Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (B)

Identification of Medium-Sized Copy Number Alterations in Whole-Genome Sequencing

Hatice Gulcin Ozer[1], Aisulu Usubalieva[1], Adrienne Dorrance[2], Ayse Selen Yilmaz[1], Michael Caligiuri[2], Guido Marcucci[2] and Kun Huang[1]

[1]Department of Biomedical Informatics, [2]Division of Hematology, Department of Medicine, The Ohio State University, Columbus, OH, USA.

ABSTRACT: The genome-wide discoveries such as detection of copy number alterations (CNA) from high-throughput whole-genome sequencing data enabled new developments in personalized medicine. The CNAs have been reported to be associated with various diseases and cancers including acute myeloid leukemia. However, there are multiple challenges to the use of current CNA detection tools that lead to high false-positive rates and thus impede widespread use of such tools in cancer research. In this paper, we discuss these issues and propose possible solutions. First, since the entire genome cannot be mapped due to some regions lacking sequence uniqueness, current methods cannot be appropriately adjusted to handle these regions in the analyses. Thus, detection of medium-sized CNAs is also being directly affected by these mappability problems. The requirement for matching control samples is also an important limitation because acquiring matching controls might not be possible or might not be cost efficient. Here we present an approach that addresses these issues and detects medium-sized CNAs in cancer genomes by (1) masking unmappable regions during the initial CNA detection phase, (2) using pool of a few normal samples as control, and (3) employing median filtering to adjust CNA ratios to its surrounding coverage and eliminate false positives.

KEYWORDS: copy number alteration (CNA), whole-genome sequencing, acute myeloid leukemia (AML)

SUPPLEMENT: Computational Advances in Cancer Informatics (B)

CITATION: Ozer et al. Identification of Medium-Sized Copy Number Alterations in Whole-Genome Sequencing. *Cancer Informatics* 2014:13(S3) 105–111 doi: 10.4137/CIN.S14023.

RECEIVED: June 04, 2014. **RESUBMITTED:** December 29, 2014. **ACCEPTED FOR PUBLICATION:** January 04, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Methodology

FUNDING: This work was funded by U.S. National Cancer Institute through Leukemia SPORE (P50-CA140158) and Cancer Center Support Grant (P30-CA016058). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ozer.9@osu.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE). Published by Libertas Academica. Learn more about this journal.

Introduction

Recent advances in next-generation sequencing technologies facilitated genome-wide discoveries such as variant analysis, expression quantification, and copy number alteration (CNA) analysis. CNAs are variations in the genome that result in either gain or loss of one or more copies of the DNA segment. The CNAs range from 1 kilobase (kb) to several megabases and are one of the essential constituents of genomic diversity.[1] In humans, CNAs have been reported to account for approximately 12% of genomic DNA.[2] While some CNAs do not have any observable effects on phenotype, some have been linked to diseases such as autism,[3] to susceptibility to HIV,[4] and to cancers such as non–small-cell lung cancer[5] and acute myeloid leukemia.[6]

Wide ranges of computational approaches have been developed to identify CNA events in whole-genome sequencing data. As previously described by Liu et al, there are three common steps in these algorithms: data preprocessing, data segmentation, and data interpretation.[7] Data preprocessing starts with normalization of read depths or counts that are considered the most common input. Then, log2 ratios of those counts are calculated and compared to a selected reference value, which typically is a matched control. Tools such as SegSeq,[8] ReadDepth,[9] HMMCopy,[10] Bayesian Information Criterion sequencing (BIC-seq),[11] Patchwork,[12] VarScan2,[13] Control-FREEC[14] use this approach. Some of the algorithms also include steps to handle systematic biases such as genomic

mappability (Control-FREEC, ReadDepth, and HMMCopy) and GC-content (Control-FREEC, Patchwork, and HMMCopy). Moreover, some of them also incorporate B allele frequency information to improve CNA detection (Patchwork).

The second step of the CNA detection algorithm is segmentation. In this step, continuous regions with similar copy numbers are combined and the CNA profile is smoothened. Circular binary segmentation (CBS) and hidden Markov model (HMM) are the most commonly adopted algorithms to implement this procedure. In addition, the HMM-based approaches simultaneously assign copy number status to each region during this segmentation step.

The final step in the CNA detection algorithms is interpretation of data in order to determine the copy number state. Typically, an empirical cutoff is applied on each segment to identify copy number changes (Control-FREEC, BIc-seq, VarScan2). Some algorithms optimize these cutoff values to a desired sensitivity and specificity (Patchwork, SegSeq).

Once the next-generation high-throughput sequencing experiments generate millions of short (36 bp–100 bp) sequence reads, mapping tools such as BWA,[15] Bowtie,[16] or SOAP2[17] align those reads to the reference genome. The uniqueness of the reference genome sequence plays a major role during this alignment stage and is one of the limitations of current CNA detections tools. Lack of sequence uniqueness (mappability) might lead to low complexity regions on the genome and therefore even the best mapping tools cannot align all reads, despite using the highest quality sequence reads. Therefore, with the technology in hand, it is not possible to thoroughly sequence the entire genome. For instance, only 86% of the human genome can be securely mapped by using 100-bp sequence reads (Table 1).

Another limitation of the current CNA detection tools is dependence on control samples. The current standard in a typical copy number analysis is to compare the tumor (case) sample with the matching normal sample from the same individual. However, in some cases, this might not be possible due to technical problems, availability of samples, or cost.

The third issue with the current CNA detection tools is the size of the detected CNAs. Typically, these tools detect alterations of around or larger than 100 kB with confidence. By tweaking parameters, one can detect shorter CNAs. However, several hundred to thousand candidate CNAs would be generated, suggesting very high false-positive rates.

We believe that it is important to understand these characteristics and limitations of the genome as well as CNAs detection tools and address them when developing new methods. In this paper, we describe a simple yet effective method that addresses these issues. Our method detects medium-sized CNAs from whole-genome sequencing data in the absence of a matching control sample by using a pool of normal samples as a baseline. It also effectively handles mappability issues in the genome. In addition, our method employs median filtering to evaluate shape of the coverage around the candidate CNAs and effectively eliminates false positives.

Methods

Genome mappability. Mappability tracks can be generated computationally based on the level of sequence uniqueness of the reference genomes. As the sequence reads get longer, the mappability increases significantly.[18] The University of California Santa Cruz (UCSC) Genome Browser provides mappability tracks from ENCODE project for different sequence read lengths (eg, 100 mer, 75 mers, 50 mers). These tracks are generated by GEM (GEnome Multitool) mappability tool and regions are scored from 0 to 1 based on the sequence uniqueness, and therefore, the higher the mappability score, the more unique is the mapping position. We downloaded mappability tracks for mouse (mm9) reference genome for sequence reads of 100, 75, 50, 40, and 36 bp. Then, we combined the close by regions that had mappability scores of less than 1 and defined them as unmappable regions of the genome. We observed that about 28%, 26%, 23%, 18%, and 16% of the mouse genome cannot be mapped with 36, 40, 50, 75, and 100-bp sequence reads, respectively. Size of unmappable regions could be as short as a few base pairs or as long as several million base pairs. For example, median width of unmappable regions with 75-bp sequence reads is 21 bp and 81% of its unmappable regions are shorter than 100 bp. To better understand the size of unmappable regions across genome, we extended unmappable regions by 1 kb, if they span more than so

Table 1. Percentage of unmappable regions of the mouse (mm9) and human (hg19) reference genomes.

SEQUENCE READ LENGTH	MOUSE REFERENCE GENOME		HUMAN REFERENCE GENOME	
	GEM SCORE <1	UNMAPPABLE	GEM SCORE <1	UNMAPPABLE
36 mer	28%	55%	29%	64%
40 mer	26%	52%	27%	60%
50 mer	23%	46%	23%	51%
75 mer	18%	33%	16%	29%
100 mer	16%	27%	14%	20%

Notes: Second and fourth columns show the percentages of the mouse and human genome with a GEM score less than 1 (this means the region is not unique). If an unmappable region covered more than 100 bp, we extended such region by 1 kb and consolidated regions. This aligns with our 90% cutoff on mappability percentage. Total size of these extended regions gives a more realistic idea about overall unmappability of the genome. We report percentages of unmappable regions in the third and fifth columns for mouse and human genomes, respectively.

that bordering large unmappable regions will be consolidated. This approach suggests that practically about 55%, 52%, 46%, 33%, 27% of the mouse genome is unmappable with 36, 40, 50, 75, and 100-bp sequence reads, respectively (Table 1).

Whole-genome sequencing data. All animal studies were performed under approved protocols following the Ohio State University Institutional Animal Care and Use Committee. We analyzed bone marrow mononuclear cells of 11 samples obtained from two healthy wild-type control mice and nine transgenic mice with *MLL*-partial tandem duplication (PTD) deficiencies that have been associated with acute myeloid leukemia.[19] Three of the transgenic mice were *Mll* (PTD/wt): *Flt3* (ITD/ITD) double knock-in mouse, three were nonleukemic *Mll* (PTD/wt) mouse, and the remaining three were *Flt3* (ITD/wt) single-knock-in mouse (unpublished data). All transgenic mice were genotyped and six mice were validated to have PTD in *Mll1* gene on chromosome 9.

The whole-genome sequencing of all mouse samples was performed by the Beijing Genomics Institute using Illumina HiSeq 2000 platform. None of the samples had cytogenetic abnormalities. The sequence reads were mapped to mouse (mm9) reference genome by BWA.[15] The mapping quality filter was applied to remove potential optical and polymerase chain reaction duplicate reads, nonspecific reads, and improper read pairs. On average, 638 million, 90-bp paired-end sequence reads per sample were aligned to mm9 reference genome providing 43x average coverage (41x minimum, 44x maximum).

Detection of CNAs. In this study, we were specifically interested in identifying medium-sized (~1 kb–30 kb) CNA events. We started with computing the total number of sequence reads per kilobase of genome (read counts) for each sample using BEDTools.[20] Next, these read counts were normalized to the total number of sequence reads per sample. Average genomic coverage was quite similar across all samples, thus, employed. Genomic coverage is not reliable around unmappable regions. Therefore, we eliminated 1-kb intervals with unmappable bases. Since our data consisted of 90-bp paired-end reads with slight quality drop toward the end of the reads, we chose to use mappability tracks with 75-bp reads. Then, we generated pairwise log2 ratios between case and control samples using normalized read counts for the remaining 1-kb intervals.

Following the calculation of log2 ratios of the coverage between case and control samples, we used R DNAcopy library from Bioconductor[21] to combine intervals into segments. This library initially implemented for the analysis of array-based DNA copy number data using a CBS algorithm.[21] This algorithm consolidates neighboring intervals with similar ratios into segments and reports mean ratio for the segment and total number of intervals that support this outcome. First, we applied a cutoff on the mean ratio to account for 25% loss or gain.

Second, since we eliminated intervals with unmappable bases at the beginning, identified regions might span through unmappable regions. Therefore, we re-evaluated mappability of the identified regions and their surroundings (including w upstream and w downstream, where w is the width of the candidate CNA) as a whole and kept the regions with 90% or more mappable bases.

Finally, we applied one-dimensional median filter on pairwise ratios and calculated adjusted ratio as a mean of differences before and after the median filtering. Median filtering is a nonlinear filtering technique generally used to reduce noise in a signal. For the log2 ratio data over a region, $r = r_1 r_2 \ldots r_n$, median filtering replaces each r_i with the median of $[r_i - d/2, \ldots, r_i, \ldots, r_i + d/2]$, where $1 < i < n$ and $d + 1$ is the sliding window size. At this step, we zoomed in the data and calculated log2 ratios at 100-bp resolution for CNA candidates. Then, median filtering was applied around each candidate CNA staring from $4w$ upstream to $4w$ downstream, with a sliding window size of $3w$, where w is the width of the candidate CNA. Average difference between original log2 ratios and median filtered log2 ratios over each candidate region is reported as median filter adjusted ratio. Finally, CNAs with 25% loss or gain after median filter adjustment and common to both WT comparisons were reported as final CNAs.

Figure 1 depicts the complete CNA identification workflow.

Results

Since it is not possible to thoroughly and accurately sequence the entire genome with short sequence reads generated by current high-throughput sequencing technology, the genome mappability was the first issue we aimed to address in this paper. We analyzed coverage tracks of 11 mouse samples. Figure 2 shows approximately 5 million base segment of their chromosome 12. A large portion of this region (about 3 million bases) is mostly unmappable (GEM score <1) with 75-bp sequence reads. The coverage is not uniform, but rather fluctuates with the mappability. This region also contains segmental duplications (duplications of >1000 bases that are not masked by RepeatMasker). It is not possible to identify medium-sized CNAs within this kind of regions. As shown in Table 1, approximately 33% of mouse genome is unmappable with 75-bp reads. Our approach handles this issue by eliminating intervals with low mappability at the beginning of the analysis.

Figure 3 shows a 3000-bp candidate CNA on chromosome 1. Intervals with unmappable bases (marked with green) were eliminated prior to running the CBS algorithm on pairwise log2 ratios between PTD/ITD samples versus WT samples. Mean log2 ratio between the first PTD/ITD sample and the first WT sample over this 3000-bp region was −0.79, suggesting 75% loss and 94% mappable bases including the surrounding region (3000 bp upstream and 3000 bp downstream). After applying median filtering on log2 ratios in 100-bp resolution, average difference between original ratios and median filtered ratios within this region was −0.89, reporting 85% loss in this region. This 3000-bp CNA made the final list since it was reported in comparisons to both WT samples with more than 25% loss.

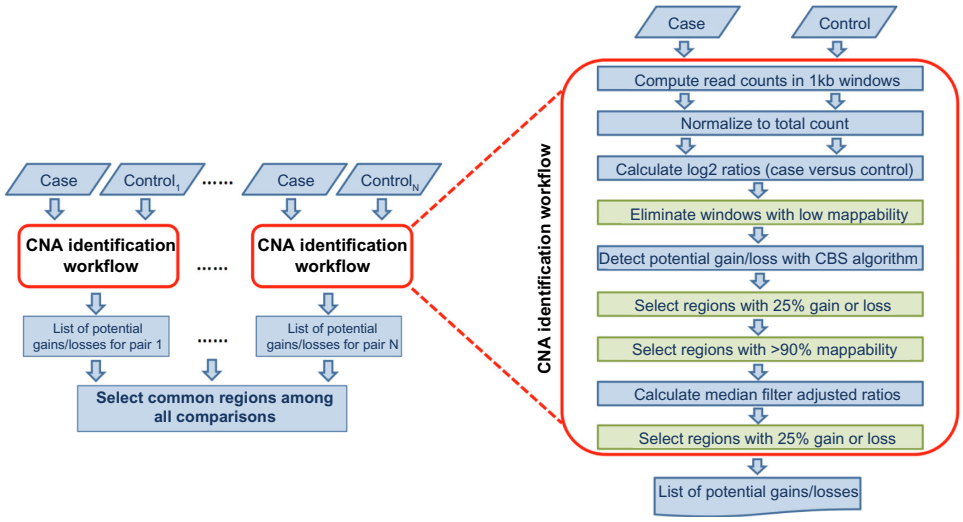

Figure 1. Complete CNA identification workflow.

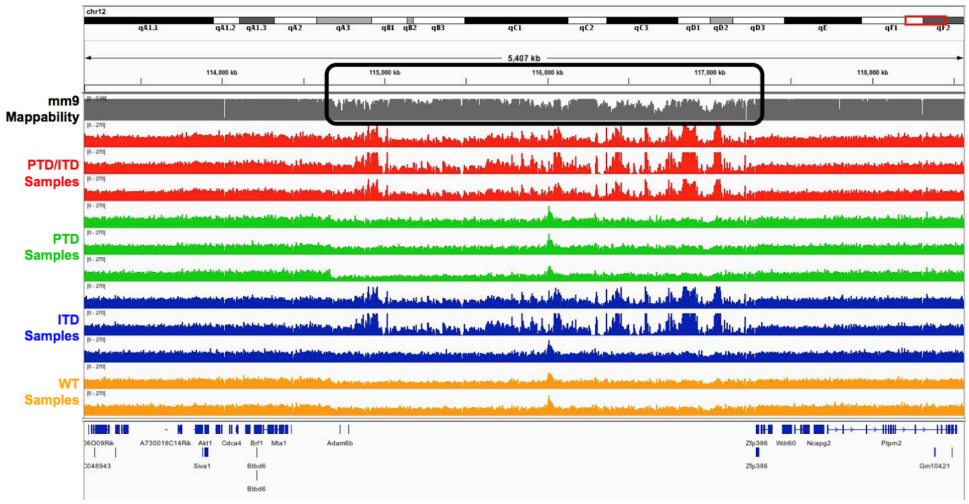

Figure 2. An example of highly unmappable region from chromosome 12. The selected region contains about 3 million bases. It is highly unmappable and also contains segmental duplications. The dark gray track shows mappability of mouse genome (mm9) with 75-bp reads. The red tracks show coverage of 3 PTD/ITD (*Mll* [PTD/wt]: *Flt3* [ITD/ITD] double knock-in) mouse samples. The green tracks show coverage of 3 PTD (nonleukemic *Mll* [PTD/wt]) mouse samples, while the blue tracks show the coverage of ITD (*Flt3* [ITD/wt] single knock-in) mouse samples.

The transgenic mice samples evaluated in this study were genotyped and six mice were experimentally validated to have PTD in *Mll1* gene. Sequencing showed the presence of multiple copies in the *Mll1* gene and therefore the alignment of sequence reads to the reference genome resulted in a larger number of reads being mapped to this region. This led to larger coverage in this duplicated region compared to the rest of the gene. This phenomenon was used as a positive control when testing our method. Figure 4 shows a slight enrichment (~30% gain) around PTD region in PTD/ITD and PTD samples coverage tracks around *Mll1* gene. Our method was able to successfully detect this enrichment, because our approach allows application of low cutoffs with confidence, while median filter adjustment allows easy distinction between true positives and false positives.

Figure 5 shows a 3000-bp candidate CNA on chromosome 1, with a mean log2 ratio of −0.57, suggesting 49% loss in that region. Median filter-adjusted ratio would be −0.20, suggesting only 15% loss. PTD/ITD 2 seems to have a slightly lower coverage across this region compared to WT 2 sample. However, this is only a slight fluctuation in coverage and median filter adjustment easily picks up this kind of fluctuations and eliminates such false positives. Although, 25% loss or gain cutoff is a very loose cutoff and results in a lot of false positives, median filter adjustment helps us to eliminate these false positives effectively and allows us to report loss or gain events confidently.

When evaluating coverage tracks of PTD/ITD mouse samples following the elimination of intervals with low

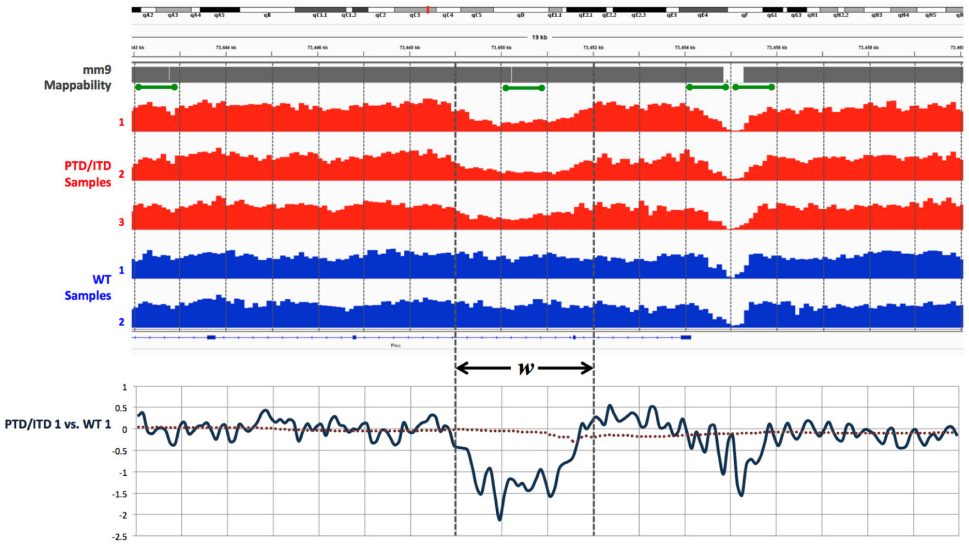

Figure 3. At the top: Integrated Genomic Viewer screenshot of a region in chromosome 1 with 3-kb candidate CNA. The dark gray track shows mappability of mouse genome (mm9) with 75-bp reads. The red and blue tracks show coverage of 3 PTD/ITD (*Mll* [PTD/wt]: *Flt3* [ITD/ITD] double knock-in) and 2 WT mouse samples in 100-bp resolution, respectively. Green intervals show 1-kb regions with unmappable bases. At the bottom: Log2 ratio for the first PTD/ITD sample versus the first WT sample (black solid line) and its median filtering (red dotted line) in 100-bp resolution across the region. Size of the detected CNA is denoted with *w*.

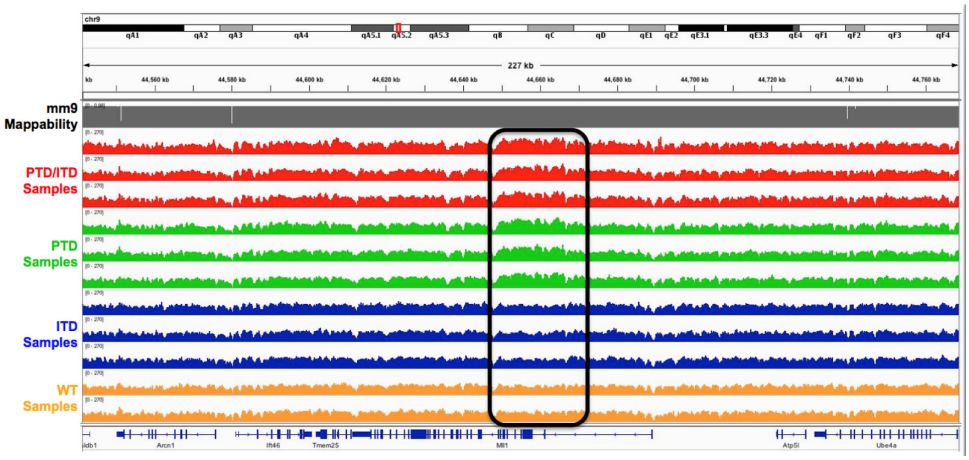

Figure 4. The region of the chromosome 9 that has *Mll1* gene. A slight enrichment is observed around PTD region in the coverage tracks of PTD/ITD and PTD samples.

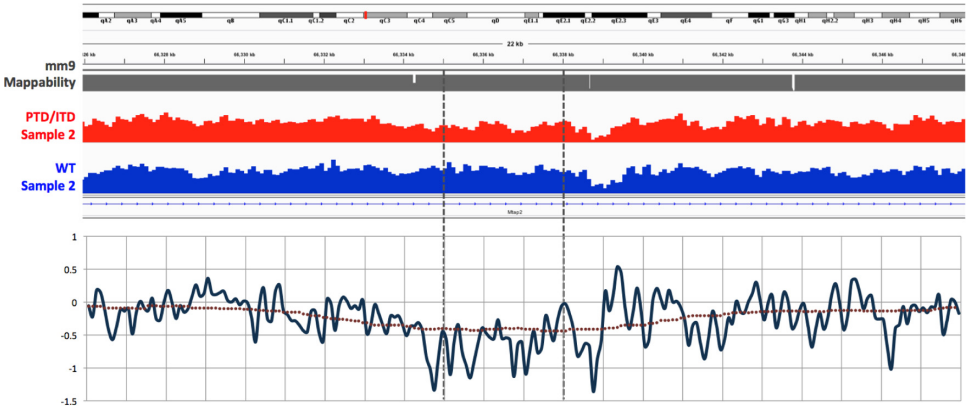

Figure 5. A false-positive example. Genomic coverage, original log2 ratios, and median filtered log2 ratios are depicted for a 3-kb candidate CNA on chromosome 1. Slight genomic coverage change originally reported as a candidate CNA suggesting 49% loss. Genomic coverage upstream and downstream of the candidate was taken into account by median filter adjustment, and adjusted ratios suggest only 15% loss.

mappability, CBS algorithm detected about 292 CNA candidates per PTD/ITD sample compared to WT. Then, 90% mappability filter eliminated 22% of these candidates. Median filter adjustment eliminated another 34%. Finally, selection of CNAs common to both control comparisons resulted in 16% of the initially reported candidates, which is 47 CNAs per PTD/ITD sample with a median size of 3000 bases (average size of 10,300 bases). Being common to both comparisons seems to be the most stringent filter. However, if we had applied that filter without median filter adjustment, we would end up with 30% of the initially reported candidates, which is 88 CNAs per PTD/ITD sample. Therefore, median filter adjustment is clearly a crucial step in this approach, as we would have reported twice as much CNA candidates without this adjustment.

Discussion

The next-generation high-throughput sequencing is the future of personalized medicine. The Next Generation Sequencing (NGS) data allow us to investigate genome-wide discoveries such as single-nucleotide polymorphisms, expression data, DNA's structural variations, and CNAs. Designing tools that facilitate efficient detection of those discoveries is the next step that brings us closer to personalized medicine. Currently, tools designed to detect CNA face challenges such as a lack of sequence uniqueness in the genome, dependence on matching control samples for comparison and inability to detect smaller size CNAs. In this paper, we propose a simple yet effective approach that takes into account these limitations of current technology and specifically targets to identify medium-sized (1–30 kb) CNAs.

First, our method addresses the issue of mappability by eliminating regions of the genome with low mappability at the beginning of the analysis. This resulted in elimination of about 33% of the mouse genome, leaving us with the remaining 67% for analyses. Although mappability-associated fluctuations in the coverage smoothen out and thus do not pose difficulties when identifying large aberrations (>500 kb) in the genome, such fluctuations become the signal when detecting the medium-sized CNAs. Therefore, unmappable regions of the genome should be masked, for the sake of avoiding high false-positive rates.

When testing for CNAs, the usual approach is to compare the tumor sample with the control sample from the same individual. In some cases, the matched control sample is not available. Similarly, in our study, we tested 11 samples, 9 of which were from transgenic mice, and therefore these samples did not have their matched controls. So, in accordance with previous studies,[22] we used the samples from the control mice that represented a pool of control samples.

Median filter adjustment of the ratios is the novel part of our method. Median filtering is applied around each candidate CNA and window size is directly determined by CNA size. Thus, smoothing around candidate CNA is performed at an appropriate scale. The difference between the median filtered ratios and original ratios suggests that candidate CNA indeed has a different coverage profile compared to its surrounding and it can be reported confidently. If median filtered ratios are not much different than the original ratios, adjusted ratios would suggest much smaller gain or loss, thus a false positive. Median filtering was used by Lee et al.[23] with a constant window size to smooth the coverage signal as part of the preprocessing before CNA detection. We employ median filtering with variable window size to adjust copy number change rate and CNA ratio.

We also tested the BIC-seq algorithm, which is one of the well-established tools for the detection of CNAs from whole-genome sequencing data,[11] with our data. The BIC-seq is a nonparametric model that segments the genome into small bins and iteratively combines adjacent bins with the same copy number. Due to its nonparametric nature, BIC-seq is robust in the identification of outliers. However, one of its limitations is that it relies on comparison of case samples with matched control samples. This limitation confines the utilization of this algorithm to analyses that only have matched control samples. Despite not having the matched control in our samples, we still tested this algorithm using our WT samples as the control. We ran the BIC-seq algorithm using PTD/ITD samples as case and each WT as a control. The BIC-seq has two main adjustable parameters: starting bin and lambda, which is used for tuning the smoothness of the CNV profile. The average number of detected CNAs decreases with increasing bin size and lambda; however, the average size of CNAs increases. We tested three combinations of bin sizes and lambda values and observed that in all the cases only about 5.5% of CNAs were coming from securely mappable regions, while the rest was from unmappable regions of the genome (Table 2). In comparison to our method, which detected 159 CNAs from 40% of securely mappable regions of the genome in all PTD/ITD samples, BIC-seq falls short in identifying CNAs from

Table 2. Combinations of various bin size and lambda parameters used in the detection of CNAs with BIC-seq toolBIC-seq parameters.

		AVERAGE # OF CNAs	% ON COMPLETELY MAPPABLE REGIONS	AVERAGE SIZE OF CNAs
Bin size = 1000	Lambda = 10	283	4.6%	19,363
Bin size = 100	Lambda = 10	330	5.3%	7,789
Bin size = 100	Lambda = 4	928	6.6%	3,714

Notes: On average only about 5.5% of CNAs were coming from securely mappable regions, while the rest was from unmappable regions of the genome.

mappable regions, while reporting high rate of CNAs, most of which might be false positives.

Another method we tested was Control-FREEC, which automatically calculates the copy number profiles from the NGS data.[14] One advantage of Control-FREEC is that the use of matched normal samples is optional. This approach also allows exclusion of regions with low mappability from the analysis by using provided mappability tracks. However, unmappable regions are not eliminated entirely but are rather skipped during the evaluation. In other words, Control-FREEC algorithm considers an unmappable region as a gap. When we tested our samples with the Control-FREEC method, it detected approximately 160 CNAs per sample with an average CNA size of 500 kb. Only about 4% of those CNAs were in completely mappable areas. Given the large sizes of the detected CNAs, it was not a surprise that Control-FREEC identifies CNAs in unmappable regions. This algorithm identified fewer CNAs from mappable regions compared to both the BIC-seq and our method.

Most of the CNA detection tools allow adjustments in the size of the detected CNA by modifying some parameters. However, one of the most common problems is that as the CNA size gets smaller, the number of detected CNAs gets larger. Unfortunately, most of these detected CNAs are false positives, because these approaches cannot manage fluctuation in coverage when the bin size gets smaller. In the larger bin sizes, these fluctuations are smoothened out, and large CNAs can be securely reported. Alas, there is no reliable and effective algorithm for detection of "medium-sized CNA" that takes these problems into consideration.

In this paper, we discussed potential issues in medium-sized CNA detection using whole-genome sequencing data and why existing approaches are not suitable for the detection of medium-sized CNAs. We tested an approach to demonstrate that by addressing underlying issues, we can effectively identify medium-sized CNAs.

We understand that our approach requires further modifications; nevertheless, the fact that our method successfully detects CNAs from mappable regions of the genome is a proof of concept. This approach can be improved by implementing a sliding window approach to better identify CNA start and end sites (rather than bin start and end). We will also work on a methodology to eliminate false positives due to nonuniform coverage in control samples.

Author Contributions

Conceived the concepts: HGO, KH. Provided biological data: GM, MC. Analyzed the data: HGO. Wrote the first draft of the manuscript: HGO, AU. Contributed to the writing of the manuscript: HGO, AU, ASY, AD, MC, GM, KH. Agree with manuscript results and conclusions: HGO, AU, AD, ASY, MC, GM, KH. Jointly developed the structure and arguments for the paper: HGO, AU, KH. Made critical revisions and approved final version: HGO, AU, ASY, KH. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437–55.
2. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–54.
3. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature*. 2008;455(7215):919–23.
4. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005; 307(5714):1434–40.
5. Cappuzzo F, Hirsch FR, Rossi E, et al. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J Natl Cancer Inst*. 2005;97(9):643–55.
6. Jacoby MA, Walter MJ. Detection of copy number alterations in acute myeloid leukemia and myelodysplastic syndromes. *Expert Rev Mol Diagn*. 2012;12(3):253–64.
7. Liu B, Morrison CD, Johnson CS, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*. 2013;4(11):1868–81.
8. Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*. 2009;6(1):99–103.
9. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*. 2011;6(1):e16327.
10. Ha G, Roth A, Lai D, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res*. 2012;22(10):1995–2007.
11. Xi R, Hadjipanayis AG, Luquette LJ, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A*. 2011;108(46):E1128–36.
12. Mayrhofer M, Dilorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol*. 2013;14(3):R24.
13. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568–76.
14. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423–5.
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
16. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
17. Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966–7.
18. Derrien T, Estellé J, Marco Sola S, et al. Fast computation and applications of genome mappability. *PLoS One*. 2012;7(1):e30377.
19. Zorko NA, Bernot KM, Whitman SP, et al. Mll partial tandem duplication and Flt3 internal tandem duplication in a double knock-in mouse recapitulates features of counterpart human acute myeloid leukemias. *Blood*. 2012;120(5):1130–6.
20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
21. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557–72.
22. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: exomeCNV. *Bioinformatics*. 2011;27:2648–54.
23. Lee J, Lee U, Kim B, Yoon J. A computational method for detecting copy number variations using scale-space filtering. *BMC Bioinformatics*. 2013;14:57.