Integrated analysis of whole-genome paired-end and mate-pair sequencing data for identifying genomic structural variations in multiple myeloma

Rendong Yang, Emory University
Li Chen, Emory University
Scott Newman, Emory University
Khanjan Gandhi, Emory University
Gregory Doho, Emory University
Carlos S Moreno, Emory University
Paula Vertino, Emory University
Leon Bernal-Mizarchi, Emory University
Sagar Lonial, Emory University
Lawrence Boise, Emory University

Only first 10 authors above; see publication for full author list.

Journal Title: Cancer Informatics
Volume: Volume 13, Number Suppl 2
Publisher: Libertas Academica Ltd. | 2014-09-21, Pages 49-53
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.4137/CIN.S13783
Permanent URL: https://pid.emory.edu/ark:/25593/mr3hh

Final published version: http://dx.doi.org/10.4137/CIN.S13783

Copyright information:
© 2014 the author(s), publisher and licensee Libertas Academica Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), which permits making multiple copies, distribution, public display, and publicly performance, distribution of derivative works, provided the original work is properly cited. This license requires copyright and license notices be kept intact, credit be given to copyright holder and/or author. This license prohibits exercising rights for commercial purposes.

Accessed December 31, 2017 3:28 PM EST
Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Integrated Analysis of Whole-Genome Paired-End and Mate-Pair Sequencing Data for Identifying Genomic Structural Variations in Multiple Myeloma

Rendong Yang1,†, Li Chen2, Scott Newman3, Khanjan Gandhi3, Gregory Doho4, Carlos S. Moreno5, Paula M. Vertino6, Leon Bernal-Mizarchi7, Sagar Lonial7, Lawrence H. Boise7, Michael Rossi4,6, Jeanne Kowalski1,3 and Zhaohui S. Qin1

1Department of Biostatics and Bioinformatics, Emory University, Atlanta, GA, USA. 2Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA. 3Winship Biostatistics and Bioinformatics Shared Resource, Atlanta, GA, USA. 4The Emory Integrated Genomics Core, Emory University, Atlanta, GA, USA. 5Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA. 6Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA. 7Department of Hematology & Medical Oncology, Emory University School of Medicine, Atlanta, GA, USA. 8Current address: Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN, USA.

ABSTRACT: We present a pipeline to perform integrative analysis of mate-pair (MP) and paired-end (PE) genomic DNA sequencing data. Our pipeline detects structural variations (SVs) by taking aligned sequencing read pairs as input and classifying these reads into properly paired and discordantly paired categories based on their orientation and inferred insert sizes. Recurrent SV was identified from the discordant read pairs. Our pipeline takes into account genomic annotation and genome repetitive element information to increase detection specificity. Application of our pipeline to whole-genome MP and PE sequencing data from three multiple myeloma cell lines (KMS11, MM.1S, and RPMI8226) recovered known SVs, such as heterozygous TRAF3 deletion, as well as a novel experimentally validated SPI1 – ZNF287 inter-chromosomal rearrangement in the RPMI8226 cell line.

KEYWORDS: structural variations, multiple myeloma, whole-genome sequencing, variant detection

Introduction

Structural changes to cancer genomes have important effects such as oncogene amplification, tumor suppressor gene disruption, and chimeric fusion gene formation. Structural changes from each of these categories contribute important neoplastic events such as amplification of HER2 in breast cancer, deletion of CDKN2A in a number of different cancers, and formation of fusion genes such as BCR-ABL in chronic myeloid and other leukemias. Identifying such structural changes is a crucial part of most diagnostic workups as the presence or absence of certain genome rearrangements can allow for risk stratification and targeted therapy.

Recent array CGH and whole-genome sequencing studies have shown that cancer genomes are often highly rearranged. For example, array CGH copy number profiles often contain tens to hundreds of discrete copy number changes. Such complexity has been difficult to define using conventional cytogenetics, and many clinical and research laboratories
now rely on array CGH as a first-line assay for structural and numerical changes to chromosomes.

However, array CGH only detects copy number changes and no structural information is implicit in this methodology. Nevertheless, cytogeneticists and researchers now face a new challenge: to make clinical sense of a complex array CGH profile. To do this, they must assign each separate copy number imbalance to one of the two categories: pathogenic or benign. Although some copy number changes such as amplification of *HER2* or homozygous deletion of *CDKN2A* are clearly pathogenic and copy number changes in regions such as the Yq heterochromatin are probably benign, the majority of copy number changes are of uncertain significance.

When structural information is available in conjunction with copy number data, variants of uncertain significance can often be classified as pathogenic or benign. For example, a 500-kb duplication containing only one gene would likely be classed as uncertain significance so long as the gene had no known role in cancer. If, however, we knew that this 500-kb region had inserted itself into the *CDKN2A* locus and disrupted one copy of the gene, we could now class the duplication as pathogenic. Knowing how individual copy number gains and losses relate to one another within the rearranged genome is potentially of great clinical utility.

The necessary structural information can come from whole-genome paired-end (PE) or mate-pair (MP) sequencing. These next-generation sequencing methodologies provide information about the genes disrupted at chromosome breakpoints. Although many tools are available to detect structural changes and their genetic consequences from whole genome and transcriptome, they are all stand alone tools that are relatively difficult for a non-specialist to integrate into their clinical analysis workflow.

Here, we describe structural variation (SV) finder a fast, lightweight, and easy to use tool that identifies structural rearrangements in cancer genomes and outputs data that can be integrated into downstream analysis or viewed in a genome browser with other type data. We show the utility of this approach using integrated genomic data from three highly rearranged multiple myeloma cell lines.

### Results

**Whole genome PE and MP sequencing data.** From Illumina PE and MP sequencing of three multiple myeloma cell lines (KMS11, MM.1S, and RPMI8226), we obtained around 15× PE and 5× MP sequence-level coverage (Table 1). The MP reads differ from PE reads, by having a larger insert size (approximately 3 kb) and an outward facing (reverse–forward) read pairs orientation due to a circularization procedure used in MP preparation. The average sequencing quality of MP and PE reads are satisfactory (over 30) as shown in Table 1. Therefore, read trimming is not carried out prior to mapping. We reverse-complemented all MP reads and aligned the PE and preprocessed MP reads with the Burrows-Wheeler Aligner (BWA) algorithm. Over 90% of PE and 50% of MP reads were mapped to human reference genome GRCh37 (hg19).

**SV identification with SVfinder.** To detect chromosomal rearrangements, we developed the SVfinder pipeline (Fig. 1). The first step of the algorithm involves classifying mapped read pairs into two groups: concordant and discordant pairs based on the bitwise flag component of the sequence alignment/map (SAM) file. Concordant pairs are defined as read pairs that mapped to the reference genome with the expected orientation and insert size. For PE reads, the SAM file bitwise flag 0x2 indicates that the reads are mapped properly, meaning that the reads are correctly oriented with respect to one another, i.e., that one of the MPs maps to the forward strand and the other maps to the reverse strand and both the ends were mapped within a reasonable distance given the expected distance (and standard deviation) that the aligner inferred. In this study, BWA is used for mapping, but the SV detection pipeline can accept SAM format output files from other aligners as well, for example Novoalign (http://www.novocraft.com).

Next, SVfinder estimates the normal insert size range by calculating the mean and standard deviation of insert sizes of concordant reads. In parallel, the discordant pairs are grouped into interconnected clusters, which are hypothesized to originate from the same SV. The clustering procedure is performed by extending each seed pair 1 kb from

<table>
<thead>
<tr>
<th>TYPE</th>
<th>CELL LINE</th>
<th>READ LENGTH (BP), DEPTH, QUALITY SCORE</th>
<th>NO. OF TOTAL READS</th>
<th>NO. OF MAPPED READS</th>
<th>FRAGMENT LENGTH MEAN ± SD (BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair-end</td>
<td>KMS11</td>
<td>2 × 100, 12X, 35</td>
<td>374,274,810</td>
<td>347,448,960</td>
<td>301 ± 75</td>
</tr>
<tr>
<td></td>
<td>MM.1S</td>
<td>2 × 100, 14X, 34</td>
<td>429,622,446</td>
<td>394,631,072</td>
<td>296 ± 78</td>
</tr>
<tr>
<td></td>
<td>RPMI8226</td>
<td>2 × 100, 13X, 34</td>
<td>394,471,850</td>
<td>366,210,450</td>
<td>291 ± 81</td>
</tr>
<tr>
<td>Mate-pair</td>
<td>KMS11</td>
<td>2 × 100, 7X, 31</td>
<td>221,885,782</td>
<td>48,782,044</td>
<td>3,268 ± 496</td>
</tr>
<tr>
<td></td>
<td>MM.1S</td>
<td>2 × 100, 3X, 35</td>
<td>86,871,758</td>
<td>43,464,282</td>
<td>3,646 ± 244</td>
</tr>
<tr>
<td></td>
<td>RPMI8226</td>
<td>2 × 100, 4X, 36</td>
<td>114,486,206</td>
<td>59,310,148</td>
<td>3,501 ± 340</td>
</tr>
</tbody>
</table>
Paired-end/mate pair mapped reads (SAM file format)

Flag number

Concordant pairs

Insert size (l) estimation

Mean (µ), Variation (σ^2)

Discordant pairs

Clustering by position

Interconnected clusters

Classifying discordant clusters (# of pairs ≥2)

(l) < µ−3σ

(l) > µ+3σ

(l) ∈ [µ−3σ, µ+3σ]

(l) = ∞

Short insert size

Long insert size

Normal insert size

No insert size

Pair orientation

Insertion

Complex

Deletion

Inversion

Tandem duplication

Inter-chromosome translocation

Genomic annotation

List of SVs with gene fusion prediction (TXT and BED format output)

Figure 1. An overview of the SVfinder pipeline. SVfinder takes a SAM file as input and divides the mapped reads into concordant and discordant group. The normal insert size range is estimated from reads in concordant group. The discordant reads are clustered and classified into different types of variation subgroup according to their insert size and read pairs orientation. Identified SVs are annotated and output in BED format, allowing for easy downstream analysis or viewing in a genome browser.

each end and merging any read pair with overlapping of the extended genomic interval. In the third step, six types of SV are determined, including deletion, insertion, inversion, tandem duplication, inter-chromosomal translocations, and complex events, based on the insert size and read-pair orientation as shown in Figure 1. A novel feature of the SVfinder pipeline is to add the genomic annotation information in the identified SV list at the final step. In this way, possible gene fusions caused by genomic structural variation (SV) are predicted if both sides of the SV breakpoint reside in coding regions. Moreover, SVfinder provides BED format output for identified SVs, which can be easily imported into a genome browser for visualization, such as Integrative Genome Viewer (IGV).

Table 2. Known and novel SVs associated with genes identified by SVfinder.

<table>
<thead>
<tr>
<th>GENE A</th>
<th>GENE B</th>
<th>POSITION A</th>
<th>POSITION B</th>
<th>CHR A</th>
<th>CHR B</th>
<th>KMS11</th>
<th>MM.1S</th>
<th>RPMI8226</th>
<th>TYPE</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHSC1</td>
<td>IGH</td>
<td>intron</td>
<td>upstream</td>
<td>chr4</td>
<td>chr14</td>
<td>PE</td>
<td></td>
<td></td>
<td>Translocation</td>
<td>20</td>
</tr>
<tr>
<td>MAF</td>
<td>IGH</td>
<td>intron</td>
<td>downstream</td>
<td>chr16</td>
<td>chr14</td>
<td>PE</td>
<td></td>
<td></td>
<td>Translocation</td>
<td>20</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>CDKN2B</td>
<td>coding</td>
<td>intron</td>
<td>chr9</td>
<td>chr9</td>
<td>PE</td>
<td></td>
<td></td>
<td>Deletion</td>
<td>21</td>
</tr>
<tr>
<td>TRAF3</td>
<td>CDC42BPB</td>
<td>coding</td>
<td>intron</td>
<td>chr14</td>
<td>chr14</td>
<td>MP/PE</td>
<td></td>
<td></td>
<td>Deletion</td>
<td>22</td>
</tr>
<tr>
<td>SPI1</td>
<td>ZNF287</td>
<td>coding</td>
<td>coding</td>
<td>chr11</td>
<td>chr17</td>
<td>MP/PE</td>
<td></td>
<td></td>
<td>Translocation</td>
<td>Novel</td>
</tr>
<tr>
<td>KDM6A</td>
<td>KDM6A</td>
<td>intron</td>
<td>intron</td>
<td>chrX</td>
<td>chrX</td>
<td>MP/PE</td>
<td></td>
<td></td>
<td>Deletion</td>
<td>Novel</td>
</tr>
<tr>
<td>KCTD8</td>
<td>KCTD8</td>
<td>intron</td>
<td>intron</td>
<td>chr4</td>
<td>chr4</td>
<td>MP/PE</td>
<td></td>
<td></td>
<td>Deletion</td>
<td>Novel</td>
</tr>
<tr>
<td>ABHD17B</td>
<td>C9orf85</td>
<td>coding</td>
<td>intron</td>
<td>chr9</td>
<td>chr9</td>
<td>MP/PE</td>
<td></td>
<td></td>
<td>Deletion</td>
<td>Novel</td>
</tr>
</tbody>
</table>
Application of SVfinder to whole-genome sequencing data in multiple myeloma. We applied SVfinder to whole-genome PE and MP sequencing data for the three multiple myeloma cell lines. These cell lines have several well-characterized mutations and SVs, allowing us to validate SVfinder. In our identified SVs list, three known variations were recovered from PE sequencing data (Table 2): the t(4;14) translocation that fuses the IGH locus with MMSET identified in KMS11 cells, the t(14;16) translocation involving the IgH and MAF genes, and CDKN2A deletion found in MM.1S. Additionally, the known TRAF3 deletion was found in both MP and PE data in RPMI8226.

We next searched for high-confidence, novel SVs – those that were predicted in both MP and PE data. Second, we filtered out possible artifactual SV predictions that resided within repetitive regions. We identified four novel SV candidates, a SPI1-ZNF287 fusion caused by an inter-chromosomal rearrangement (Fig. 2A), and deletions of KDM6A, KCTD8, and ABHD17B (Table 2). All these variations were identified in RPMI8226. We selected the SPI1-ZNF287 rearrangement for further experimental validation. PCR primers designed to span the rearrangement were used to amplify genomic DNA from RPMI8226 cells and the product subjected to Sanger sequencing. The results confirmed that the exon 4 of SPI1 is rearranged with ZNF287 intron region (Fig. 2B).

Conclusions
We present a tool – SVfinder to detect SV from whole-genome PE or MP sequencing data. Compared with other existing SV detection methods, SVfinder is easy to use and integrates the genomic annotation and repetitive region information to filter false positives as well as predicting gene fusions. By applying our method to multiple myeloma whole-genome sequencing data, we were able to recover known recurrent translocations and deletions in multiple myeloma as well as identify several novel SVs. SVs predicted by SVfinder could be experimentally validated. Our pipeline outputs an SV list in BED format, which provides a convenient gateway for downstream analysis by integrating with other software workflows and allowing direct visualization of the results in any genome browser.

Methods
Sample preparation and data processing. MP and PE libraries were constructed according to standard protocols using Mate Pair Library Prep Kit v2 Genomic DNA sample prep kits. Sequencing for each sample and library was over a single lane of an Illumina HiSeq 2000 instrument. Raw data in Fastq format were aligned to the hg19 human reference genome using the BWA version 0.5.8a. The raw sequencing data have been achieved at the Sequence Read Archive (SRA) with accession number SRP039529.
SV annotation, filtering, and validation. Each SV is annotated based on the positions of breakpoints and their overlap with gene regions. Gene annotation files were downloaded from UCSC genome browser with the track “UCSC genes” from human reference genome GRCh37 (hg19). The genomic zone of SVs included 5’ distal, 5’ proximal, 5’ UTR, coding, 3’ UTR, 3’ proximal, 3’ distal, intergenic. Identified SVs were excluded if they reside in repetitive regions. Genomic repeat information was obtained from UCSC genome browser RepeatMasker track. Of the novel SV candidates, the SPI1 and ZNF297 rearrangement was PCR amplified from genomic DNA isolated from RPMI8226 and the product was subject to Sanger sequencing. The primers used for PCR amplification and sequencing were 5’-CTCGGCCTCCTCTCATCTGAGCT (SPI1) and 3’-AAGGCCATGCA-TCTGTCAT (ZNF287).

Software availability. SVfinder is written in Python, and the source code and manual are available from: https://github.com/cauyrd/SVfinder

Author Contributions
Conceived and designed the experiments: RY, MR, LHB, SL, JK, ZSQ. Analyzed the data: RY, LC, SN, KG, GD. Wrote the first draft of the manuscript: RY, SN. Contributed to the writing of the manuscript: PMV, LHB, SL. Agree with manuscript results and conclusions: CSM, LBM. Jointly developed the structure and arguments for the paper: RY, SN. Made critical revisions and approved final version: RY, SN, LBM, SL, LHB, MY, JK, ZSQ. All authors reviewed and approved the final manuscript.

REFERENCES