolution CGH arrays. (**a**) Distribution of RD of coverage of NA10851 sequencing for CNV segments identified from CGH arrays of 73 individuals. (**b**) Identifying CNVs of NA10851 using RD of sequence coverage on putative regions determined by CGH arrays. (**c**) Examples of four categories of candidate CNVs by visual inspection. Top row, Apparent CNV; second row, Indistinct region; third row, Indeterminable region due to extremely unstable RD; fourth and bottom, Nested CNV, the CNV candidate in the bottom row is removed since it is included in the CNV illustrated in the fourth row.

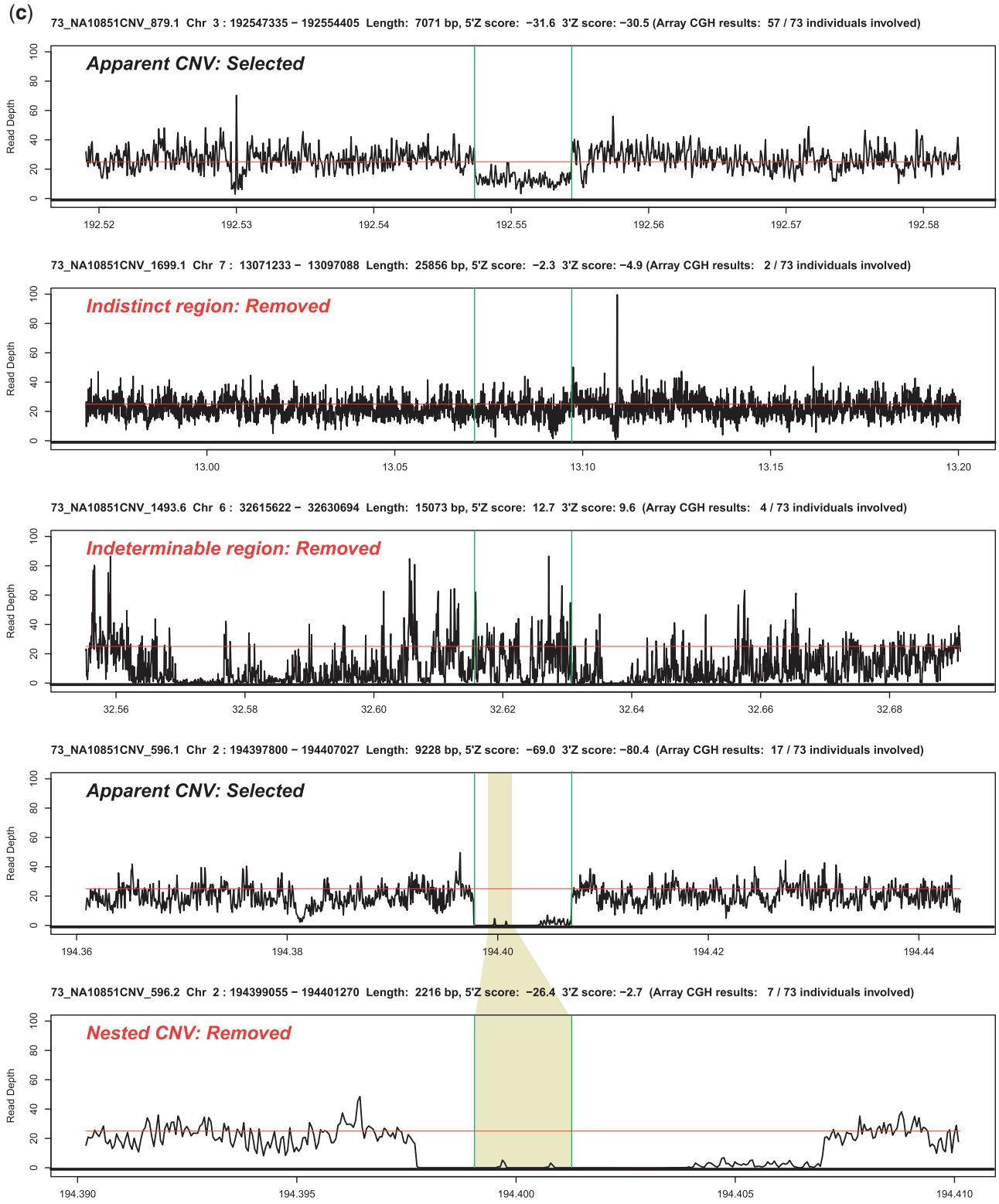

Figure 1. Continued.

have a 0 RD of coverage were determined to be homozygous CN losses.

The genomic overlaps between the filtered CNVs and human genes as well as repetitive sequence (such as segmental duplication and simple repeats) were examined using information downloaded from UCSC genome browser (http://genome.ucsc.edu). We used refGene.txt (July 2009), genomicSuperDups.txt

(February 2010) and simpleRepeat.txt (February 2010) for human genes, segmental duplications and simple repeats, respectively. When a CNV overlap ≥1 bp of a genic region, we considered it gene related. Likewise, CNVs with >50% overlap by segmental duplication or simple repeats were so categorized.

NA10851 genome browser and CARA software

We developed a genome browser for NA10851 showing genome-wide RD information as well as ultra-high resolution CGH array data (http://cara.gmi.ac.kr). All information on genomic variations information as well as our CARA software can be downloaded from this website.

CARA software was developed using a tool of Microsoft Visual Studio .Net 2008. An importing module of high-resolution CGH array data was implemented using a Microsoft 'mscorlib' library, and CNV data of NA10851 used in the adjusting module were deposited in an ASCII flat file format in CARA software. The efficiency of CARA in adjusting CNV regions and in processing large amounts of data was confirmed by simulation using a set of Agilent 24 M high resolution CGH array data of AK1 (9).

Validation of the accuracy of CARA using AK1 data

We validated the accuracy of the CARA by showing its improvement in the genome-wide concordance between the RD of sequence coverage and CGH array data using AK1 (9,17). The 24 M probe CGH array data of AK1 was CNV segmented using the ADM2 algorithm with Agilent Genomic Workbench Standard Edition 5.0.14 as described previously (9). Absolute CNVs were obtained in the same way after CARA was applied to the CGH array data with a centralizing coefficient $\alpha = 0.7$ (see 'Results' section for details). We used the filtering criteria $|\log_2\text{ratio}| \geq 0.5$ and $P\text{-value} \leq 1 \times 10^{-22}$ to remove false positives. The RD for each CNV segment was calculated using the short-read data reported previously (17), and the concordance was determined between its deviation from the RD of its flanking regions and the corresponding $\log_2$ratio from CGH array data.

Comparison of CNVs between before and after applying CARA

We compared relative and reference-independent CNV sets of AK1. The CNV segments in each group that did not overlap any of the other groups by 1 bp overlap were categorized as unique to that group. Overlapping ones are further categorized as follows: when the magnitude of overlapping length was <80% reciprocally, they were classified as 'size altered'; those with ≥80% reciprocal overlap but a change in $\log_2$ ratio ≥ 0.3 were classified as '$\log_2$ ratio altered'; and those CNVs with ≥80% reciprocal overlap but without a significant change in $\log_2$ ratio (<0.3) were categorized as 'no change'.

RESULTS

Cataloging genome-wide NA10851 CNVs

We analyzed 84.9 Gb of massively parallel sequencing data from three paired-end libraries constructed from the genomic DNA of NA10851 (Table 1). The short reads were aligned to the reference human genome assembly build 36.3 (hg18) using the GSNAP alignment tool as described previously (17,23). Overall, we covered 25.01 times the haploid genome. In total, we identified 3 683 016 SNPs and 319 174 short insertion deletion polymorphisms (Supplementary Tables S1–S3).

To use genome-wide RD information, we calibrated the effect of GC contents on sequence coverage (14). The genome-wide read depth of NA10851 is shown on the genome browser on our website (http://cara.gmi.ac.kr). CNV detection from RD data alone is not sufficiently accurate, since the RD normally fluctuates even without CNV in the genomic region (17). In order to solve this problem, we first obtained putative NA10851 CNV regions using multiple ultra-high resolution CGH arrays, and subsequently confirmed these regions using their RD. We analyzed the 68 105 CNV segments from 73 individuals detected by two different kinds of ultra-high resolution CGH array platforms that used NA10851 as a reference (see 'Materials and Methods' section for details) (8,9). The RD ratios of all the CNV segments were analyzed to identify the CNVs on putative NA10851 CNV regions (see 'Materials and methods' section for details). The distribution of RD ratios showed three clear peaks, near 0, 0.5 and 1.0, equivalent to 0, 1 and 2 CN regions, respectively, of NA10851 (Figure 1a). Some CNV segments showed RD ratio significantly >1.0, indicating CN gains in NA10851.

We removed the CNV segments corresponding to the NA10851 CN normal regions by initial filter criteria (segments with RD ratio ≤1.15 and ≥0.85 were removed; see 'Materials and Methods' section 'Identifying NA10851 CNV regions' for more detailed method). Then we collapsed the 46 626 remaining segments into 6499 minimally redundant CNVEs for the first-filtered NA10851

Table 1. Summary of massively parallel sequencing of NA10851

Library	Read length	Insert size	Total reads	Aligned reads	Aligned bases	Genome covered	RD	Total SNPs	Total Indels
Library #1	2 × 36 bp	500 bp	557 060 528	499 554 933	35 966395 532				
Library #2	2 × 76 bp	500 bp	159 462 248	125 776 263	19 116 529 680				
Library #3	2 × 101 bp	500 bp	101 921 217	81 211 013	16 403 136 593	99.71%	25.01×	3 683 016	319 174
Overall	–	–	818 443 993	706 542 209	71 486 061 805				

RD: read-depth of sequence coverage.

CNV candidates. Subsequently, CNVEs which contained single segments ($N = 3002$) were removed, since singletons are likely not true NA10851 CNV regions, inasmuch as CNV regions of the reference DNA cause the regions to appear as frequent relative CNV regions in multiple independent CGH array experiments (Figure 1b). Moreover, false positive CNVs on CGH array were likely to be detected as singletons among populations, since they tended to be randomly distributed throughout the entire genome and are therefore difficult to overlap.

Next, the RDs of the 3497 remaining CNVEs were systematically compared with the RDs of their 5′ and 3′ flanking regions. This step identified 2621 suggestive CNVEs for the second-filtered NA10851 CNV regions. Finally, the RDs of the 2621 CNVEs with their flanking regions were visually inspected, in order to remove unreliable regions (Figure 1c; see 'Materials and Methods' section). Thus, of a total of 2621 CNV candidates, 599 were considered indistinct and removed. We also excluded 328 candidates located in 16 genomic regions where the RDs are highly ambiguous due to extreme segmental duplications, making determinations of CNs for NA10851 unreliable (Supplementary Table S4). In addition, 385 nested CNVs were removed. Finally, 1309 apparent CNVs and their RD ratios, which would be inverted and used as adjusting coefficient in CARA, were determined (Figure 2a and b, Supplementary Tables S5 and S6, Supplementary Figure S2; see 'Materials and Methods' section for details).

The total and median lengths of the NA10851 CNVs were 23.34 Mb and 2.7 kb, respectively. CN loss was observed in 1023 regions and CN gain in 286. Among the 1023 losses, 178 were considered homozygous. CN gains and losses showed different size distributions (Figure 2b), consistent with our previous report (9). Of the 1023 CN losses and 286 gains, 594 (58.0%) and 59 (20.6%), respectively, were not involved by segmental duplications or simple repeats (Figure 2c), indicating that segmental duplications are more enriched in CN gains than in losses.

CARA: algorithm for reference-independent CNV detection

We developed CARA to detect of reference-independent CNVs in CGH arrays (Supplementary Figure S3). Using the RD ratios of NA10851 CNV regions, this algorithm modifies the signal intensity (SI) for the reference sample (NA10851) and the $\log_2$ ratio of target probes within the range of the 1309 NA10851 CNV regions (Figure 3). These corrections thus normalize the biased-$\log_2$ ratio as if the reference sample has normal CNs. Autosomes and sex-chromosomes are analyzed independently, since the median signal intensity for sex chromosomes is nearly half that of autosomes. If an overlapping NA10851 CNV region is not a homozygous CN loss, we corrected the SI for the reference sample using the adjusting coefficient (ρ), where $\rho = \text{median RD}_{\text{NA10851, whole genome}}/\text{median RD}_{\text{NA10851, CNVE}}$ (inverse of RD ratio). Then, the reference SI for each probe is adjusted according to the simple equation, $\text{SI}'_{\text{ref, probe}} = \text{SI}_{\text{ref, probe}} \cdot \rho$, where $\text{SI}_{\text{ref,probe}}$ is the signal intensity of each probe for the reference sample. The $\text{SI}'_{\text{ref, probe}}$ can be considered the signal intensity for two (normal) CN state. Then the reference-independent $\log_2$ratio for the probe is calculated using the formula:

$$\log_2 \text{ratio}' = \log_2\left(\frac{\text{SI}_{\text{test, probe}}}{\text{SI}'_{\text{ref, probe}}}\right) = \log_2\left(\frac{\text{SI}_{\text{test, probe}}}{\text{SI}_{\text{ref, probe}}}\right) - \log_2 \rho$$
$$= \log_2 \text{ratio}_0 - \log_2 \rho$$

where $\log_2\text{ratio}_0$ is the unadjusted $\log_2$ratio for the probe.

For regions of homozygous loss, the above formula is inapplicable, since in theory, $\text{SI}_{\text{ref, probe}} = 0$ and $\rho = \text{infinity}$. We therefore utilized another strategy, by replacing $\text{SI}_{\text{ref,probe}}$ with the median SI of the reference sample on the corresponding arrays ($\text{SI}'_{\text{ref, probe}} = \text{median SI}_{\text{ref, array}}$). To reduce the false positive rate due to the arbitrary replacement, the signal intensity of the test sample of the corresponding probes was also modified using a pre-determined centralizing coefficient (α): $\text{SI}'_{\text{test, probe}} = \text{median SI}_{\text{test, array}} + \alpha.(\text{SI}_{\text{test, probe}} - \text{median SI}_{\text{test, array}})$. The coefficient α can range from 0 (most conservative centralization) to 1 (not at all conservative) according to the researcher's choice. This centralization method reflects the characteristics of each probe by controlling the magnitude of signal intensity deviation from the median value in the test sample. Our validation suggested that 0.7 was optimal value for α in our experimental setting (Supplementary Figure S4). Then the reference-independent $\log_2$ratio for the probe can be calculated as:

$$\log_2 \text{ratio}' = \log_2\left(\frac{\text{SI}'_{\text{test, probe}}}{\text{SI}'_{\text{ref, probe}}}\right)$$
$$= \log_2\left(\frac{\text{median SI}_{\text{test, array}} + \alpha.\ (\text{SI}_{\text{test, probe}} - \text{median SI}_{\text{test, array}})}{\text{median SI}_{\text{ref, array}}}\right)$$

We developed and released the first version of user-friendly and freely available software implementing CARA, which can be downloaded from the website (http://cara.gmi.ac.kr). Data from Agilent (feature extracted data) or NimbleGen platforms (.pair files) can be analyzed with the current version of CARA. In a simulation, the time required to adjust a 24×1 M probes Agilent platform was ~40 min using 8 GB of main memory on an Intel(R) Core(TM)2 Duo CPU E8500 3.16 GHz machine.

Validation of the accuracy of CARA using AK1 CGH array data

To validate its accuracy, we applied the CARA to the previously reported 24 M probe CGH array data for AK1 genome (9,17). The relative and reference-independent CNV sets (before and after CARA, respectively) were compared with the RD of AK1 sequence coverage. The total numbers of CNV segments found were similar (Figure 4a; $N = 535$ and 598, respectively), but, their contents were quite different. When we compared the two sets of CNV segments, we found that

Figure 2. Personal CNVs of NA10851. (**a**) Personal CNV distribution throughout the entire genome. (**b**) Numbers and lengths of CN losses and CN gains of NA10851. (**c**) Size distribution of 1309 NA10851 CNV regions. (**d**) Repetitive context of CN gains and losses.

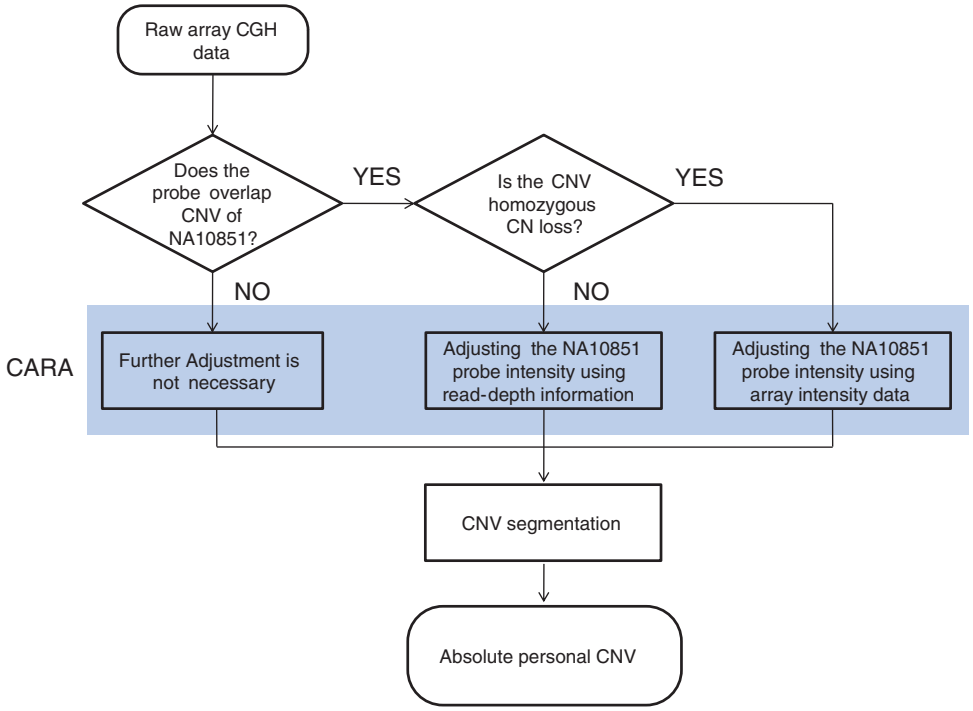

Figure 3. Flowchart showing the CARA for detection of reference-independent CNVs using CGH arrays.

39.1% of relative CNVs were removed (Figure 4b) and were considered CNVs of NA10851 instead of AK1. In addition, 21.7% of the CNV segments showed changed in their sizes as well as altered $\log_2$ ratios from the relative set. Only 39.3% of the relative CNVs were not related to the CNVs of NA10851, so that their $\log_2$ ratios and sizes did not change markedly. After applying CARA, 280 CNVs were rescued; these were undetected in the relative sets because both the test and reference samples have identical CNVs. As a result, the concordance rate between the CGH array and RD was markedly enhanced from 49.7 to 88.8% (Figure 4a and c). To our knowledge, this is the highest concordance rate between CGH array results and massively parallel sequencing data.

DISCUSSION

We have shown here an accurate personal CNV map of NA10851. This sample has been widely used as a reference sample for CGH array experiments (10,21). The early phases of CNV discovery studies have focused on determining CN variable genomic regions among entire human populations, for which detection of absolute CNVs has not been critical. As high-resolution CGH array platforms have become available, more precise CNV maps of human populations have been generated (4–6,8,9). To employ CNVs in personalized medicine, it is imperative to identify personal CNVs accurately. The accuracy of CGH arrays has been often compromised because the effects of CNVs of the reference sample were not removed, the final results, therefore, have been biased. Thus, it is critical to identify reference CNVs and also develop a streamlined approach to remove their influence from CGH arrays. We utilized a systematic approach to identify CNVs of the common reference sample by combining information from the personal genome sequence obtained from massively parallel sequencing data and ultra-high resolution CGH arrays obtained from 73 individuals using the sample as a reference. Compared with the personal genomes ascertained by a single technology, such as massively parallel sequencing or CGH array only, the NA10851 genome revealed by a combination of these methods allowed us to obtain the most accurate estimates of personal CNVs to date.

The predominance of numbers of CN losses over gains is in good agreement with previous reports (8,9). Although we believe that this is the predominant characteristic feature of human CNVs, some technical issues are worth considering. Generally, CN losses are easier to identify than CN gains, using both the hybridization and the RD methods. Especially, high proportion of CN gains are placed on duplicated genomic regions, therefore it is difficult to design unique probes of good quality for CGH array or to align short-reads in resequencing methods. In addition, the numbers of copies of DNA segments in repetitive regions, such as microsatellites, vary almost continuously among human populations, making detection of integer CN difficult. Moreover, the insertion of DNA sequences, which are not found on the human reference genome, cannot be detected using general CGH array. Approaches that do not depend on the human reference genome, such as *de novo* assembly are therefore needed to identify all the CN gains and their exact integer CNs in a personal genome.

NA10851 is the most widely employed individual genome in CGH arrays. Therefore, information on its

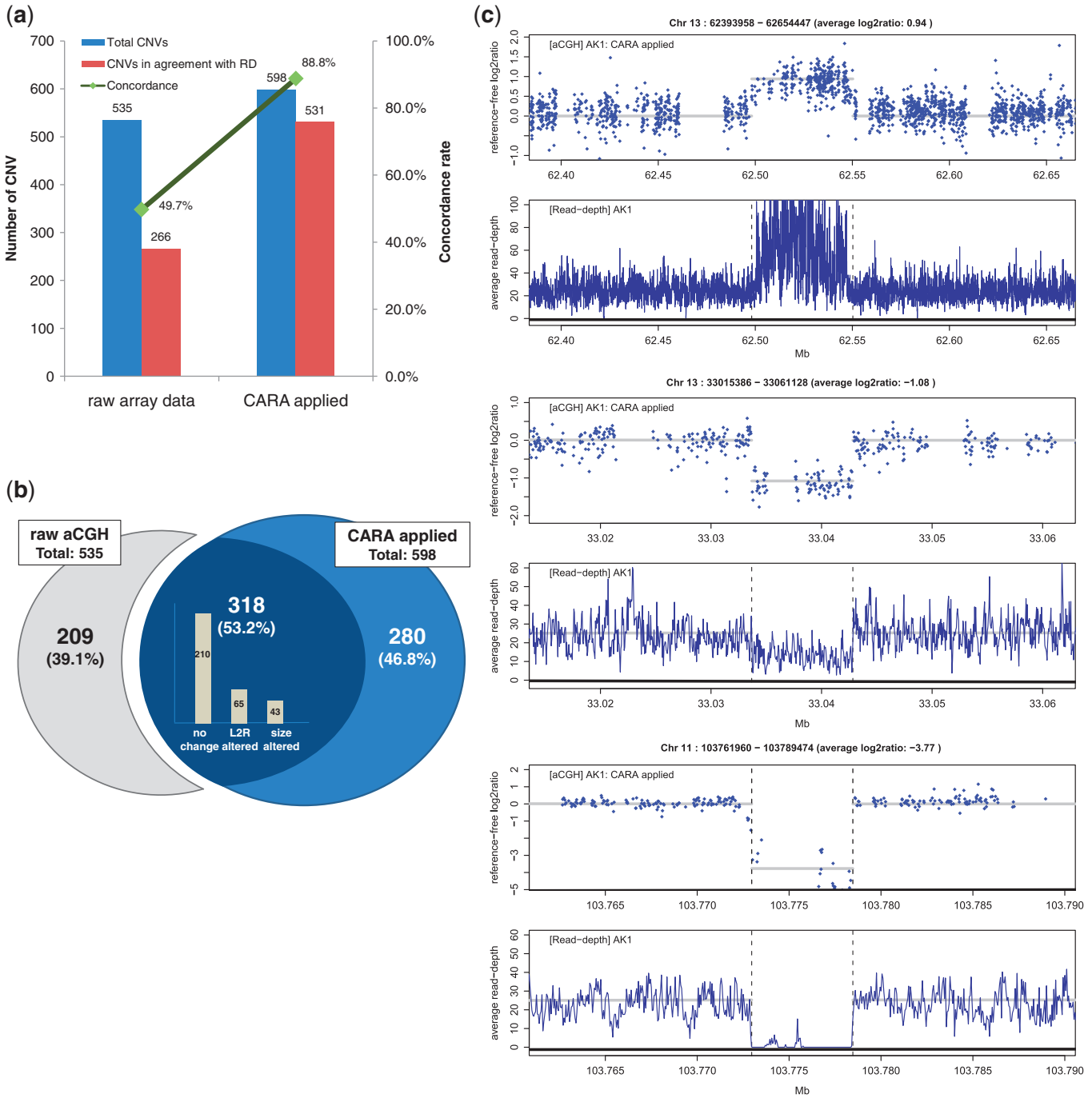

Figure 4. Evaluation of the utility of CARA. (**a**) Comparisons between CNV sets of AK1 before and after application of CARA. (**b**) Alterations in CNV segments upon application of CARA. (**c**) Concordance between CGH arrays and RD of sequence coverage after application of CARA.

genomic variants and CARA will contribute toward accurately estimating CNVs and their utility in personalized medicine. Most human genomic variations have been analyzed, cataloged and annotated in public databases based on the 'Human Reference Genome', which has been sequenced and assembled by Human Genome Project (24). CARA enables the determination of personal CNVs based on the human reference genome rather than on an arbitrary sample NA10851. The high concordance rate after CARA between CGH array and RD shows the utility of CARA for accurately identifying personal CNVs. Using CARA, absolute CNVs from a variety of DNA samples, including cancer cells and mosaic samples, can be assessed only if NA10851 is used as a reference for CGH arrays. A more accurate determination of the genomic variants of NA10851 can increase the accuracy of adjustment from CARA. Therefore, it is critical to collect and release information on the genomic variants on NA10851, such as newly detected CNVs, or more precise CNV breakpoints. In addition, further

deeper sequencing of NA10851, which will provide much higher RD and more accurate RD ratio, will be also valuable for finer adjustment. We have opened the database of NA10851 genomic variants on the website (http://cara.gmi.ac.kr) and we have released all relevant information, such as CNVs, SNPs, short-indels and RD of NA10851. We hope to get new information from the community. As the information is updated, new versions of CARA will be released.

Along with CGH arrays, massively parallel sequencing is a powerful tool for identifying CNVs. However, systematic differences in CNVs calls due to the use of an arbitrary reference sample in CGH arrays have interfered with complete sample-matched CNV comparisons between the two technologies (12,13,17–18). By correcting the demerits of CGH arrays using CARA, we were able to obtain the highest sample-matched concordance between the technologies.

To assess the impact of human CNVs, integer CNs (e.g. 0, 1, 2, 3) of each segment should be genotyped. Although CARA enables the detection of normal CN, as well as CN gains and losses, integer CN cannot be assessed, especially in CN gains. The methodology for accurately identifying personal structural variations will be improved continuously as new algorithms are developed by ultra-high-resolution CGH arrays and massively parallel sequencing. Ultimately, a rapid algorithm for detecting the integer CN of genes will be developed by combining all the CNV data available. These efforts will enable the identification of disease-related CNVs, as well as understanding their role in the pathophysiology of complex human diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank D. R. Govindaraju at Boston University for his personal comments regarding the manuscript.

FUNDING

Korean Ministry of Education, Science, and Technology (grant number 2010-0013662), Green Cross Therapeutics (0411-20080023). Funding for open access charge: Korean Ministry of Education, Science, and Technology (grant number 2010-0013662).

Conflict of interest statement. None declared.

REFERENCES

1. Hurles,M.E., Dermitzakis,E.T. and Tyler-Smith,C. (2008) The functional impact of structural variation in humans. *Trends Genet.*, **24**, 238–245.
2. Lucito,R., Healy,J., Alexander,J., Reiner,A., Esposito,D., Chi,M., Rodgers,L., Brady,A., Sebat,J., Troge,J. *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, **13**, 2291–2305.
3. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
4. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
5. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
6. Perry,G.H., Ben-Dor,A., Tsalenko,A., Sampas,N., Rodriguez-Revenga,L., Tran,C.W., Scheffer,A., Steinfeld,I., Tsang,P., Yamada,N.A. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, **82**, 685–695.
7. Zhang,F., Gu,W., Hurles,M.E. and Lupski,J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
8. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C., Campbell,P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
9. Park,H., Kim,J.I., Ju,Y.S., Gokcumen,O., Mills,R.E., Kim,S., Lee,S., Suh,D., Hong,D., Kang,H.P. *et al.* (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, **42**, 400–405.
10. Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
11. Tuzun,E., Sharp,A.J., Bailey,J.A., Kaul,R., Morrison,V.A., Pertz,L.M., Haugen,E., Hayden,H., Albertson,D., Pinkel,D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
12. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
13. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
14. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
15. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Guo,Y. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
16. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
17. Kim,J.I., Ju,Y.S., Park,H., Kim,S., Lee,S., Yi,J.H., Mudge,J., Miller,N.A., Hong,D., Bell,C.J. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
18. Yoon,S., Xuan,Z., Makarov,V., Ye,K. and Sebat,J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
19. Alkan,C., Kidd,J.M., Marques-Bonet,T., Aksay,G., Antonacci,F., Hormozdiari,F., Kitzman,J.O., 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.
20. Chiang,D.Y., Getz,G., Jaffe,D.B., O'Kelly,M.J., Zhao,X., Carter,S.L., Russ,C., Nusbaum,C., Meyerson,M. and Lander,E.S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
21. Scherer,S.W., Lee,C., Birney,E., Altshuler,D.M., Eichler,E.E., Carter,N.P., Hurles,M.E. and Feuk,L. (2007) Challenges and

standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
22. Perry,G.H., Yang,F., Marques-Bonet,T., Murphy,C., Fitzgerald,T., Lee,A.S., Hyland,C., Stone,A.C., Hurles,M.E., Tyler-Smith,C. *et al.* (2008) Copy number variation and evolution in humans and chimpanzees. *Genome Res.*, **18**, 1698–1710.
23. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
24. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.