Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2015 ; 12(4): 799–806. doi:10.1109/TCBB.2014.2359450.

MMBIRFinder: A Tool to Detect Microhomology-Mediated Break-Induced Replication

Matthew W. Segar,
Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis 46202, IN. msegar@iupui.edu

Cynthia J. Sakofsky,
Department of Biology, University of Iowa, Iowa City 55242-1324, IA

Anna Malkova, and
Department of Biology, University of Iowa, Iowa City 55242-1324, IA

Yunlong Liu
Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis 46202, IN

Abstract

The introduction of next-generation sequencing technologies has radically changed the way we view structural genetic events. Microhomology-mediated break-induced replication (MMBIR) is just one of the many mechanisms that can cause genomic destabilization that may lead to cancer. Although the mechanism for MMBIR remains unclear, it has been shown that MMBIR is typically associated with template-switching events. Currently, to our knowledge, there is no existing bioinformatics tool to detect these template-switching events. We have developed MMBIRFinder, a method that detects template-switching events associated with MMBIR from whole-genome sequenced data. MMBIRFinder uses a half-read alignment approach to identify potential regions of interest. Clustering of these potential regions helps narrow the search space to regions with strong evidence. Subsequent local alignments identify the template-switching events with single-nucleotide accuracy. Using simulated data, MMBIRFinder identified 83 percent of the MMBIR regions within a five nucleotide tolerance. Using real data, MMBIRFinder identified 16 MMBIR regions on a normal breast tissue data sample and 51 MMBIR regions on a triple-negative breast cancer tumor sample resulting in detection of 37 novel template-switching events. Finally, we identified template-switching events residing in the promoter region of seven genes that have been implicated in breast cancer. The program is freely available for download at https://github.com/msegar/MMBIRFinder.

Index Terms

Biology and genetics; life and medical sciences

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

1 Introduction

The advent of new sequencing technologies has radically lowered the cost of sequencing large-scale genomes. Next-generation sequencing technologies have revolutionized the way in which biological data is collected, analyzed, and interpreted. Due to the unprecedented amount of biological data now available, the difficulty is not in collection, but rather in trying to interpret and understand the ever-increasing amount of biological information.

It is widely known that the buildup of mutations and other structural genetic changes are all-important properties of genomic instability that can eventually lead to complex diseases, such as cancer [1], [2], [3], [4], [5]. Microhomology-mediated break-induced replication (MMBIR) is one of the mechanisms that can lead to various complex chromosomal rearrangements including copy number variations [6], [7]. Originally, MMBIR was proposed to explain complex duplications and triplications of non-continuous chromosome regions joined by microhomologies that were observed in patients with Pelizaeus-Merzbacher disease [8]. Soon after, many other studies were performed in cancer patients, yeast, and plants, which also documented MMBIR [9], [10], [11], [12]. Despite the important role that MMBIR plays in genomic instability, the mechanism of MMBIR remains unclear. This gap in our knowledge results in part from difficulties associated with detection of unselected MMBIR. The goal of this work was the development of an algorithm for detection of MMBIR events based on results of whole-genome sequencing.

Template-switching mutations are the hallmark of MMBIR [13]. As outlined in Fig. 1, the mutation event closely resembles the reference sequence with one key difference. A template-switching mutation event contains a template (the bold nucleotides in Fig. 1A) that matches the reference sequence and a downstream insertion that is the reverse complement of the template (the bold nucleotides in Fig. 1B). The inserted region is characterized as the template switching event or the MMBIR region. Additionally, small (3–8 b.p.) microhomology tags are found on the 3′ end of the template and the 5′ end of the insertion (the italicized, non-bolded nucleotides in Fig. 1). The insertion distance, the distance between the template and insertion (i.e. the distance between region A and B in Fig. 1), is usually between 5 and 100 b.p.

Today, most genetic variant methods focus on single-nucleotide polymorphisms (SNPs), small insertion deletion (microINDELs), and structural variants (SVs) [14], [15], [16]. More recent approaches aim to detect short tandem repeats (STRs) in DNA [17]. Most of the existing tools operate on the same general principle: scan the aligned reads for specific SV artifacts, reprocess (align) the specific regions, and calculate the resulting variant. However, while our approach does not deviate from the effective model, the uniqueness of MMBIR as a structural variant makes it difficult to detect using previous tools. Since MMBIR can result in the deletion and insertion of DNA tracks with unknown lengths followed by a possible inversion, other methods are not designed to capture the unique mutation pattern associated with MMBIR. To our knowledge, there is no existing bioinformatics tool to detect these template-switching events.

We have developed MMBIRFinder to detect template-switching events in both small and large-scale genomes. We first tested the program on the *Saccharomyces cerevisiae* genome. Using a simulated data set with 5,000 inserted MMBIRs, the tool detected 83 percent within five nucleotides and 90 percent within 10 nucleotides. To study the biological relevance, we further tested the tool on triple-negative breast cancer samples. The normal breast tissue sample contained 33 possible MMBIR regions while the triple-negative tumor sample contained 62 MMBIR events.

2 Methods

The MMBIRFinder method consists of three major steps. First, the BWA alignment tool (version 0.7.3) [18] is used to perform an initial alignment on the full genome. Additionally, unaligned reads from the initial alignment are extracted and half-reads are created and again aligned using BWA. Second, the aligned half-reads are then used to create a list of candidate MMBIR regions where one-half of the read is aligned, or anchored, to a specific location and the other half remains unaligned. The anchored read positions are used to cluster the reads into candidate regions of potential interest. A successive base-calling of the clustered reads creates a consensus read that is the most common nucleotide at each genomic location. Third, a series of local alignments on the consensus is performed and the MMBIR region and its matched template are recorded. A detailed analysis of the overall method is given below.

2.1 Identification of Reads Spanning MMMBIR Regions

To identify the candidate MMBIR region, the BWA alignment tool is conducted twice. First, the full set of reads is mapped against the reference genome. The parameters used in BWA ensure a highly accurate alignment with only one mismatch or error per read. The output of the first step is a SAM file containing all the aligned and unaligned reads [19]. Since MMBIR events contain regions that are sufficiently different from the reference (Figs. 1B and 2A), those reads that align to the reference are not included in the further analysis. Therefore, the unaligned reads are extracted in order to perform split-read mapping. In split-read mapping, the unaligned read is split at the halfway point (the X′s in Fig. 2A) and allows for increased coverage around the structural variant (Supplementary Fig. 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2359450 available online). BWA is again used against the unaligned split reads and the reference genome. Finally, the anchored reads are extracted into a structure of candidate reads; an anchored read is defined as a read where one half of the split-read is aligned to the reference genome while the other half remains unaligned. This is shown in Fig. 2A by the two-colored reads. In the figure, the dark grey reads indicate the anchored read. Since the anchored read is aligned to the genome at a specific location, it is now known where the unalignable half of the read is located on the reference genome. If both halfs of the read remain unaligned (the solid, light grey read in Fig. 2A), then the read is discarded due to lack of information. Similarly, if the two half-reads are aligned to non-consecutive locations on the genome, then the read is also discarded due to the ambiguity of the genomic location.

2.2 Identification of Potential MMBIR Regions

Anchored reads (where half of the original read can be aligned) within a pre-defined distance are further clustered using a simple distance-from-neighbor method. Clustering ensures that we are focusing on regions of high interest Since an anchored read only exists when half of the read is alignable, a region with a large number of anchored half reads must contain a localized unalignable region as well.

Let D represent the set of all anchored reads that were output from the SAM file in the previous step. Additionally, let r_i be genomic starting position of the ith read in the anchored half-read structure D, $d_{i;i+1}$ be the distance between the starting points of two reads r_i and r_{i+1}, and t be a pre-defined distance threshold. Furthermore, let C be a structure of clustered reads c such that c contains all overlapping reads r. Before clustering can commence, D must be sorted by both nucleotide position and chromosome number. We utilize the *gcc sort* method to ensure a worst-case scenario run time of $O(n * logn)$ [20].

Conceptually, we are clustering a pair of reads if there exists an overlap between the two reads. Algorithmically, for each two consecutive reads r_1 and r_2 we calculate $d_{1,2} = r_2 r_1$. If the distance $d_{1,2}$ between the two reads r_1 and r_2 is less than a pre-defined threshold t, then r_1 and r_2 will be both clustered into c_1. Subsequently, if the distance $d_{2,3}$ between r_2 and r_3 is less than t, r_3 is clustered into c_1 as well. This procedure continues until the distance between r_j and r_{j+1} is greater than t. Once $d_{j,j+1} > r_{j+1}\ r_j$, then r_{j+1} is clustered into c_2 and the process repeats.

After all the reads have been analyzed, the clusters with an insufficient number of reads below a minimum threshold (as specified in the user configuration file—default value of 20) are removed. The loci of these read clusters are considered potential MMBIR regions. The clustering method is fast as it operates, after sorting, in $O(n)$ time.

Within each potential MMBIR cluster, we conduct base-calling to derive the de *novo* sequence that best represents the event, based on all the aligned and unaligned portions of the anchored reads (indicated in Fig. 2B as the dark grey and light grey reads respectively). The base calling results in a new consensus read.

From above, let R_i be a read with length m in the cluster c at position i. Let s represent an arbitrary nucleotide in sequence $R_i = s_1, s_2, \dots s_m$. At each position j along the nucleotide sequence, the probability of each nucleotide is calculated and the highest probability nucleotide is the new nucleotide for base call read n at position j. In other words, n_j is the nucleotide with the highest probability at position j. However, the nucleotide with the highest probability must be at least 10 percent greater than the next highest probability nucleotide. Otherwise, an ambiguous ‘N’ is designated at n_j. For example, if the probability at n_j is 54 percent for nucleotide ‘C’ and 46 percent for ‘T’ then n_j is ‘N’ and is treated as any of the four nucleotides in the remaining steps.

2.3 Identification of Template-Switching Events

Once a consensus read has been created, the read must be compared to the original reference genome. Since a MMBIR region has a unique pattern of mismatches anchored on both sides by a long stretch of nucleotide matches Fig. 1) and the location of the consensus read is known (due to split-read mapping), a Smith-Waterman local alignment is performed against the reference genome (at the known location) and consensus read (Fig. 2C). The resulting alignment is then traversed in order to extract the location of the MMBIR region.

A finite state machine (FSM) is employed containing three states: pre-region, in-region, and post-region. The FSM operates by looking for regions of mismatches anchored on both sides by a long stretch of matches. Starting in the pre-region, the FSM travels along the alignment until it reaches a mismatch. If the number of consecutive mismatches is greater than or equal to the opening value (user-defined), the FSM transition into the in-region state and the location of the first mismatch is stored. Returning to the location of the first mismatch, the FSM continues until it encounters a match. Similarly, if the number of consecutive matches is greater than or equal to the closing value (user-defined), the FSM transition into the post-region state and the location of the first match is stored. An evaluation of the finite state machine parameters is given in Section 3.1.1.

Using Fig. 2C as an example, let the opening value of the FSM be 3 and the closing value be 6. Traversing along the alignment, the first reported mismatch is located at position 4 (note the small numbers above the reference genome). Continuing, locations 5, 6, and 7 are all consecutive mismatches. Since the number of consecutive mismatches (4) is greater than or equal to the defined opening value (3), the first nucleotide of the candidate MMBIR region is at location 4. Returning to location 4, the FSM traverses until location 8 where a match is found. Since location 9 is a mismatch and the number of consecutive matches (1) is less than the closing value (6), the FSM continues. Continuing, the nucleotides at locations 20–25 are all consecutive matches. Since the number of consecutive matches (6) is greater than or equal to closing value (6), location 19 is recorded as the last nucleotide in the candidate MMBIR region.

The nucleotides between the two stored locations constitute the candidate MMBIR region. One final check is required to ensure that the candidate region is in fact an actual MMBIR region. Once the candidate MMBIR region is identified, finding the template is straightforward. Since the MMBIR by definition contains the reverse complement of an upstream template region (Fig. 1), the reverse complement of the MMBIR is aligned against a long stretch (user-defined) of the reference genome upstream from the start of the candidate MMBIR region. If a region is found that is a near perfect match (i.e. within 1–2 nucleotide mismatches), the locations of the template and MMBIR regions are stored.

2.4 Detection and Removal of False Positives

Finally, before a potential MMBIR region is deemed accurate, an error-correcting step is conducted to help mitigate false positives. Due to the highly repetitive nature of human DNA, majority of the false positives reside in these regions. Therefore, a method that detects simple repeats was employed. Our method observes the frequency of dinucleotides, defined as two consecutive nucleotides, within a given MMBIR region.

From above, let m be the length of the MMBIR region and s_i and s_j be be nucleotides at position i and j respectively. Formally, a dinucleotide is defined as the concatenation of s_i

and s_j such that $j = i + 1$. By iteration over the MMBIR region we count the occurrence of each dinucleotide D_d where D is the frequency count of dinucleotide d. If D_d is greater than 90 percent of $m/2$ then the MMBIR is deemed to reside in a highly repetitive region and has a high probability of being a false positive. Only those regions that pass the error-correcting step are stored. It is this MMBIR region with a matching template that constitutes a template-switching event.

3 Results

The MMBIRFinder tool was primarily developed in C++ using the GNU gcc framework, version 4.7.3 [20]. The input to MMBIRFinder is simply a FASTA or FASTQ file of reads [21], a FASTA reference genome, and a configuration file (config.txt). The configuration file allows for alteration of the various thresholds, tolerances, finite state machine parameters, and input and output files. The program is freely available for download at https://github.com/msegar/MMBIRFinder.

3.1 Performance on the Simulated Data Set

We evaluated MMBIRFinder by inserting 5,000 artificial MMBIR regions in to the *Saccharomyces cerevisiae* genome (strain S288C). The length of the genome was 12,157,058 b.p. Paired-end reads were simulated with an average length of 75 b.p. at 50× coverage with a Gaussian distributed insert size of 200 b.p. Sequencing error followed normal Illumina data with an overall error rate around 1 percent [22]. The MMBIR regions had an average length of 10 b.p. with an average insert distance (the distance between the MMBIR and template) of 20 b.p.

The MMBIRFinder tool was evaluated using the simulated genome with default parameters. Majority of the run time was spent on the first stage, the BWA alignments, with an average alignment time of nearly 47 minutes. Once the alignments between the full- and half-reads were complete, MMBIRFinder took only 17 seconds to parse, cluster, and align the results. Of the 5,000 MMBIR events inserted into the genome, MMBIRFinder predicted 4,826 locations (Table 1). Of these, 4,519 were within 10 nucleotides of the exact location, 4,365 were within seven nucleotides, 4,150 were within five nucleotides, and 3,798 were within three nucleotides, for an accuracy of 90, 87, 83, and 76 percent respectively (Table 1 and Supplementary Fig. 2, available online). The tolerance in Table 1 is the maximum number of nucleotides between the predicted and actual MMBIR location to constitute an accurate prediction. Thus, a lower tolerance signifies a more accurate prediction.

We further examined whether the performance of the algorithm varies depending on the length of the MMBIR region. As indicated in Table 2, the algorithm consistently detects between 77 and 84 percent regardless of the MMBIR region length. This suggests that there is no bias with respect to MMBIR length and each region is equally likely to be detected.

3.1.1 Examination of False Negatives—Even though MMBIRFinder correctly identified between 76 and 90 percent of the MMBIR events in the simulated genome, determining why false negatives occurred can be useful in algorithm refinement. Above all, two key parameters add considerable variability in the MMBIR detection. First, since aligning and evaluating every candidate MMBIR region is unwieldy, clustering helps minimize the number of expensive computations at a cost of possibly removing important information. Typically, optimal sensitivity is achieved when setting the algorithm coverage, defined as the minimum number candidate reads to constitute a read cluster (Fig. 2B), to one-third the machine read coverage. For example, if the DNA sample was sequenced at 60× coverage, the algorithm reported optimal results using a coverage value of 20. However, as seen in the triple-negative breast cancer data, a high coverage parameter can cause potential MMBIR regions to be missed. Changing the coverage parameter to a lower value will increase the number of reported MMBIRs at a cost of decreased accuracy.

Second, changing the FSM variables to open and close a potential MMBIR region can drastically change the detection rate. Referring to Section 2.3, the opening value was defined as the number of consecutive mismatches in the alignment to start, or open, a candidate MMBIR region, while the closing value was defined as the number of consecutive matches in the alignment to end, or close, a candidate MMBIR region. In order to evaluate the effects of different FSM open and close parameters on the algorithm, we computed precision—the fraction of true positive over all predictions, and recall—the fraction of true positives over all variants, for various opening and closing value combinations (Fig. 3).

The value to open the MMBIR region caused the greatest difference in detection rate. Too many consecutive mismatches to open a candidate MMBIR region decreases the ability of the program to detect true positives (i.e. recall). Conversely, the value to close the MMBIR did not change the precision of the program. The precision within the same opening value group did not change much relative to the closing value. In other words, the recall and precision for an opening value of 2 and closing value of 4 was very similar to the 2:5 and 2:6 parameters. Similarly, the 3:4, 3:5, and 3:6 had equivalent precision and recall. Therefore, a value of four consecutive matches to close the potential MMBIR region was sufficient to detect a large majority of the implanted MMBIRs. However, it is unknown if a value less than 4 would result in the same precision. Conceptually, the FSM must account for random matches within the MMBIR region. Thus, a value of 1 to close the MMBIR would not suffice. It is also not uncommon to see two consecutive matches. Therefore, a value of at least three consecutive matches is the theoretical minimum threshold between random matches in the MMBIR region and the end of the MMBIR region itself. As evident from the graph, the 2:4, 2:5, and 2:6 parameters for the FSM optimize the MMBIR detection. Since only a fraction of MMBIR events are identifiable using a half-read approach, we do not expect the precision to be exactly 1.

3.2 Results for Triple Negative Breast Cancer Data

Further evaluation of the MMBIRFinder tool was conducted using real data acquired from the Susan G. Koman Tissue Bank at the Indiana University Simon Cancer Center. The whole-genome data was sequenced at the DNA sequencing Core Facility in Indianapolis, IN using Illumina Hi-Seq technology at 50× coverage for the normal sample and 70× coverage for the tumor sample using NimbleGen sequence capture and standard high-throughput sequence library preparation. Sequencing resulted in 117.2 and 299.6 million paired-end 100 b.p. reads for the normal and tumor sample respectively. Read alignment was against the hg19 reference genome. Comparing a normal breast data set against a tumor sample allows us to identify novel variants.

3.2.1 Results for Normal Sample—We ran MMBIRFinder on the normal breast sample obtained from TCGA. As summarized in Table 3, the alignment took around 50 hours and resulted in over 50 million aligned half-reads and nearly 67 million unaligned half-reads. Of these aligned half-reads we identified 26,576 clusters that had at least 20× read overlaps. Finally, 24 potential MMBIRs were identified. However, upon error-correction, eight were calculated to be false positives. The resultant 16 regions were considered to be accurate MMBIRs. A possible explanation for the miscalculation is the result of highly repetitive regions in DNA (Section 2.4). Despite an error-correcting step that eliminates the most obvious repetitive regions of DNA, the only way to determine the difference between a potential MMBIR and a false positive is through manual inspection. Further discussion on false positives is given in Section 4.

3.2.2 Results for Tumor Sample—Similarly, we ran MMBIRFinder on the tumor sample following the same procedure. As predicted, the tumor sample contained many more unaligned reads requiring a longer processing and run time. The BWA alignment took over 116 hours and resulted in nearly 118 million aligned half-reads and 182 million unaligned half-reads. We further identified 383,509 clusters that had at least 20× read coverage that resulted in 62 possible MMBIRs. Of the 62, five were filtered in the error-correcting step for a total of 57 potential MMBIRs identified (Table 3). Futhermore, in the 57 MMBIRs reported in the tumor sample, 14 were also present in the normal sample. Therefore, the number of potential novel MMBIRs detected that were present only in the tumor data set was 43. Manual inspection of the 43 potential MMBIRs resulted in 37 verified novel MMBIRs introduced in the tumor sample. The six additional false positive MMBIRs were located in highly repetitive regions of DNA that contained more than dinucleotide repititions (Section 2.4) and the two MMBIRs detected in the normal sample and not in the tumor sample were the result of too little coverage in the tumor data set (i.e. less than 20× coverage). Changing the coverage parameter in the config.txt file to 15× identified the missing two MMBIR regions.

The relatively few number of detected MMBIRs may reflect the large number of pathways for repairing DNA double-strand breaks. Ideally, a double-strand break would be repaired by an error-free pathway. However, double-strand breaks resulting from replication fork collapse or eroded telomeres are typically repaired by the highly mutagenic BIR pathway [23].

In all, the stark contrast between the normal and tumor sample supports the evidence that there is a strong correlation between cancer and complex mutations present at the genomic level. The number of reads that were unable to be aligned was nearly three times that of the normal sample.

3.3 Biological Relevance

While the detection of MMBIRs is novel, the true value is determining the biological impact of mutagenic template-switching events. The 37 detected MMBIRs unique to the tumor sample were further analyzed to determine their location relative to the nearest gene. Only those MMBIRs within −5000/+3000 were considered to be within the promoter region of a gene. A total of seven genes were found to reside with a promoter region. The recorded genes include *COL6A5*, a gene associated with cell-binding in soft tissue, and *FMO5*, a unique FMO associated with the cytochrome P450 drug-metabolism pathway [24]. A complete listed of the genes and their descriptions are listed in Table 4. While the correlation of template-switching events associated with MMBIR certainly does not imply causation, the detection of these minute genetic mutations can further enhance the knowledge of erroneous repair mechanisms within a DNA data set.

4 Conclusion and Discussion

Identifying novel variants is not trivial. Furthermore, identifying small microhomologies less than 20 b.p. in a 3,000 Mb human genome is likened to finding a needle in a haystack. MMBIRFinder is a versatile tool that makes searching for the outcomes indicative of MMBIR in whole genome sequencing data a much easier task. The use of half-read alignment and read clustering allows us to focus on regions of interest. Additionally, by allocating computational resources to regions of high confidence, we drastically lower the time it takes to analyze an entire human genome.

Our analyses showed that MMBIRFinder provides a highly accurate and specific framework for detecting template-switching events associated with microhomology-mediated break-induced replication. However, because of the naturally repetitive nature of DNA, DNA itself can be the cause of considerable variability in algorithm performance. Due to the small number of nucleotides in DNA and high probability of developing palindromes, our method falsely identifies areas of highly repetitive regions as potential MMBIRs. For example, the occurrence of microsatellites in the human genome allows for long stretches of highly repetitive DNA.

A microsatellite, also known as short tandem repeats and simple sequence repeats (SSRs), are repeating sequences between two and six base pairs in length [25]. For example, a common microsatellite is of the form $(TA)_n$ where n is the number of alleles. Therefore, when n is six the microsatellite sequence is TATATATATATA. Conversely, a misalignment at the three end of the sequence will be identified as a potential candidate region. Therefore, the reverse complement of the microsatellite (ATATATATATAT) and the microsatellite itself are nearly identical and are considered a positive MMBIR identification.

Microsatellites can trigger false positives when the repeated sequence is sufficiently large and greater than the minimum insertion distance (typically 5–10 b.p.). The challenge then lies in differentiating between a highly repetitive region and a template-switching event. Further research will aim to help distinguish between repeats and important genetic events.

Notably, our method is not sensitive enough to differentiate between somatic and germline mutations in cancer. In the simulation study, the reference genome was mutated by inserting MMBIR regions. Since the reference genome considers all cells to be the same, each mutation is assumed to be a germline mutation. In real world data samples, it is increasingly difficult to differentiate between a somatic and germline mutation.

Finally, as with most other NGS studies, a prediction can only be proven as a true positive with experimental validation. While at this time there is too much difficulty to verify our findings using PCR/Sanger sequencing, it should be noted that our method lends itself well to experimental validation.

Acknowledgments

The authors would like to thank the Center for Computational Biology and Bioinformatics at the Indiana University School of Medicine. This research is based upon work supported by the US National Science Foundation under Grant No. ABI-1062432 to Indiana University and the National Center for Genome Analysis Support. Additionally, AM is supported by NIH Grant GM084242. Finally, they thank Dr. Milan Radovich for use of the triple-negative breast cancer sequenced data.

References

1. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability-An evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. 2010 Mar; 11(3):220–228. [Online] Available: www.ncbi.nlm.nih.gov/pubmed/20177397. [PubMed: 20177397]
2. McEachern MJ, Haber JE. Break-induced replication and recombinational telomere elongation in yeast. Annu. Rev. Biochem. 2006 Jan.75:111–135. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16756487. [PubMed: 16756487]
3. Sakofsky CJ, Ayyar S, Malkova A. Break-induced replication and genome stability. Biomolecules. 2012 Dec; 2(4):483–504. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23767011. [PubMed: 23767011]
4. Malkova A, Haber JE. Mutations arising during repair of chromosome breaks. Annu. Rev. Genetics. 2012 Dec.46:455–73. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23146099. [PubMed: 23146099]
5. Deem A, Keszthelyi A, Blackgrove T, Vayl A, Coffey B, Mathur R, Chabes A, Malkova A. Break-induced replication is highly inaccurate. PLoS Biol. 2011 Jan.9(2):e1000594. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3039667&tool=pmcentrez&rendertype=abstract. [PubMed: 21347245]
6. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genetics. 2009 Jan.5(1):e1000327. [Online]. Available: www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2621351&tool=pmcentrez&rendertype=abstract. [PubMed: 19180184]
7. Bosco G, Haber JE. Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. Genetics. 1998 Nov; 150(3):1037–1047. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1460379&tool=pmcentrez&rendertype=abstract. [PubMed: 9799256]
8. Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007; 131(7):1235–1247. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867407015413. [PubMed: 18160035]
9. Kryh H, Abrahamsson J, Jegeras E, Sjoberg R-M, Devenney I, Kogner P, Martinsson T. MYCN amplicon junctions as tumor-specific targets for minimal residual disease detection in neuroblastoma. Int. J. Oncol. 2011 Nov; 39(5):1063–1071. [Online]. Available: http://www.researchgate.net/publication/51487326_MYCN_amplicon_junctions_as_tumor-specific_targets_for_mini-mal_residual_disease_detection_in_neuroblastoma. [PubMed: 21750863]
10. Lawson ARJ, Hindley GFL, Forshew T, Tatevossian RG, Jamie GA, Kelly GP, Neale GA, Ma J, Jones TA, Ellison DW, Sheer D. RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. Genome Res. 2011 Apr;

21(4):505–514. [Online]. Available: http://genome.cshlp.org/content/21/4/505.short. [PubMed: 21393386]

11. Cappadocia L, Parent J-S, Zampini E, Lepage E, Sygusch J, Brisson N. A conserved lysine residue of plant Whirly proteins is necessary for higher order protein assembly and protection against DNA damage. Nucleic Acids Res. 2012 Jan; 40(1):258–269. [Online]. Available: http://nar.oxfordjournals.org/content/40/1/258. [PubMed: 21911368]
12. Yatsenko SA, Hixson P, Roney EK, Scott DA, Schaaf CP, Ng Y-T, Palmer R, Fisher RB, Patel A, Cheung SW, Lupski JR. Human subtelomeric copy number gains suggest a DNA replication mechanism for formation: Beyond breakage-fusion-bridge for telomere stabilization. Human Genetics. 2012 Dec; 131(12):1895–1910. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3493700&tool=pmcentrez&rendertype=abstract. [PubMed: 22890305]
13. Smith CE, Llorente B, Symington LS. Template switching during break-induced replication. Nature. 2007 May; 447(7140):102–105. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17410126. [PubMed: 17410126]
14. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods. 2009 Nov; 6(11):S13–S20. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19844226"> http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1374.htmlhttp://www.ncbi.nlm.nih.gov/pubmed/19844226. [PubMed: 19844226]
15. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep; 28(18):i333–i339. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3436805&tool=pmcentrez&rendertype=abstract. [PubMed: 22962449]
16. Jiang Y, Wang Y, Brudno M. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics. 2012 Oct; 28(20):2576–2583. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22851530. [PubMed: 22851530]
17. Gymrek M, Golan D, Rosset S, Erlich Y. A microhomology-mediated break-induced replication model for the origin of human copy number variation. Genome Res. 2012 Jun; 22(6):1154–1162. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2621351&tool=pmcentrez&rendertype=abstract. [PubMed: 22522390]
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760. [Online]. Available: http://www.pubmed-central.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract. [PubMed: 19451168]
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009 Aug; 25(16):2078–2079. [Online] Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract. [PubMed: 19505943]
20. Stallman RM. Using and porting the GNU compiler collection. Free Softw. Found. 1989; 2(59):02 111–1307.
21. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 Apr; 38(6):1767–1771. [Online]. Available: http://wwwpubmedcentral.nih.gov/articlerender.fcgi?artid=2847217&tool=pmcentrez&rendertype=abstract. [PubMed: 20015970]
22. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS One. 2012 Jan.7(2):e30087. [Online]. Available: www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3277595&tool=pmcentrez&rendertype=abstract. [PubMed: 22347999]
23. Llorente B, Smith C, Symington L. Break-induced replication: What is it and what is it for? Cell Cycle. 2008; 7(7):859–864. [Online]. Available: http://www.landesbioscience.com/journals/6/article/5613/. [PubMed: 18414031]
24. Mosca E, Alfieri R, Merelli I, Viti F, Calabria A, Milanesi L. A multilevel data integration resource for breast cancer study. BMC Syst. Biol. 2010 Jun.4(1):76. [Online]. Available: www.ncbi.nlm.nih.gov/pubmed/20177397. [PubMed: 20525248]
25. Turnpenny, P.; Ellard, S. Emery's Elements of Medical Genetics. 12th. London, UK: Elsevier; 2005.

Biographies

Matthew W. Segar received the BA degree in computer science from Bucknell University in 2012 and the MS degree in bioinformatics from Indiana University in 2014. Since 2012, he has been a researcher with the Center for Computational Biology and Bioinformatics at the Indiana University School of Medicine. He is currently working towards the MD degree at the Indiana University School of Medicine. His current research interests include *de novo* sequence assembly, next-generation sequencing, and DNA methylation epigenetics.

Cynthia J. Sakofsky received the BS degree from New York University in 2000 and the MS degree from Indiana University Bloomington in 2004. She received the PhD degree from University of Cincinnati in 2011 focusing on DNA repair. She is currently a postdoctoral research fellow in Dr. Anna Malkova's lab at the University of Iowa. Her research interests include understanding mechanisms of genome instability, specifically aimed at understanding mechanisms that lead to mutagenesis, chromosomal rearrangements and other destabilizing events similar to those found in cancer and other diseases.

Anna Malkova received the PhD degree in Genetics in 1993 from the St. Petersburg University, Russia. She is currently an associate professor in the Biology Department, College of Liberal Arts and Sciences, the University of Iowa. She completed postdoctoral training in molecular genetics at Brandeis University. From 2003 till 2013, she worked as a faculty at Indiana University-Purdue University Indianapolis, and joined the University of Iowa in 2014. Her current research is focused on understanding the mechanisms of DNA repair.

Yunlong Liu received the BS degree from Harbin Engineering University in 1996, the MS degree from Tsinghua University in 1999, and the PhD degree in biomedical engineering from Purdue University in 2004. He is currently an associate professor in the Department of Medical and Molecular Genetics and Division of Biostatistics at the Indiana University School of Medicine. He completed his postdoctoral training at the Indiana University's School of Medicine and has since been the Bioinformatics core director at the Center for Computational Biology and Bioinformatics. His research interests include transcriptional, post-transcriptional and epigenetic regulation and informatics pipelines in analyzing data from NGS technologies.

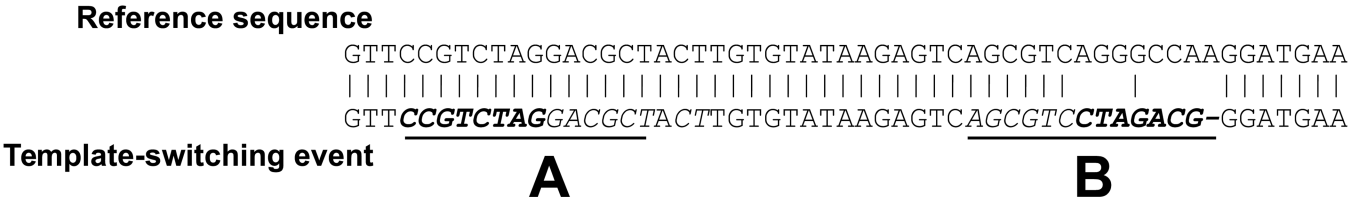
Reference sequence
GTTCCGTCTAGGACGCTACTTGTGTATAAGAGTCAGCGTCAGGGCCAAGGATGAA
Template-switching event
GTTCCGTCTAGGACGCTACTTGTGTATAAGAGTCAGCGTCCTAGACG-GGATGAA
A
B

Fig. 1.

Examples of template-switching events identified in yeast. The reference sequence indicates the donor DNA strand that is being copied during MMBIR. In the template-switching event, the —indicates a deleted nucleotide. The bold nucleotides in region **A** denote the sequences that served as the template during the template-switching event. The bold nucleotides in region **B** indicate insertions. The italicized, non-bolded nucleotides in both sections indicate micro-homologies located at junctions of template-switching events.

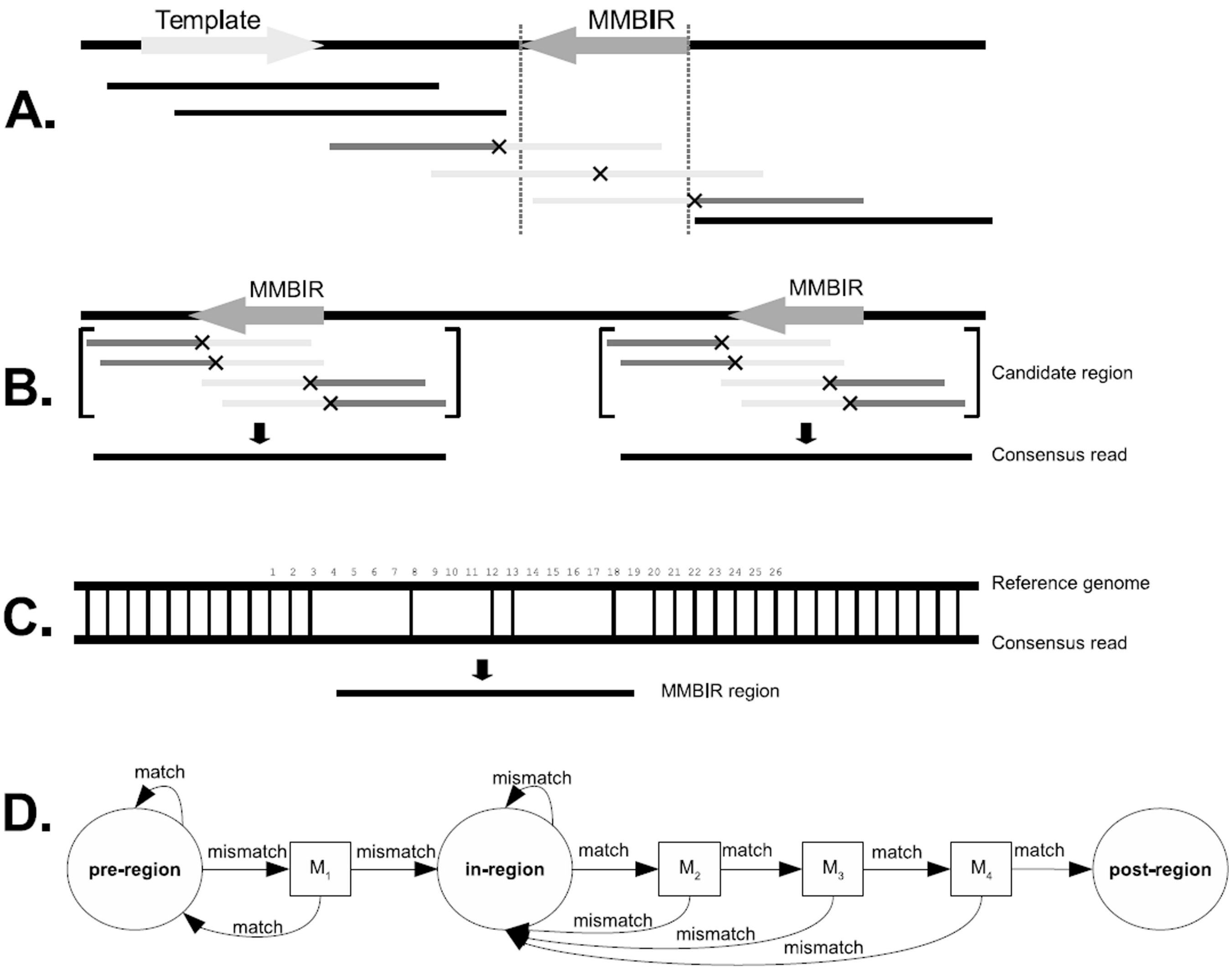

Fig. 2.

The main steps of the MMBIRFinder tool. **A**. BWA is used to conduct two consecutive alignments. First, the full reads are aligned against the reference genome. Next, the unaligned half-reads are mapped. The top black line is the reference sequence with the light grey arrow representing the template and the dark grey arrow representing the MMBIR region. The MMBIR region will not be alignable to the reference genome. The smaller black lines represent the full reads. The Xs are the halfway point where the unaligned reads (light grey lines) are split to create half-reads. Dark grey lines represent anchored, or alignable, half-reads. **B**. Anchored half-reads are consolidated into clusters where there exists an overlap between the reads. Again, the grey arrows represent multiple MMBIR regions on the reporter sequence. Clusters are represented using [ ] brackets. Each cluster is then base called to create a consensus read. Due to the randomness of DNA or the allowance of mismatches in BWA, it is possible for a small portion of the half read to align to the MMBIR region. Split-read alignment does not guarantee an exact border between the SV and the original genome, it rather limits the search space and reduces the time necessary to find the possible MMBIR region. **C**. The final step is the local alignment of the consensus read with the reference. If a potential MMBIR region is found, the reverse complement is used to find the template upstream from the MMBIR region. The vertical bars indicate matches between the reference and the consensus read. The numbers represent the nucleotide locations. As indicated, the MMBIR region starts at nucleotide position 4 and ends at nucleotide position 19. Notice how there is a large gap of unaligned nucleotides to the reference in the MMBIR region. **D**. The finite-state machine to identify the MMBIR region. The FSM consists of three main state: pre-region, in-region, and post-region represented by the circles. In the example above, two consecutive mismatches are required to identify the start of the MMBIR region and four consecutive matches are required to identify the end of the MMBIR region. When comparing the reference genome and the consensus read (Fig. 2C above), the FSM starts in the pre-region and identifies nucleotide matches. If a mismatch occurs the FSM transitions to M_1. If a match occurs the FSM transitions back to the pre-region. Otherwise, after two consecutive mismatches, the FSM transitions to the second state, in-region, and the position location is stored. In the in-region state, if four consecutive mismatches occur, the FSM transitions from the in-region to the post-region and the location (constituting the end of the MMBIR region) is stored.

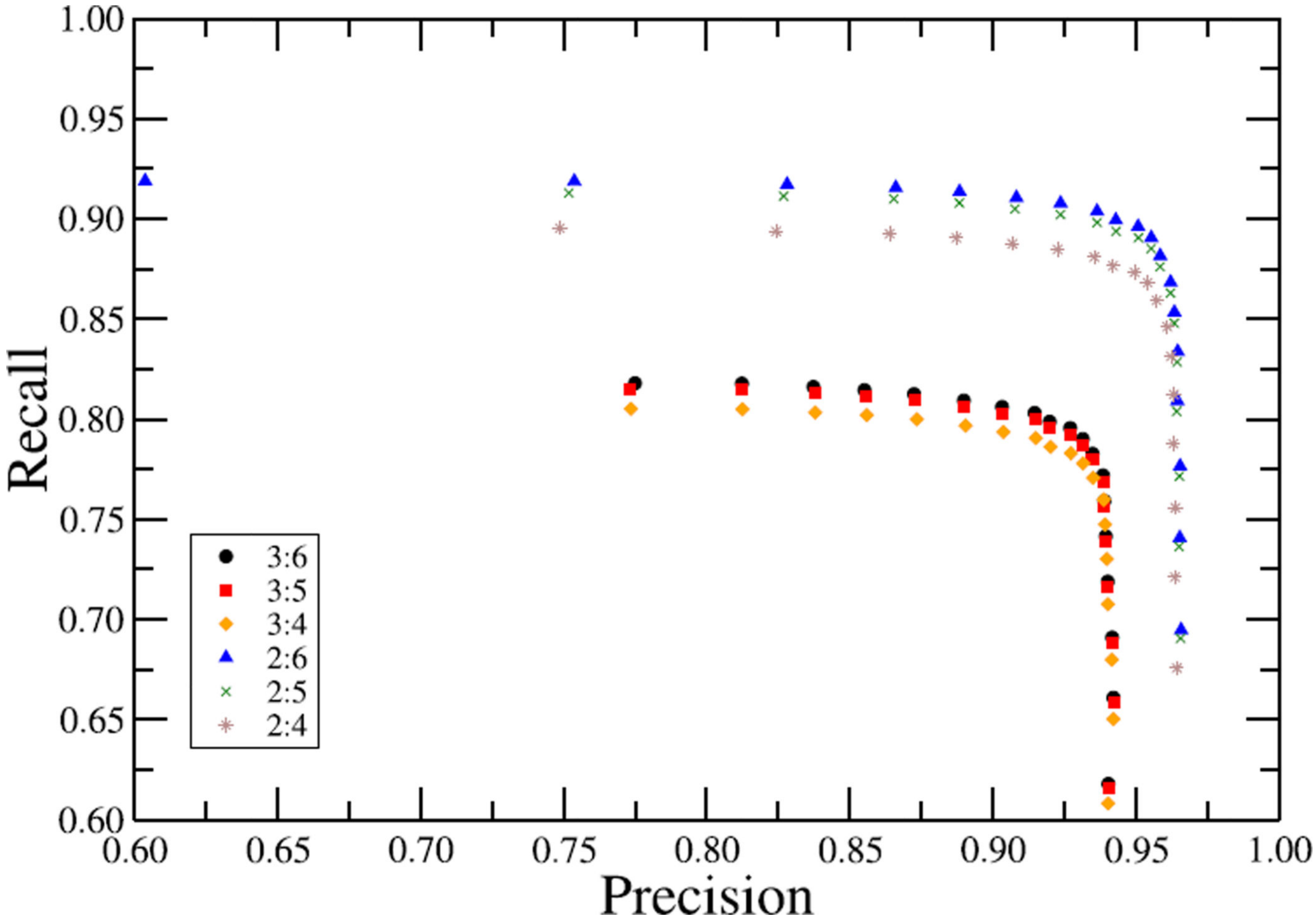

Fig. 3.

Precision versus recall graph of the FSM parameters in the form of *x*:*y*, where *x* is the *opening value*—the number of consecutive unaligned nucleotides to open the MMBIR region, and *y* is the *closing value*—the number of consecutive aligned nucleotides to close the MMBIR region. Each point for each FSM parameter (*x*:*y*) represents a changing tolerance from 0 to 17.

TABLE 1

Summary of MMBIRFinder Statistics

+/− Tolerance	TP	FP	FN	Sensitivity
10	4,519	307	481	90.4%
7	4,365	461	635	87.3%
5	4,150	676	850	83.0%
3	3,798	1,028	1,202	76.0%

Tolerance is the number of nucleotides between the predicted and actual MMBIR location to constitute an accurate prediction. A lower tolerance means a more accurate prediction.

TABLE 2

Detection Statistics Based on Length of MMBIR Region

Length	Detected	Actual	Percentage
14–15	363	470	77%
16–17	818	996	82%
18–19	812	976	83%
20–21	806	1,012	79%
22–23	878	1,042	84%
24–25	402	504	79%

Length is the length of the actual MMBIR region. Detected is the number of true positive identifications outputted from MMBIRfinder. Actual is the number of inserted MMBIR regions in the simulated sample.

TABLE 3

Comparison of the Normal and Tumor Sample on Real Triple-Negative Breast Cancer Data

	Normal sample	Tumor sample
Total number of reads (mil)	1,245	2,021
Coverage	50×	70×
BWA alignment time (hrs)	50	116
Aligned half-reads (mil)	50.3	117.9
Unaligned half-reads (mil)	66.9	181.7
Clusters	26,576	383,509
Clustering time (s)	918	28,083
MMBIR identification time (s)	260	7,227
Predicted MMBIRs (before e.c.[*])	24	62
Predicted MMBIRs (after e.c.[*])	16	57

[*] e.c. = error-correcting step

TABLE 4

The Seven MMBIRs that Resided within the Promoter Region of a Gene

Name	Description	Insert
FMO5	Flavin containing monooxygenase 5	−1,612
COL6A5	Collagen, Type VI, Alpha 5	−2,355
CNOT8	CCR-NOT transcrption complex, subunit 8	−2,480
LINC00951	Long intergenic non-protein coding RNA 951	2,679
DRAP1	DR1-Associated Protein 1	−1,841
STOML3	Stomatin (EPB72)-like 3	2,862
ANKRD30B	Ankyrin repeat doman 30B	−2,695

A Description of the Gene and the Distance from the MMBIR Starting Location is Listed