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: 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 Yang[1,†], Li Chen[2], Scott Newman[3], Khanjan Gandhi[3], Gregory Doho[4], Carlos S. Moreno[5], Paula M. Vertino[6], Leon Bernal-Mizarchi[7], Sagar Lonial[7], Lawrence H. Boise[7], Michael Rossi[4,6], Jeanne Kowalski[1,3] and Zhaohui S. Qin[1]

[1]Department of Biostatics and Bioinformatics, Emory University, Atlanta, GA, USA. [2]Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA. [3]Winship Biostatistics and Bioinformatics Shared Resource, Atlanta, GA, USA. [4]The Emory Integrated Genomics Core, Emory University, Atlanta, GA, USA. [5]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA. [6]Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA. [7]Department of Hematology & Medical Oncology, Emory University School of Medicine, Atlanta, GA, USA. [†]Current 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

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Yang et al. Integrated Analysis of Whole-Genome Paired-End and Mate-Pair Sequencing Data for Identifying Genomic Structural Variations in Multiple Myeloma. *Cancer Informatics* 2014:13(S2) 49–53 doi: 10.4137/CIN.S13783.

RECEIVED: March 9, 2014. **RESUBMITTED:** April 28, 2014. **ACCEPTED FOR PUBLICATION:** April 29, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: Research reported in this publication was supported in part by the Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI under award number P30CA138292. NIH grants R01 HG005119 and R21 HG 004751. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Other funding sources support this work, including the Team Science Seed Funding from the Winship Cancer Institute of Emory University, Byron-Davis Research Fund and Leukemia and Lymphoma Translational Research Program Award.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: zhaohui.qin@emory.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

Structural changes to cancer genomes have important effects such as oncogene amplification, tumor suppressor gene disruption, and chimeric fusion gene formation.[1] Structural changes from each of these categories contribute important neoplastic events such as amplification of *HER2* in breast cancer,[2] deletion of *CDKN2A* in a number of different cancers,[3] and formation of fusion genes such as *BCR-ABL* in chronic myeloid and other leukemias.[4] 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.[5–7] 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,[7–18] all are 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.[19] 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 0×2 indicates that the reads are mapped properly, meaning that the reads are correctly oriented with respect to one another, ie, 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 1. Summary of sequencing data.

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

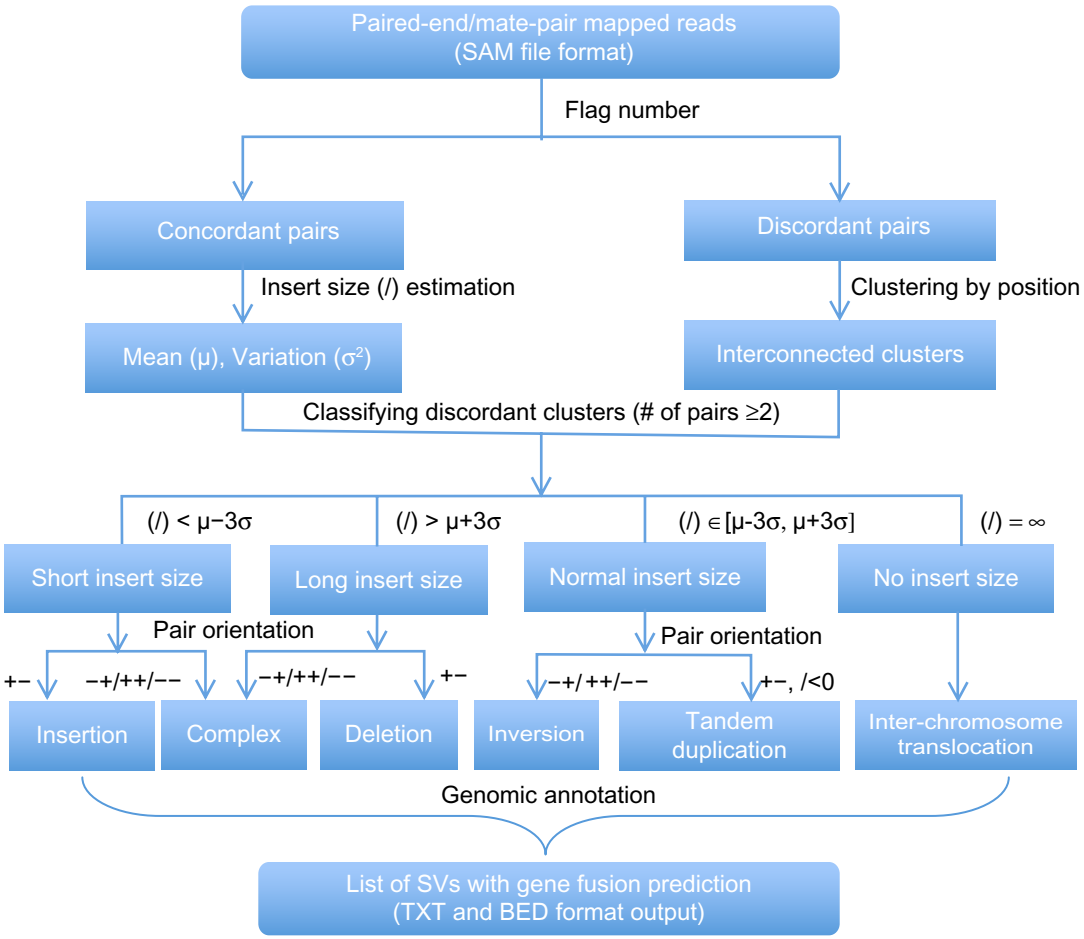

Figure 1. An overview of the SVfinder pipeline. SVfinder takes a SAM file as input and divides the mapped reads into concordant and discordant group. The normal insert size range is estimated from reads in concordant group. The discordant reads are clustered and classified into different types of variation 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).[20]

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

GENE A	GENE B	POSITION A	POSITION B	CHR A	CHR B	KMS11	MM.1S	RPMI8226	TYPE	REFERENCE
WHSC1	IGH	intron	upstream	chr4	chr14	PE			Translocation	20
MAF	IGH	intron	downstream	chr16	chr14		PE		Translocation	20
CDKN2A	CDKN2B	coding	intron	chr9	chr9		PE		Deletion	21
TRAF3	CDC42BPB	coding	intron	chr14	chr14			MP/PE	Deletion	22
SPI1	ZNF287	coding	coding	chr11	chr17			MP/PE	Translocation	Novel
KDM6A	KDM6A	intron	intron	chrX	chrX			MP/PE	Deletion	Novel
KCTD8	KCTD8	intron	intron	chr4	chr4			MP/PE	Deletion	Novel
ABHD17B	C9orf85	coding	intron	chr9	chr9			MP/PE	Deletion	Novel

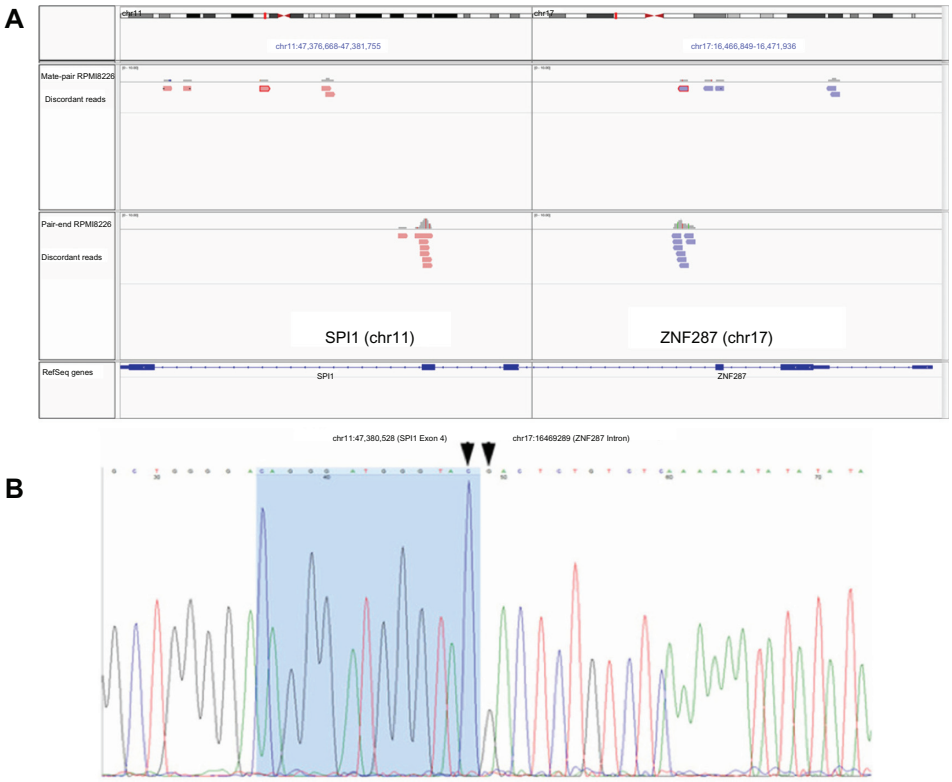

Figure 2. Discovery of novel SPI1 and ZNF287 t(11;17) inter-chromosomal rearrangement in RPMI8226 multiple myeloma cell line. (**A**) Mapped discordant read pairs in mate pair (upper panel) and paired end (lower panel) sequencing shown in integrative genome viewer (IGV). (**B**) Identification of t(11;17) translocation and breakpoint using Sanger sequencing.

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,[21] and CDKN2A deletion[22] found in MM.1S. Additionally, the known TRAF3 deletion[23] 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′-CTCGCCCTCCTC-CTCATCTGAGCT (SPI1) and 3′-AAGGCCATGCAT-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

1. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*. 2007;7(4):233–45.
2. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*. 1987;235(4785):177–82.
3. Cairns P, Polascik TJ, Eby Y, et al. Frequency of homozygous deletion at p16/CDKN2 in primary human tumours. *Nat Genet*. 1995;11(2):210–2.
4. Druker BJ, Tamura S, Buchdunger E, et al. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med*. 1996;2(5):561–6.
5. Greenman CD, Pleasance ED, Newman S, et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res*. 2012;22(2):346–61.
6. Newman S, Howarth KD, Greenman CD, Bignell GR, Tavaré S, Edwards PAW. The relative timing of mutations in a breast cancer genome. *PLoS One*. 2013;8(6):e64991.
7. Ng CK, Cooke SL, Howe K, et al. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J Pathol*. 2012;226(5):703–12.
8. Carrara M, Beccuti M, Lazzarato F, et al. State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int*. 2013;2013:340620.
9. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677–81.
10. Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods*. 2011;8(8):652–4.
11. Quinlan AR, Clark RA, Sokolova S, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*. 2010;20(5):623–35.
12. Zeitouni B, Boeva V, Janoueix-Lerosey I, et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*. 2010;26(15):1895–6.
13. Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol*. 2010;11(12):R128.
14. Xi R, Kim TM, Park PJ. Detecting structural variations in the human genome using next generation sequencing. *Brief Funct Genomics*. 2010;9(5–6):405–15.
15. Dalca AV, Brudno M. Genome variation discovery with high-throughput sequencing data. *Brief Bioinform*. 2010;11(1):3–14.
16. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12(5):363–76.
17. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*. 2009;6(11 suppl):S13–20.
18. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*. 2011;27(20):2903–4.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
20. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–92.
21. Pratt G. Molecular aspects of multiple myeloma. *Mol Pathol*. 2002;55(5):273–83.
22. Kuehl WM, Bergsagel PL. Multiple myeloma: evolving genetic events and host interactions. *Nat Rev Cancer*. 2002;2(3):175–87.
23. Bergsagel PL. TRAF3 in B cells: too much, too little, too bad. *Blood*. 2009;113(19):4481–2.