Sensitive and accurate detection of copy number variants using read depth of coverage

Seungtai Yoon,[1] Zhenyu Xuan,[1] Vladimir Makarov,[1] Kenny Ye,[2,3] and Jonathan Sebat[1]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [2]*Albert Einstein College of Medicine, Bronx, New York 10461, USA*

Methods for the direct detection of copy number variation (CNV) genome-wide have become effective instruments for identifying genetic risk factors for disease. The application of next-generation sequencing platforms to genetic studies promises to improve sensitivity to detect CNVs as well as inversions, indels, and SNPs. New computational approaches are needed to systematically detect these variants from genome sequence data. Existing sequence-based approaches for CNV detection are primarily based on paired-end read mapping (PEM) as reported previously by Tuzun et al. and Korbel et al. Due to limitations of the PEM approach, some classes of CNVs are difficult to ascertain, including large insertions and variants located within complex genomic regions. To overcome these limitations, we developed a method for CNV detection using read depth of coverage. Event-wise testing (EWT) is a method based on significance testing. In contrast to standard segmentation algorithms that typically operate by performing likelihood evaluation for every point in the genome, EWT works on intervals of data points, rapidly searching for specific classes of events. Overall false-positive rate is controlled by testing the significance of each possible event and adjusting for multiple testing. Deletions and duplications detected in an individual genome by EWT are examined across multiple genomes to identify polymorphism between individuals. We estimated error rates using simulations based on real data, and we applied EWT to the analysis of chromosome 1 from paired-end shotgun sequence data (30×) on five individuals. Our results suggest that analysis of read depth is an effective approach for the detection of CNVs, and it captures structural variants that are refractory to established PEM-based methods.

[Supplemental material is available online at http://www.genome.org.]

Structural variants (SVs) in the human genome (Iafrate et al. 2004; Sebat et al. 2004; Feuk et al. 2006a), including copy number variants (CNVs) and balanced rearrangements such as inversions and translocations, play an important role in the genetics of complex disease. Analysis of CNV in diseases such as cancer (Lucito et al. 2000; Pollack et al. 2002; Albertson and Pinkel 2003), and in developmental and neuropsychiatric disorders (Feuk et al. 2006b; Sebat et al. 2007; Kirov et al. 2008, 2009; Marshall et al. 2008; Mefford et al. 2008; Rujescu et al. 2008; Stefansson et al. 2008; Stone et al. 2008; Walsh et al. 2008; Zhang et al. 2008), has led to the identification of novel disease-causing mutations, thus contributing important new insights into the genetics of these disorders.

Our current power to detect SVs in disease studies is limited by the resolution of microarray analysis. Currently available array platforms that consist of more than 1 million probes have a lower limit of detection of ~10–25 kb (McCarroll et al. 2008; Cooper et al. 2008). More comprehensive studies of individual genomes using sequencing-based approaches are capable of detecting CNVs <1 kb in size (Tuzun et al. 2005; Korbel et al. 2007; Bentley et al. 2008; Wang et al. 2008). Thus, new sequencing technologies promise to enable more comprehensive detection of SVs as well as indels and point mutations (Mardis 2008).

New computational methods are needed that can reliably identify SVs using next-generation sequencing platforms. To date, multiple approaches have been developed for the detection of SVs that are based on paired-end read mapping (PEM), which detects insertions and deletions by comparing the distance between mapped read pairs to the average insert size of the genomic library (Tuzun et al. 2005; Korbel et al. 2007). Advantages of this approach include the sensitivity for detecting deletions <1 kb in size, and localizing the breakpoint within the region of a small fragment. This approach also has certain limitations. In particular, PEM-based methods have poor ascertainment of SVs in complex genomic regions rich in segmental duplications and have limited ability to detect insertions larger than the average insert size of the library (Tuzun et al. 2005).

We sought to develop an alternative approach to the detection of SVs from sequence data that compliments existing methods. Here we used the depth of coverage in sequence data from the Illumina Genome Analyzer to look for genomic regions that differ in copy number between individuals. This method is based on the depth of single reads and, hence, is orthogonal to methods that are based on the mapping of paired-end sequences.

To detect CNVs based on read depth (RD), we developed a pipeline consisting of three steps, as illustrated in Figure 1: (1) First, we estimated the coverage or RD in nonoverlapping intervals across an individual genome, (2) we implemented a novel CNV-calling algorithm to detect events, and (3) we compared data from multiple individuals to distinguish events that are polymorphic (i.e., CNVs) from those that show similarly increased or decreased copy number in all individuals in this study (i.e., mononomorphic events). Here we demonstrate the feasibility of this approach and its unique advantages in comparison with other methods of SV detection.

Methods

Data sets included in this study

Genome sequence data from five individuals were analyzed in this study. These include a CEU trio of European ancestry (NA12878,

[3]Corresponding author.
E-mail kye@aecom.yu.edu; fax (718) 430-8699.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.092981.109. Freely available online through the *Genome Research* Open Access option.

19:1586–1592 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

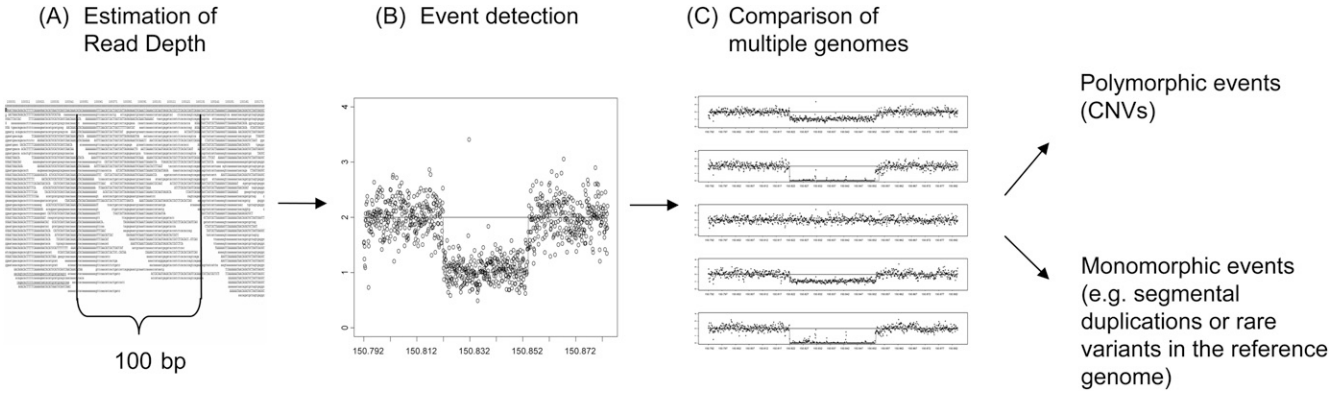

Figure 1. Pipeline for the detection of CNVs based on analysis of read depth (RD). (*A*) RD was determined by counting the start position of reads in nonoverlapping windows of 100 bp. (*B*) Events were detected using a custom CNV-calling algorithm, event-wise testing (EWT). (*C*) Each event was examined in multiple genomes in order to distinguish polymorphic events (CNVs) from the majority of events that were found to show a similar copy number change in all five genomes in this study (i.e., monomorphic events).

NA12891, and NA12892) sequenced as part of the pilot project of the 1000 Genomes Project (http://www.1000genomes.org; L Brooks, pers. comm.), and two additional published genomes, including a Yoruban individual NA18507 (Bentley et al. 2008) and a Chinese individual (Wang et al. 2008). For the CEU trio, we obtained complete genome sequence data in the form of ".bam" alignment files from ftp://ftp-trace.ncbi.nih.gov/1000genomes/. These mappings were generated with the MAQ alignment method using the default settings (Li et al. 2008) as described in the documentation of the December 2008 data release (ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/release/2008_12/). The complete sequence data (.fastq files) on two additional genomes (NA18507 and YH) were obtained from ftp sites designated by Bentley et al. (2008) and Wang et al. (2008), and reads were mapped to the human reference genome in HG18 using the same methods. Alignment (.bam) files were parsed out using SAMtools (samtools.sourceforge.net), and we then filtered out reads of low mapping quality (<Q30).

Estimation of coverage from genome sequence data

For each sample, RD was measured by counting the number of mapped reads in 100-bp windows, assigning each read only once by its start position. We chose a window size of 100 bp for multiple reasons. A larger window size (e.g., of 1000 bp) would provide less precision in defining the breakpoints of CNVs. A larger window size could also make the detection of small (~1000 bp) CNVs problematic, because in many cases these CNVs would only partially span one or two windows. In addition, at 30× coverage, the distribution of read counts of 100-bp windows are well approximated by a normal distribution, thus permitting us to assume normality in our statistical calculations (see below), while read counts in smaller window sizes are not (Supplemental Fig. 1).

Sequence coverage on the Illumina Genome Analyzer platform is influenced by GC content, as first described by Bentley et al. (2008). We have observed a similar effect in all of the data sets from this study (see Results section). Therefore, we sought to adjust the 100-bp window read counts based on the observed deviation in coverage for a given G+C percentage. For G+C percentages of 0, 1, 2, 3,. . ., 100%, we determined the deviation of coverage from the genome average. Then a simple adjustment was made according to the equation $\tilde{r}_i = r_i \cdot \frac{m}{m_{GC}}$, where r_i are read counts of the *i*th window, m_{GC} is the median read counts of all windows that have the same G+C percentage as the *i*th window, and m is the overall median of all the windows. Our subsequent analysis was carried out on such GC-corrected read counts.

The tasks of parsing alignment files and determining read counts were implemented in the Java programming language on a Linux cluster at Cold Spring Harbor Laboratory. Processing time was ~20 h for the whole genomes of five individuals.

Event detection: Event-wise testing

We use the GC-adjusted RD within 100-bp windows as a quantitative measurement of genome copy number. A deletion or duplication is evident as a decrease or increase in coverage across multiple consecutive windows, as illustrated in Figures 2 and 3. This is perfectly analogous to the detection of CNVs from microarray intensity data. Therefore, events such as these can be detected using the same types of segmentation algorithms that are used for microarray data (Colella et al. 2007; Wang et al. 2007, 2009; Cahan et al. 2008; Korn et al. 2008). However, RD data and microarray CGH data differ in certain characteristics. Most notably, in microarray CGH data, variance in probe ratios is lowest for the "normal" state (two copies) and probe variance increases for copy number changes in both directions. By contrast, in RD data, variance is lowest for "deletion" states (zero or one copy) and variance increases proportionally with increasing copy number (see Bentley et al. 2008 and results presented here in the following sections). Therefore, modification to some of these methods is necessary in order for them to work optimally on RD data.

We have developed a novel CNV-calling algorithm that is designed for the analysis of RD. The event-wise testing (EWT) method is based on significance testing. EWT rapidly searches the entire genome for specific classes of small events that meet criteria of statistical significance, and then clusters of small events are grouped into larger events. Since the number of iterations in EWT is far less than the number of windows (e.g., 19 iterations for all 2.4 million 100-bp windows on chromosome 1), we can perform an exhaustive, fast, and robust search of very large data sets. We evaluated the performance of EWT on simulations constructed from real data and found that it has good statistical power and controls the type-I error well.

The basic idea of our approach is to identify regions of consecutive 100-bp windows with significantly increased or reduced

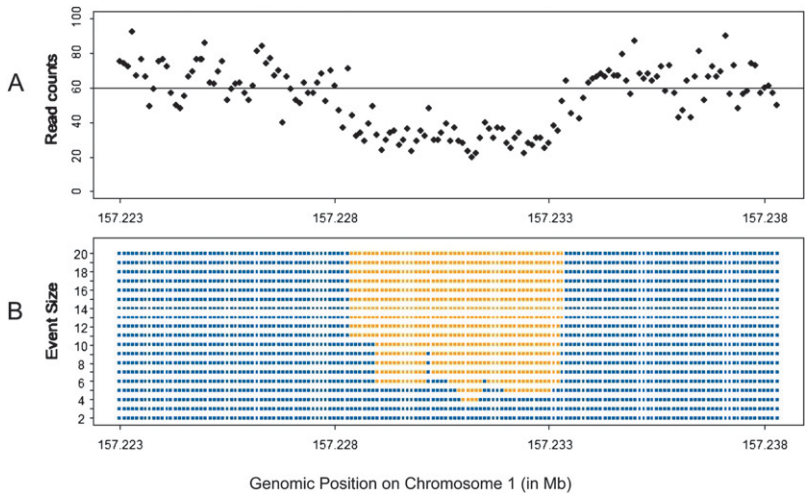

Figure 2. Illustration of the event-wise testing (EWT) method for detecting CNVs based on depth of coverage. Panel *A* illustrates the read depth by 100-bp window for a 15-kb (150 windows) genomic region in sample NA12891, where a 4.9-kb (49 windows) deletion was detected (chr1:157,227,901–157,232,800). The heatmap in *B* illustrates test results for all 100-bp windows of this region for each of the 19 event types (i.e., size 2, 3, 4,. . ., up to size 20) for deletion. The *y*-axis is event size (*l*). An orange dot represents a significant test result for an *l*-sized event, and a blue dot represents a nonsignificant test result.

RD. To detect such an event, we first convert the read count of a window into a *Z*-score z_i by subtracting the mean of all windows and dividing by the standard deviation. The *Z*-score is then converted to its upper-tail probability $p_i^{Upper} = P(Z>z_i)$, and its lower-tail probability $p_i^{Lower} = P(Z<z_i)$. For an interval of consecutive windows *A* with *l* windows, we call it an unusual event if $\max\{p^U{}_i | i \in A\} < \left(\frac{FPR}{L/l}\right)^{\frac{1}{l}}$ for duplications or $\max\{p^L{}_i | i \in A\} < \left(\frac{FPR}{L/l}\right)^{\frac{1}{l}}$ for deletions, where FPR is the nominal false-positive rate (FPR) desired for the entire chromosome (deletion and duplications are treated separately), *L* is the number of windows of a chromosome, and *l* is the size of the interval *A*.

It can be easily seen that if all probes in *A* are from the normal state, the probability of *A* being called an unusual event is less than $\frac{FPR}{L/l}$. Note that instead of using Bonferroni correction to divide the nominal FPR by the total number of intervals of windows, $L-1$, we divide the FPR by the number of nonoverlapping intervals window of size *l*, which is L/l. The former would be overconservative in controlling the type-I error since the statistical tests of overlapping intervals of windows are not independent. The search for deletion events and duplication events are performed separately. For each one, we first search with two-window events. Then iterate the procedure by increasing the size of event by 1. Note that as *l* increases, the cutoff $\left(\frac{(FPR)}{L/l}\right)^{\frac{1}{l}}$ increases as well. We stop the iteration at $N-1$, when $\left(\frac{(FPR)}{L/N}\right)^{\frac{1}{N}}$ exceeds 0.5.

We applied the above procedure to chromosome 1 of five individual genomes using FPR of 0.05.

As is typically the case, additional filtering criteria must be applied to a set of calls made on real data. The additional criteria included the following: First, clusters of small events (within 500 bp) with a copy number change in the same direction were merged. Events with a low absolute difference from the average, that is, a median RD of between 0.75 and 1.25 times the overall mean, were filtered out. Then we tested the significance of each merged event by performing a one-sided *Z*-test. Merged events were filtered using a significance level at 10^{-6}. This threshold corresponds approximately to the significance level for detecting a size 2 event by EWT using a FPR of 0.05. Finally, a threshold of 10^{-6} was deemed to be adequate based on manual inspection of many events at all significance levels. The additional filtering steps that we applied to the merged events substantially increase stringency and reduce sensitivity as a consequence. Thus, we tested the FPRs and false-negative rates of the EWT calls before and after filtering using a set of simulations that are presented below.

Pairwise comparison of RD among five individuals

There is one feature of the MAQ alignment algorithm that is important to point out here. When a single read has multiple exact matches in the genome, it is assigned to a single location randomly. Consequently, coverage across a repetitive or segmentally duplicated region does not differ from the mean if the copy number of those regions in the sample is the same as the copy number of the reference genome. Therefore, the observed events in our data constitute regions of copy number difference between the sample and the reference genome. These events may represent CNVs. They may also represent fixed segmental duplications that are not correctly mapped in the genome, or they may represent a region where the reference genome has a rare allele. Therefore, we must compare the RD of the region in multiple genomes in order to distinguish between events that are clearly polymorphic and those that are not.

As a final step in our pipeline, we conducted a comparison of events between multiple individuals. Many of the deleted or duplicated regions in our filtered call set clearly differed in copy number among the five individuals we examined (see, e.g., Fig. 3). Interestingly, however, the majority showed similarly increased or decreased copy number in all five individuals (which we refer to as "monomorphic" events). Therefore, we sought to distinguish events that were polymorphic from those events that were monomorphic. For each region called by EWT in a given sample, we compare the read counts of 100-bp windows in the region between that sample and each of the other four samples by *t*-tests. CNVs were identified based on the *t*-test *P*-value and the absolute difference between median read counts (D). Events where at least one of four comparisons had *P*-value < 0.001 and D > 0.5 were designated as polymorphic, the remainder, as monomorphic.

Lastly, the copy number of each event in the filtered call set was inferred by rounding the average normalized read counts in each individual to the nearest integer. The normalized read count is defined as 2 × (read count)/(mean read count over the genome).

Simulated data sets for evaluating performance of EWT

To evaluate the performance of EWT, we generated simulated data from chromosomes 1 and X of the male individual NA18507. First, we filtered out all gaps, segmental duplications, telomeres/centromeres, and regions with known CNVs from five publicly available sets of calls, that is, those from Kidd et al. (2008), the

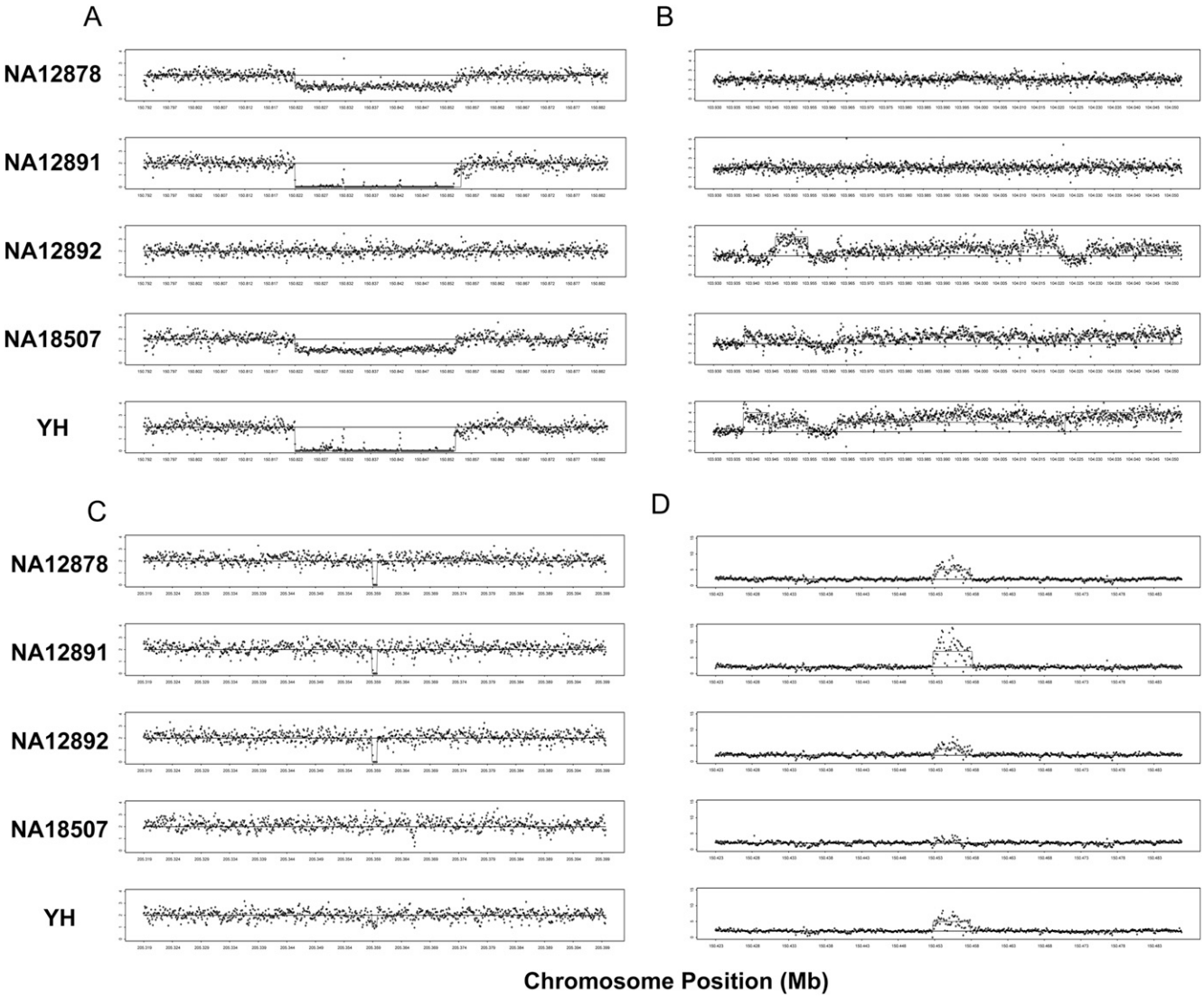

Figure 3. Examples of CNVs detected by analysis of RD. We present four examples of polymorphic gains and losses detected by EWT in five individuals. The *x*-axis represents genomic coordinates (in Mbp) and the *y*-axis represents RD, which is median-normalized to copy number 2. In each panel, plots are for NA12878, NA12891, NA12892, NA18507, and YH from *top* to *bottom*. The coordinates of *A*, *B*, *C*, and *D* are chr1:150,792,101–150,884,101, chr1:103,930,401–104,053,201, chr1:205,319,001–205,399,701, and chr1:150,422,701–150,486,501, respectively.

Database of Genomic Variants (projects.tcag.ca/variation/), Iafrate et al. (2004) and Bentley et al. (2008), the Genome Structural Variation Consortium (GSV), and McCarroll et al. (2008). After filtering, we obtained 1,975,278 100-bp windows from chromosome 1 and 1,238,607 100-bp windows from chromosome X.

To evaluate type-I error, we generated 1000 simulations of normal copy number by randomizing the positions of all windows from chromosome 1, and we made CNV calls by EWT for each replicate.

To evaluate type-II error, we generated deletions by sampling 100-bp windows from chromosome X and implanting them into segments of normal copy number randomly sampled from chromosome 1. For each of 1000 replicates, we selected 100,000 "normal" windows corresponding to 10 Mbp from chromosome 1 and a set of nine "deletion" segments of size 200 bp, 300 bp, 400 bp, 500 bp, 700 bp, 1 kbp, 2.5 kbp, 5 kbp, and 10 kbp sampled from chromosome X. Then we inserted "deletions" into nine fixed positions.

Results

Baseline distribution of read counts

The basic assumption for detecting CNVs using RD analysis is that the reads are randomly sampled with equal probability from any location on the test genome. Under this assumption and an additional assumption that the reads are sampled independently, the count of reads that are mapped into a window of the reference genome follows a Poisson distribution. Note that by the central limit theorem, the distribution should approach normality as coverage of the genome increases, or as the window-size is increased (thus giving larger counts). Bentley et al. (2008) has reported previously that coverage by the Illumina Genome Analyzer platform follows a pattern of Poisson distribution with slight overdispersion. We have observed a similar pattern in all individuals from this study (see Supplemental materials). We confirmed this behavior by directly examining several broad genomic regions

Table 1. Summary of CNVs detected in chromosome 1 in five genomes by analysis of read depth

	NA12878	NA12891	NA12892	NA18507	YH
Polymorpic events (CNVs)	142	132	72	500	304
Deletions	76	75	24	400	180
Duplications	66	57	48	100	124
Combined length (bp)	2,048,700	1,753,300	1,594,200	2,411,100	2,045,300
Validation rate	106/142 (75%)	100/132 (76%)	64/72 (89%)	163/500 (33%)	149/304 (49%)
Monomorphic events	684	534	342	651	1401
Deletions	254	151	85	252	888
Duplications	430	383	257	399	513
Combined length (bp)	3,516,300	3,508,500	3,272,100	3,658,700	4,779,900
Validation rate	235/684 (34%)	209/534 (39%)	138/342 (40%)	236/651 (36%)	327/1401 (23%)

The number of gains and losses detected by EWT in each individual are listed separately for polymorphic events and for monomorphic events. We also list the validation rate of EWT calls in each individual, which is the proportion of EWT calls that overlap (by at least 1 bp) with CNV regions in the GSV validation call set.

with means corresponding to copy numbers of 1, 2, and 3 (for illustrative examples, see Fig. 3), and observed a linear relationship between coverage and copy number. We also observed an overdispersion factor of ~4 (Supplemental Table 1). Furthermore, as expected, counts in 100-bp windows from 30× coverage are sufficiently high that the counts begin to approximate a normal distribution as shown in the Supplemental Figure 1.

We also investigated the relationship between RD and G+C content and observed a nonlinear relationship, where RD is decreased at both extremes. This pattern was observed in all individuals in this study, and it is similar to what was described previously by Bentley et al. (2008). Hence, we corrected for this GC-related effect as described in the Methods.

Type-I and type-II error calculations of EWT on simulated data

We ran an analysis of RD on each replicate in the simulated data, as described in the Methods. Then we evaluated the FPR and the false-negative rate of EWT calls before and after the final merging and filtering steps. In our simulations of type-II error, a deletion was considered to be detected if there was any overlap between the deletion and the segment detected by EWT. We summed the number detected for each event size in 1000 replicates. These results are described in Supplemental Table 1. Results indicate that EWT has good sensitivity to detect CNVs of 1 kb and larger. In the unfiltered call set, 3997/4000 (99.9%) of simulated deletions ≥1000 bp were detected. In the stringently filtered call set, 2934/4000 (73.3%) of deletions were detected in this size range. In our simulations of type-I error, the estimated rate of false-positives was low for CNVs of this size, less than one per individual per chromosome. In the stringently filtered call set, the rate of false-positives was even lower, 0.002 per individual per chromosome. Simulations do not perfectly represent patterns of RD in real data; therefore, a FPR of less than one per chromosome is most likely an underestimate. We attempt to further address the sensitivity and accuracy of our method by validating the EWT calls made in real data, as described in the following section.

CNV detection in genome sequence data from five individuals

We applied the EWT calling method to RD data from five genomes in order to identify regions of copy number difference relative to the reference genome. Figure 2 provides an illustrative example of a 4.9-kb deletion that was detected by EWT. We detected a total of 826, 666, 414, 1151, and 1705 events (including polymorphic and monomorphic events combined) in NA12878, NA12891, NA12892, NA18507, and YH, respectively (Table 1).

Subsequently, we identified the subset of these regions that vary in copy number among the five individual samples (as described in the Methods). Of all events detected, 142 (17.19%), 132 (19.82%), 72 (17.39%), 500 (43.44%), and 304 (17.83%) varied in copy number among five individuals and are referred to as CNVs hereafter. Multiple examples of polymorphic events are shown in Figure 3. Table 1 summarizes the CNV calls in chromosome 1 of five samples by size and copy number. All CNVs detected in this study are listed in Supplemental Table 3, and all monomorphic events detected in this study are listed in Supplemental Table 4.

For validation of EWT calls, we compared these calls to an independent call set consisting of common CNV regions provisionally released by the GSV. The GSV call set consists of CNV regions detected in 40 individuals (20 CEU Caucasian and 20 Yoruban samples) using a NimbleGen tiling array set of 42 million probes (http://projects.tcag.ca/variation/ng42m_cnv.php), and includes 748 CNV regions from chromosome 1. Our CNV calls on each individual were compared to this validation set, allowing for any overlap of 1 bp or greater, and we examined the validation rate (i.e., the number of EWT calls that overlap with GSV/the total number of EWT calls in each individual). A high rate of validation was obtained for EWT calls on the 1000 Genomes Project samples, which were 75%, 76%, and 89% for samples NA12878, NA12891, and NA12892, respectively (Table 1). Validation rates for CNV calls on the two published genomes were lower, which were 33% and 49% for samples NA18507 and YH, respectively. This is most likely due to a higher rate of false-positives in the NA18507 and YH genomes. Analysis of the RD data in these genomes shows a higher level of variance relative to the mean, with variance/mean ratios of 6.1 and 4.6, respectively, compared with 3.5, 3.9, and 3.3 in the CEU samples NA12878, NA12891, and NA12892, respectively (Supplemental Table 1). Other factors that may contribute to the difference in validation rate include the overlap of one CEU individual from our study (NA12878) with the GSV sample; in addition, because the development of the EWT method was done using 1000 Genomes Project data, EWT may perform slightly better in this data set.

As to be expected, the monomorphic events had substantially lower validation rates (23%–40%). This result is consistent with monomorphic events being variants in the reference genome that have low frequencies in the population. Therefore we expect fewer to be present in the GSV sample of 40 individuals. In total, 397/748 (53%) of the GSV calls were detected in the combined set

Table 2. Comparison of nonoverlapping calls from RD-based and PEM based analysis of NA18507

	EWT-specific calls	PEM-specific calls
No. of calls	1051	810
Size (bp)		
Mean	4598	985
Median	1100	414
Segmental duplications		
Count	416 (40%)	17 (2%)
Intersection (bp)	76%	21%
LINE/SINE elements		
Count	493 (47%)	459 (57%)
Intersection (bp)	29%	31%
Simple repeats		
Count	576 (55%)	611 (75%)
Intersection (bp)	6%	12%

The EWT- and PEM-specific calls are compared in terms of size. In addition, we compared them in terms of the content of segmental duplications, LINE elements, SINE elements, and simple repeats. For each element that was tested, we list the total number of events that intersect with at least one element (i.e., "count"), and we list the total fraction of base pairs of the events that intersect with the element (i.e., "intersection").

of EWT calls from five genomes, suggesting that a substantial fraction of common CNVs that are detectable by microarray CGH can be captured using the RD-based approach.

Comparison of CNVs detected using RD and PEM methods

In the initial publications describing the two published genomes included in this study, each performed an analysis of SVs using different PEM approaches (Bentley et al. 2008; Wang et al. 2008). The availability of PEM call sets on these genomes allows us the opportunity to compare the RD-based and PEM-based approaches on the same data. We obtained the PEM call set from Bentley et al. (2008), which was available from the Database of Genomic Variants (projects.tcag.ca/variation/; Iafrate et al. 2004), and we obtained the call set from Wang et al. (2008) from the Beijing Genomic Institute (yh.genomics.org.cn). We sought to determine the proportion of known CNVs detected by the EWT and PEM approaches and to understand the advantages of each approach.

Using the validation data set as a set of true-positives (which reported 748 CNV regions on chromosome 1), we examined the proportion of common CNVs that were detected by each approach in each sample. In samples NA18507 and YH, EWT detected 294 (39%) and 317 (42%) of the CNVs from the validation data set, respectively, and the PEM-based approaches detected 109 (15%) and 49 (7%). Both methods detected a minor fraction of common CNVs, which is not surprising because analysis was limited to only two genomes. In comparison to all PEM-based approaches, EWT captured a greater fraction of regions from the GSV. This result suggests that EWT has good sensitivity to detect CNVs that have been identified previously by microarray CGH. However, this result does not necessarily indicate differential sensitivity of the two methods for all CNVs. For example, there are many small (<1000 bp) CNVs that were detected by PEM and were not detected by EWT or by the GSV 42 million-probe array.

We observed a striking amount of nonoverlap between the PEM- and RD-based calls. For instance 1051/1151 (91%) of the EWT calls and 810/942 (86%) of the PEM calls on sample NA18507 were unique to each set. Likewise, 1638/1705 (96%) of EWT calls and 135/194 (70%) of PEM calls on the YH genome were unique. In order to determine what factors account for this nonoverlap, we examined the nonoverlapping calls in NA18507 in terms of the size of events and their content of segmental duplications and repeats. The median size of PEM-specific events and EWT-specific events was 414 bp and 1100 bp, respectively (Table 2). By examining the intersection of these call sets with annotated segmental duplications, LINE elements, SINE elements, and simple repeats, we found that a much greater fraction (40%) of EWT-specific events overlapped with annotated segmental duplications compared with 2% of PEM-specific events. Conversely, PEM-specific events showed a greater enrichment of simple repeats, which accounted for 12% of the total base pairs of PEM-specific events compared with 6% of the total base pairs of EWT-specific events. These results suggest that PEM-based and RD-based approaches have unique advantages in detecting different classes of SVs.

Discussion

Here we describe a novel computational approach for the detection of CNVs from next-generation sequence data. We carried out a systematic analysis of depth of coverage on chromosome 1 in five individuals. Based on simulations our CNV calling algorithm, EWT, is able to capture 99.9% of deletions >1000 bp in size. After applying additional filtering criteria to CNV calls, our method is able to capture 73% of deletions >1000 bp in size. We applied our method to the detection of CNVs on chromosome 1 from 30× coverage data on five individuals. On one chromosome, we detected an average of 952 events per individual, of which ~230 (24%) were copy number variable in the five individuals in this study.

A majority of events detected in this study were monomorphic in five individuals, suggesting that the structural allele represented in the reference genome differs from most chromosomes in the population. Similar results have been observed in earlier studies that have looked at PEM calls in multiple genomes (Kidd et al. 2008). Greater than 30% of the monomorphic events overlapped with CNV regions identified by the GSV; therefore, many of these events are simply CNVs with lower frequencies. Other monomorphic events may correspond to fixed segmental duplications in the genome that have not been previously identified or regions of the reference genome that have been misassembled. Thus, in addition to CNVs that we identify by EWT, the "nonvariants" are also an important subject of interest. Additional efforts are needed to map the genomic structure of the major alleles at these loci.

A RD-based approach has distinct advantages over other approaches in detecting certain classes of SVs. In comparing our EWT calls and PEM calls made by others on a previously published genome (Bentley et al. 2008), both methods identified a similar number of events. However, only a minority of the calls overlapped between the two methods. Compared with the PEM-specific events, the EWT-specific events were greatly enriched in segmental duplications. This is to be expected. Complex regions rich in segmental duplications are more difficult to ascertain using PEM because many of the reads in these regions do not map to unique locations in the genome (Tuzun et al. 2005). Segmental duplications are less of a confounder for RD analysis because it is not critical for reads to map uniquely to a region in order to estimate the sequence coverage of that region.

While RD analysis overcomes some of the limitations of other methods, it has limitations of its own. RD analysis is not able to ascertain balanced rearrangements. In addition, ascertainment of SVs that involve highly repetitive sequences is limited. RD analysis cannot determine the precise location of an insertion, nor can it find novel insertions that are not already in the reference genome. These classes of SVs are more easily detectable using a PEM-based approach. Given the relative strengths of PEM and EWT, the two methods are quite complementary. Using the two approaches in combination will enhance the detection of a variety of SVs from next-generation sequence data.

Knowledge of structural variation in the human genome will improve rapidly as many more complete genome sequences become available through the 1000 Genomes Project and related efforts. The public availability of these data sets will further enable the development of new bioinformatic tools. Ultimately, methods for SV detection such as RD analysis and PEM will be combined with methods for local de novo assembly in order to resolve the structure of SVs at the nucleotide sequence level.

Acknowledgments

This work was supported by NIH grants HG00422 (J.S.) and MH76431 (J.S.), which reflect co-funding from Autism Speaks and the Southwest Autism Research & Resource Center. Additional support for J.S. was provided by Stanley Medical Research Institute and the Simons Foundation. We thank the 1000 Genomes Consortium, which provided the data used in this study (http://www.1000genomes.org), and the Genome Structural Variation Consortium for making their data publicly available in advance of publication (http://www.sanger.ac.uk/humgen/cnv/42mio/), and the 1000 Genomes Project analysis group for many helpful discussions. We thank Dheeraj Malhotra, Shane McCarthy, Mary Kusenda, and Abhishek Bhandari for technical assistance.

References

Albertson DG, Pinkel D. 2003. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12** (Spec. no. 2)**:** R145–R152.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Cahan P, Godfrey LE, Eis PS, Richmond TA, Selzer RR, Brent M, McLeod HL, Ley TJ, Graubert TA. 2008. wuHMM: A robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Res* **36:** e41. doi: 10.1093/nar/gkn110.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. 2007. QuantiSNP: An objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35:** 2013–2025.

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

Feuk L, Carson AR, Scherer SW. 2006a. Structural variation in the human genome. *Nat Rev Genet* **7:** 85–97.

Feuk L, Marshall CR, Wintle RF, Scherer SW. 2006b. Structural variants: Changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* **15** (Suppl. 1)**:** R57–R66.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949–951.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, 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.

Kirov G, Gumus D, Chen W, Norton N, Georgieva L, Sari M, O'Donovan MC, Erdogan F, Owen MJ, Ropers HH, et al. 2008. Comparative genome hybridization suggests a role for *NRXN1* and *APBA2* in schizophrenia. *Hum Mol Genet* **17:** 458–465.

Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P, Craddock N, Owen MJ, O'Donovan MC. 2009. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* **18:** 1497–1503.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40:** 1253–1260.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

Lucito R, West J, Reiner A, Alexander J, Esposito D, Mishra B, Powers S, Norton L, Wigler M. 2000. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res* **10:** 1726–1736.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24:** 133–141.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, et al. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82:** 477–488.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker 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.

Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* **359:** 1685–1699.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci* **99:** 12963–12968.

Rujescu D, Ingason A, Cichon S, Pietilainen OP, Barnes MR, Toulopoulou T, Picchioni M, Vassos E, Ettinger U, Bramon E, et al. 2008. Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* **18:** 988–996.

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.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445–449.

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* **455:** 178–179.

Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DH, Corvin A, Craddock NJ, Gill M, Hultman CM, Lichtenstein P, et al. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455:** 237–241.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727–732.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320:** 539–543.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17:** 1665–1674.

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.

Wang LY, Abyzov A, Korbel JO, Snyder M, Gerstein M. 2009. MSB: A mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res* **19:** 106–117.

Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, Greenwood T, Nievergelt C, Barrett TB, McKinney R, Schork N, et al. 2008. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry* **14:** 376–380.

Received February 19, 2009; accepted in revised form July 15, 2009.