SVAtools for junction detection of genome-wide chromosomal rearrangements by mate-pair sequencing

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Sarah H Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

George Vasmatzis

June 2017

Dedication

To Mr. Schaberg, who made learning science fun, in the 6th grade and for a lifetime.
"Hello boys and girls, let's learn!"

Abstract

Mate-pair sequencing (MPseq), using long-insert, paired-end genomic libraries, is a powerful next-generation sequencing-based approach for the detection of genomic structural variants. SVAtools is a set of algorithms to detect both chromosomal rearrangements and large (>10kb) copy number variants (CNVs) in genome-wide MPseq data. SVAtools can also predict gene disruptions, gene fusions, and characterize the genomic structure of complex rearrangements.

To illustrate the power of SVAtools' junction detection methods to provide comprehensive molecular karyotypes, MPseq data was compared against a set of samples previously characterized by traditional cytogenetic methods. Karyotype, fluorescence in situ hybridization (FISH) and chromosomal microarray (CMA), performed for 29 patients in a clinical laboratory setting, collectively revealed 285 breakpoints in 87 rearrangements. The junction detection methods of SVAtools detected 87% of these breakpoints compared to 48%, 42% and 57% for karyotype, FISH and CMA respectively. Breakpoint resolution was also reported to 1 kb or less and additional genomic rearrangement complexities not appreciable by standard cytogenetic techniques were revealed. For example, 63% of CNVs detected by CMA were shown by SVAtools' junction detection to occur secondary to a rearrangement other than a simple deletion or tandem duplication. SVAtools with MPseq provides comprehensive and accurate whole-genome junction detection with improved breakpoint resolution, compared to karyotype, FISH, and CMA combined. This approach to molecular karyotyping offers considerable diagnostic potential for the simultaneous detection of both novel and recurrent genomic rearrangements in hereditary and neoplastic disorders.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and background

Structural variants (SVs), including balanced and unbalanced chromosomal rearrangements and copy number variants (CNVs) are a significant contributor to both neoplastic and hereditary disorders. Germline SVs can produce an adaptive advantage, but most often have a negative consequence leading to aneuploidy, infertility or disease; cancer is often a result of somatic SVs (Feuk, Carson, and Scherer 2006; Stankiewicz and Lupski 2010; Weischenfeldt et al. 2013). Characterizing and understanding the mechanisms of SVs is dependent on the ability to accurately detect SVs (Weckselblatt and Rudd 2015; Chen et al. 2010; South 2011).

SVs are routinely detected by standard cytogenetic methods: karyotype, fluorescence in situ hybridization (FISH) and genomic copy number microarrays, or chromosomal microarrays (CMA). Karyotyping by conventional, G-banded, giesma stain to identify banding patterns on chromosomes, provides a whole-genome analysis and is suitable for detecting large rearrangements. However, karyotyping requires dividing cells arrested in metaphase, which can be difficult to retrieve from some sample sources. Additionally, because metaphase chromosomes are highly condensed, karyotyping has variable and relatively low resolution, at best 3-10 Mb. FISH involves fluorescently labeled DNA probes which hybridize to a genomic region (~200 kb) of interest. The resulting fluorescent patterns on interphase or metaphase preparations reveal genomic regions of increased or decreased copy number and/or specific rearrangements. The utility of FISH is primarily limited by the *a priori* need to know which genomic region to interrogate and limited to the number of probes that can be multiplexed in a single analysis. CMA, through various methodologies, provides a whole-genome copy number scan by comparing the relative ratio of patient DNA to control. For most CMA assays in clinical use, CMA has the ability to detect regions of copy number change at a resolution of approximately 25-400 kb throughout the genome, down to <10 kb for targeted regions, depending on the number of probes on the array (Kearney et al. 2011). The primary limitation of CMA is the inability to detect balanced rearrangements or clarify the structural configuration of detected CNVs, e.g., a copy number gain may be the result of a direct tandem duplication, inverted duplication, or an insertional translocation elsewhere, and making this distinction may have significant interpretive consequence (Newman et al. 2015).

Whole-genome, read-pair next generation sequencing (NGS) acquires the DNA sequence of an organism. DNA is fragmented, sequenced from both ends, and mapped to a reference genome.

Mate-pair sequencing (MPseq) represents a variation of standard paired-end sequencing utilizing large genomic inserts, 2-5 kb, which are circularized and fragmented to the standard paired-end protocol length, 200-500 bps, Figure 1.1. Both MPseq and standard paired-end are whole genome, can provide base pair resolution (given adequate sequencing depth) and are capable of detecting balanced and unbalanced rearrangements (Korbel et al. 2007; Alkan, Coe, and Eichler 2011). MPseq however, offers additional advantages over standard paired-end sequencing. The larger fragments improve mapping accuracy in repeat areas (Medvedev, Stanciu, and Brudno 2009) and allow for significantly more efficient capture of junctions associated with SVs, thus providing a less expensive technique for SV detection.

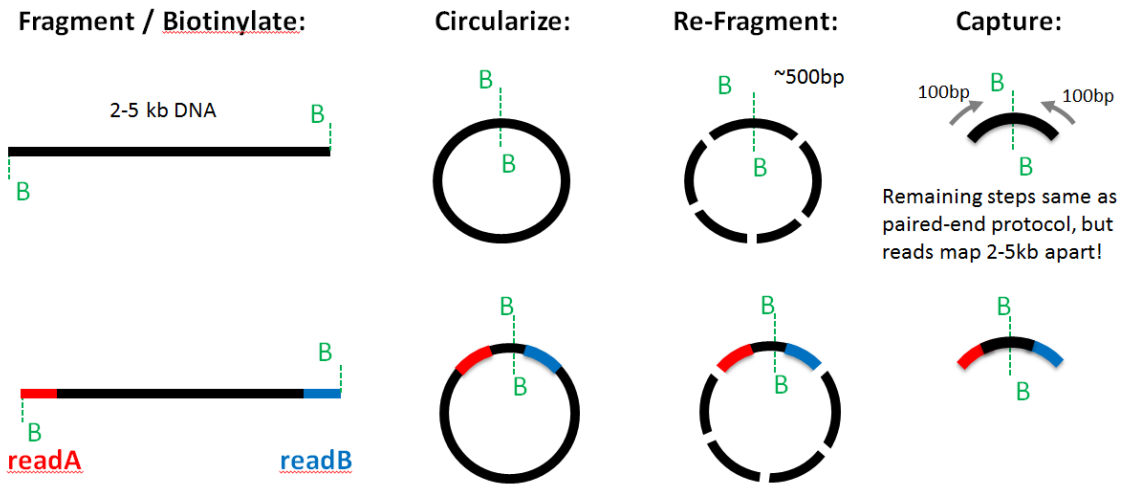


Figure 1.1: MPseq fragment circularization
DNA is prepared for MPseq using the Nextera mate-pair library protocol and sequenced on the Illumina HiSeq platform. To make mate-pair libraries, DNA is fragmented to 2-5kb pieces and the terminal ends are labeled with biotin (B). These fragments are circularized and re-fragmented to smaller ~500 bps pieces. A capture step selects the biotin labeled fragments and these ends are then sequenced. The red and blue segments represent the 101 bps reads that are sequenced, and the mapped strand orientation of that read, red for reverse strand, blue for forward strand. This long insert size gives MPseq the ability to detect structural variants with less sequencing (cheaper) than traditional whole genome sequencing with paired-end sequencing.

The SV detection pipeline established by Biomarker Discovery (BMD) at Mayo Clinic (*http://mayoresearch.mayo.edu/center-for-individualized-medicine/biomarker-discovery-program.asp*) includes guidelines for DNA preparation, library preparation, multiplex sequencing and sequencing analysis with BIMA and SVAtools. The BMD-SV pipeline has demonstrated considerable utility since its first implementation in 2011 (Feldman et al. 2011; Vasmatzis et al. 2012; Murphy et al. 2012; Murphy et al. 2014; Boddicker et al. 2016; Harris et al. 2016 ; Catic et

al. 2017). The BMD-SV pipeline does not require matched tumor/normal samples and is suitable for calling SVs from fresh or frozen DNA collected from solid tumors and soft tissue cancers, and blood or bone marrow sampled from constitutional disorders or hematological malignancies.

This thesis introduces SVAtools and describes the junction detection algorithm ($SVAtools_{JD}$) for analyzing MPseq data. The sensitivity and junction detection performance of $SVAtools_{JD}$ is compared to traditional cytogenetic techniques. Also demonstrated are the additional advantages of MPseq and $SVAtools_{JD}$ over traditional cytogenetic techniques, including ability to: a) resolve the genomic architecture of the SVs detected; b) predict gene fusions and gene truncations; and c) obtain high resolution (~200 bp) of breakpoints. The CNV detection of SVAtools (Smadbeck et al.) would improve the overall sensitivity of MPseq, but was out of scope for this thesis

Chapter 2: Methods

2A: Terminology and junction signatures

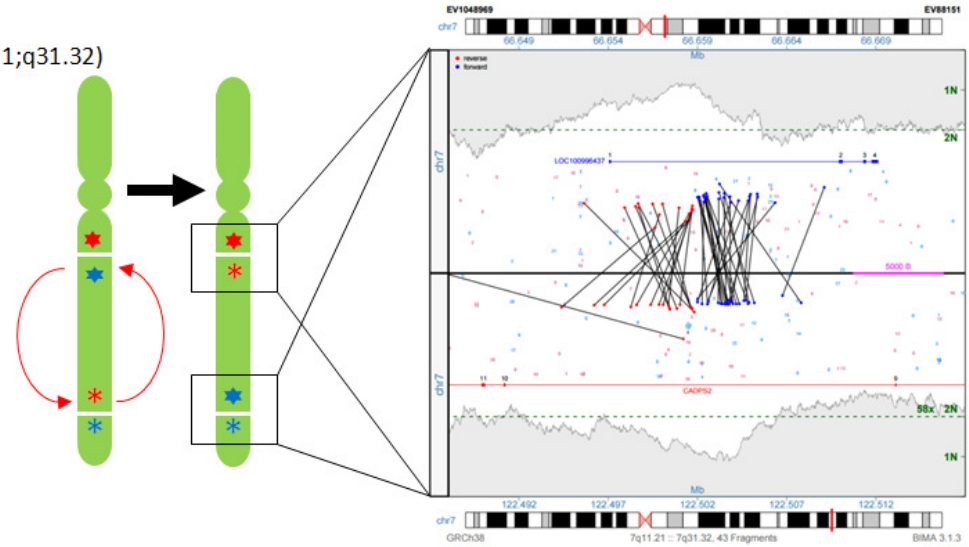
Structural variants include rearrangements and CNVs. Rearrangements, such as inversions, deletions, tandem duplications, and translocations, involve a number of breaks and reunions. To facilitate discussion, we use the following nomenclature:

- Breakpoint: location left or right of a break in the genome
- Junction: the reunion of a break, two distal breakpoints are now adjacent
- Rearrangement: chromosome structure abnormality due to one or more junctions, such as a deletion, inversion, translocation, etc.
- Bridged Coverage: average number of fragments (read-pairs) spanning a position in the diploid genome

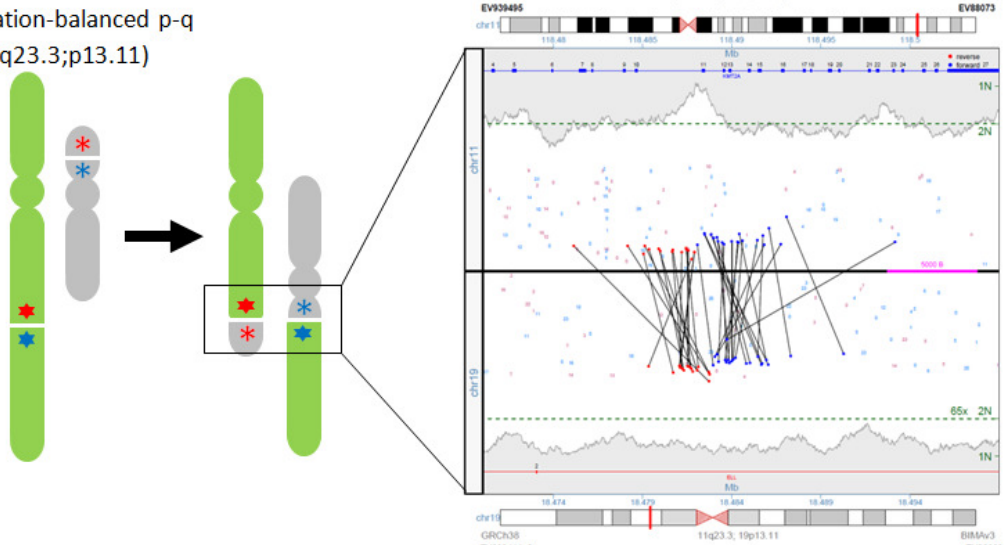
The breakpoints and junctions of non-complex rearrangements are illustrated in Figure 2A.1. For each rearrangement, a chromosome schematic is shown adjacent to an example junction plot produced by SVAtools.

A

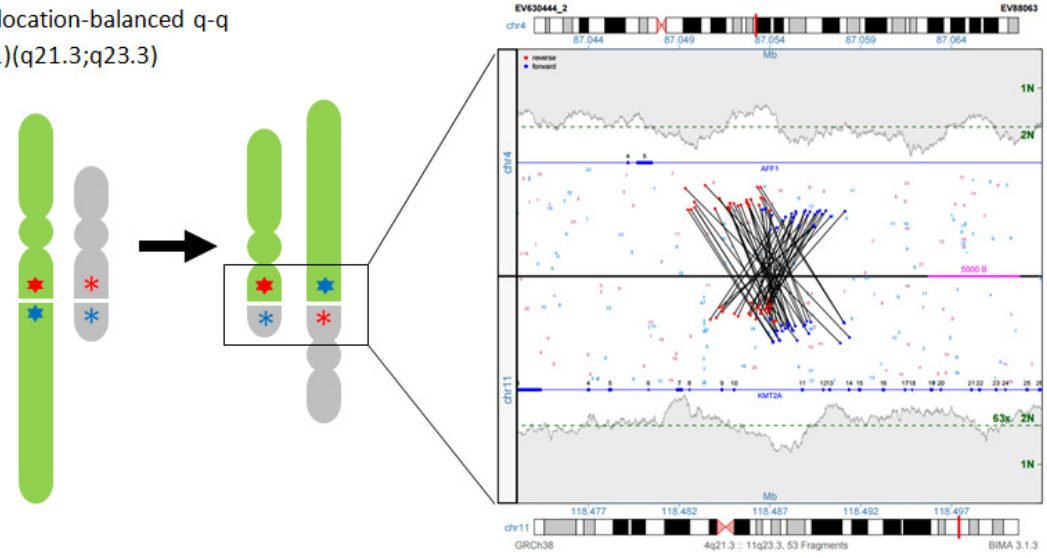
Inversion:
inv(7)(q11.21;q31.32)



Translocation-balanced p-q
t(11;19)(q23.3;p13.11)

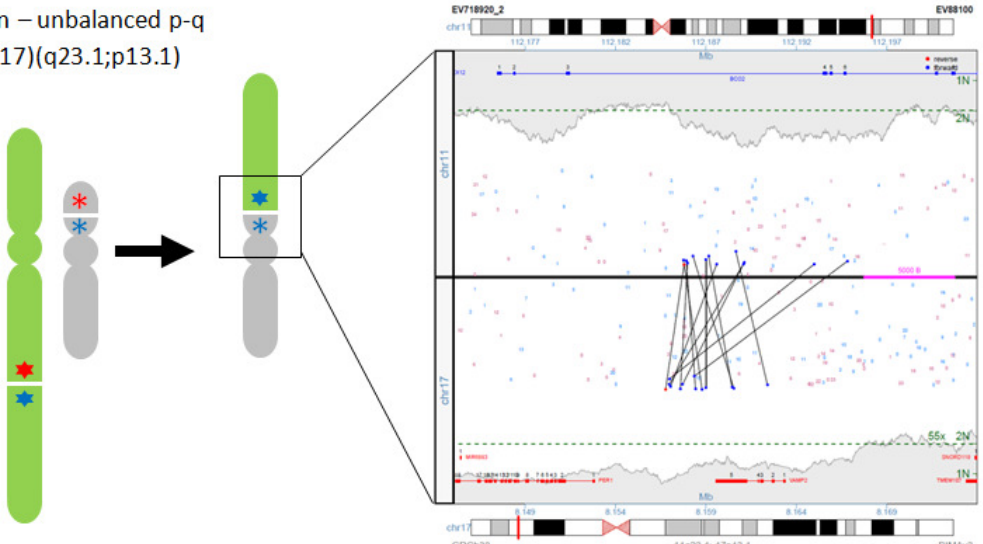


Translocation-balanced q-q
t(4;11)(q21.3;q23.3)

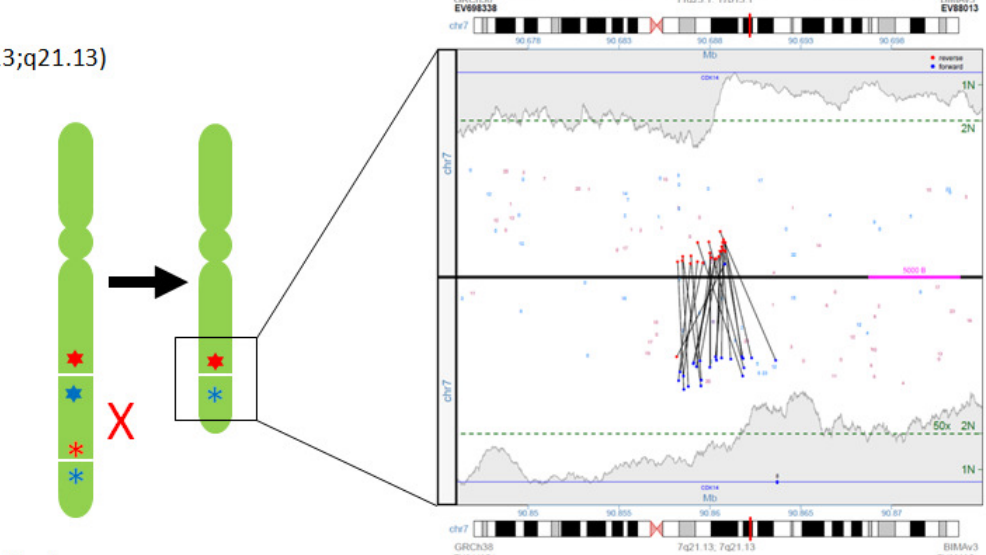


B

Translocation – unbalanced p-q
der(17)t(11;17)(q23.1;p13.1)



Deletion:
del(7)(q21.13;q21.13)



Tandem Duplication:
dup(2)(q22.1;22.1)

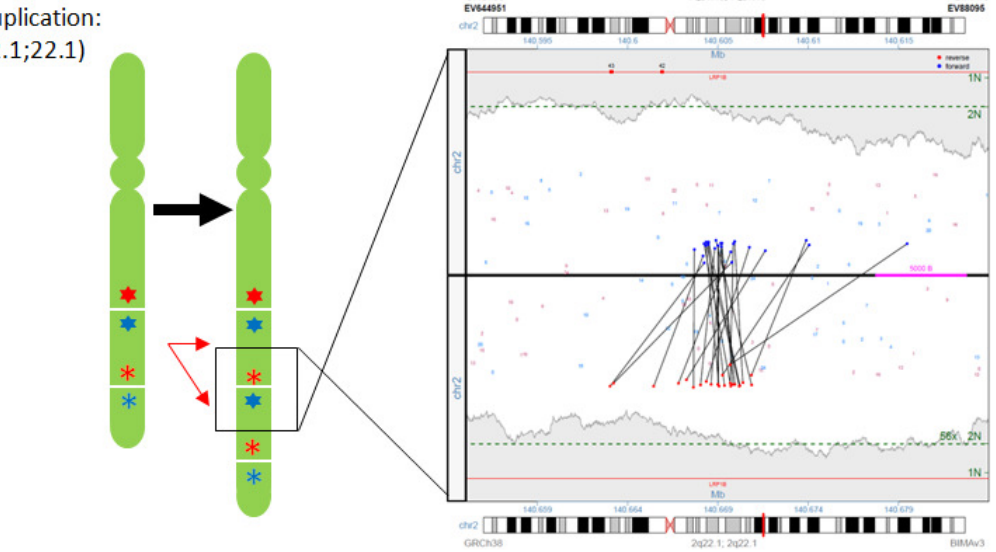


Figure 2A.1: Junction signatures for non-complex rearrangements
Rearrangements occur due to one or more breaks in the genome. Junctions are the reunion of the breaks. Each junction will have two breakpoints (stars), one on each side of the break and reunion. Junction plots have a distinct pattern indicative the rearrangement type. Non-complex rearrangements are depicted by the chromosome schematic (left) and junction plot (right). The orientation of the junction in the chromosome schematic (black box), corresponds directly to the orientation of the junction in the junction plot. A) Rearrangements with four breakpoints and two junctions include: inversions and reciprocal translocations. B) Rearrangements with two breakpoints and one junction include: derivative translocations, deletions and duplications.

By the break, reunion, and junction convention defined here, the terminal edge of a terminal CNV was not included in the breakpoint count and the non-terminal edge was included only if the CNV occurred with a junction. For example, the centromeric breakpoint of a whole arm CNV due to a translocation was included, while whole chromosome CNVs, such as monosomy and trisomy, were excluded. While not discussed here, terminal CNVs are detectable by SVAtools' CNV detection methods; full details will be provided in the upcoming paper by Smadbeck et al.

Coverage estimations provide thresholds to confidently call variants. Base coverage (often referred to as depth of coverage, read depth, or coverage) is dependent on the count and length of reads sequenced, and therefore, on total sequenced nucleotides. SVAtools uses a count of the number of fragments spanning a given position, "bridged coverage" to establish confidence for each SV detected. Bridged coverage will depend on the number of fragments and the length the fragments (read lengths plus insert length). Figure 2A.2 illustrates a pile-up of 16 fragments mapped to a reference. For the given position X, the bridged coverage is 14x, while the base coverage is only 2x. Because the insert size is large, often 2-5kb, the bridged coverage will be much higher than the base coverage. For example, given 100,000,000 fragments with 2 reads of length 100bp and assuming uniform mapping to GRCh38 (3.2×10^9 bp), base coverage is 6.25x. If the insert size is 2800, for a total fragment length of 3000, the bridged coverage is 93.75x, a 15-fold increase for the same amount of sequencing.

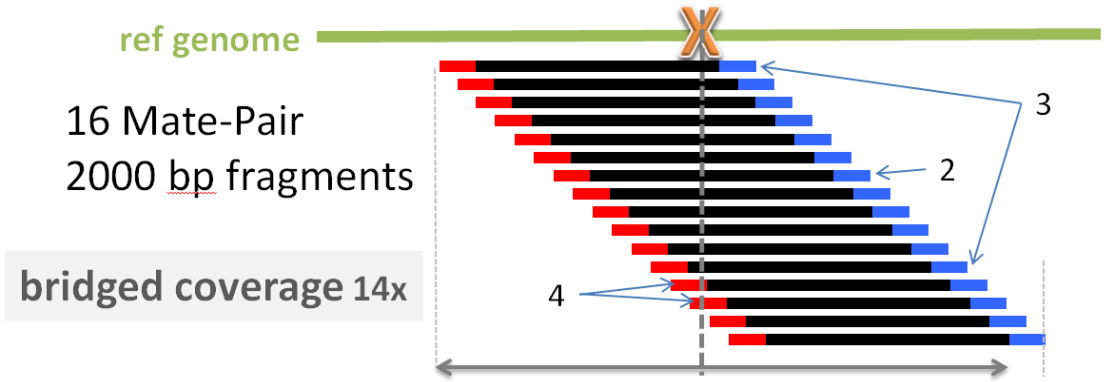


Figure 2A.2 Schematic pile up of 16 mate-pair fragments mapped to a reference. The red and blue segments of each fragment represent the two mapped reads, black is the unsequenced and inferred insert region. The number of fragments spanning position X determines the bridged coverage of that position. To determine the resolution of the breakpoint position based on bridged coverage consider: 1) the breakpoint will be in-between two discordant reads from the same fragment, thus the breakpoint must be no further than the fragment length; 2) in a worst case scenario the breakpoint is exactly in the middle of a spanning fragment, therefore the breakpoint accuracy is at least half the fragment length; 3) odds are the breakpoint will be close to one read from one fragment and close to another read from a different fragment, providing accuracy within a couple hundred bps; 4) SVAtools$_{JD}$ includes split reads in the junction detection; when available these split reads will contain the actual breakpoint.

2B: Cytogenetic Analysis

A test set of representative breakpoints were identified from the results of three cytogenetic methods. The sample selection included patients referred to the Mayo Clinic Cytogenetics Laboratory for karyotype and/ or FISH testing of hematologic neoplasms such as acute myeloid leukemia, acute lymphocytic leukemia and chronic lymphocytic leukemia. All samples were processed using standard clinical protocols for the three methods: karyotype, FISH and CMA (Affymetrix CytoScan HD array, analyzed using ChAS and manually reviewed). Two samples were excluded because they did not have at least one rearrangement from at least one of the three methods. One sample with an exceedingly complex genomic structure, 42-45,XX,psu dic(9;12)(p13;p11.2)[4],add(11)(p11.2),-16,-17[4],-19[3], +der(?)t(?;16)(?;q11.1),+1-3mar[cp5]//46,XY[15] was also excluded as it was not feasible to delineate specific rearrangements for an organized point by point technical comparison in the way chosen for this report. All clinically reported findings for each method were included in this report (Table 1). The final sample selection of 29 cases included a range of typical and straightforward cases to cases with challenging and complex clonal rearrangements.

Table 2B: Clinically reported results for karyotype, FISH, and CMA for 29 representative samples of hematologic neoplasm.

Sample	Reason for Referral	Karyotype	Abnormal FISH	CMA (converted to GRCh38)
EV88044	AML	46,XX,t(15;17)(q22;q21)[17]/46,XX,der(15)t(15;17)(q22;q21),ider(17)(q10)t(15;17)(q22;q21)[3]	(PMLx3),(RARAx3),(PML con RARAx2)[398/500]/ (PMLx4),(RARAx4),(PML con RARAx3)[79/500]	15q24.2q26.3(75655926_101991189)x3[0.15] 17p13.3p11.2(150732_19941963)x1[0.15] 17p11.2q12(22266396_39610173)x3[0.15]
EV88048	Myelofibrosis	46,XX[20]		2q31.3q32.1(181249078_183436911)x3
EV88059	CLL	46,XY,t(5;6)(q13;q23)[10]/47,XY,+12[2]/46,XY[8]	(D13S319x1),13q34(LAMP1x2)[112/200]	6q14.3q22.1(85398516_116240537)x1[0.4], 6q25.2q25.3(154631808_157214436)x1[0.4], 13q14.2q14.2(47898938_49957631)x1[0.5], 13q14.2q14.3(50108920_50922184)x1[0.5]
EV88060	CLL	46,Y,inv(X)(p?21q?21)?c[13]	(D13S319x0),13q34(LAMP1x2)[158/200]	13q14.2q14.3(49925862_50516172)x1[1.7], 13q14.3q14.3(50519349_51037765)x1[0.5]
EV88061	CLL	46,XY,der(2)t(2;7)(q31;q32),der(7)add(7)(p13)t(2;7)[2]/46,XY[15]	(D13S319x1),13q34(LAMP1x2)[107/200]	2p23.1p22.3(31802162_32393289)x1[0.3] 2p13.1p13.1(73555116_74697219)x1[0.4] 7p21.3p21.3(7398184_8546126)x1[0.4] 13q14.11q14.11(40904307_43361934)x1[0.25] 13q14.11q14.2(43368704_48382057)x1[0.5] 13q14.2q14.3(48386316_51084046)x1 13q14.3q21.1(51085062_56092156)x1[0.5] 13q21.1q21.32(56104664_66805369)x1[0.25]
EV88063	ALL	46,XX,t(4;11)(q21;q23)[10]/46,idem,i(7)(q10)[10]	(AFF1x3),(KMT2Ax3),(AFF1con KMT2A)[396/500]	(7p)x1[0.3] (7q)x3[0.3] 9p21.3p21.3(21178271_22202761)x1[1.5]
EV88064	acute leukemia	46,XX[8]		6p12.1p12.1(53161778_53731152)x3[1.25] 6p11.2p11.2(57965290_58306189)x3[0.75]
EV88065	CML	46,XY,t(9;22;15)(q34;q11.2;q15)[20]	(ABL1x3),(BCRx3),(ABL1 con BCRx1)[464/500]	18p11.32p11.32(1194375_1640812)x1
EV88070	MPD with eosinophilia	46,XY[20]	(FIP1L1x2,CHIC2x1,PDGFRAx2)[133/200]	4q12q12(53427547_54274695)x1[0.6]
EV88073	AML with monocytic different-iation	46,XX,t(11;19)(q23;p13.1)[20]	(KMT2Ax3),(ELLx3),(KMT2A con ELLx2)[491/500]	12q24.23q24.23(118516990_118588890)x1[0.4] 16q23.1q23.1(78229384_78352174)x1[0.8] 17q25.1q25.1(73839502_74684979)x3[1.2]

Sample	Reason for Referral	Karyotype	Abnormal FISH	CMA (converted to GRCh38)
EV88074	CML	46,XY,t(11;13)(q23;q14)[17]/46,XY[3]	(D13S319x1,LAMP1x2)[126/200]	11q23.3q23.3(119067603_121113378)x1[0.7] 13q14.2q14.3(47767418_51606370)x1[0.7] 18q21.2q21.31(55080639_56205472)x1[0.7] 22q11.22q13.2(23101973_43489488)x1[0.1]
EV88076	new aggressive lymphoma	46,XY[5]	(MYCx3),(IGHx2),(MYC con IGHx1)[219/500]	1q21.1q25.3(146149859_184080669)x3[0.3]
EV88078	leukocytosis	46,XX,t(9;22)(q34;q11.2)[20]	(ABL1x3),(BCRx3),(ABL1 con BCRx2)[486/500]	arr(1-22,X)x2
EV88079	post transplant lymphopro-liferative disorder	46,XY,t(9;11)(p22;q23)[20]	(MLLT3x3),(KMT2Ax3),(MLLT3 con KMT2Ax2)[465/500]	arr(1-22)x2,(X,Y)x1
EV88081	non-hodgkin lymphoma	46,XX[20]	(AFF1x3),(KMT2Ax3),(AFF1 con KMT2Ax1)[167/500]	1p36.33p36.23(1959099_7949703)x1[0.4] 1p13.1p13.1(115806740_116739470)x1[0.5] 4p15.1p15.1(30992588_31966163)x1[0.4] 6q14.1q22.1(80892556_114303640)x1[0.5] 9p24.3p21.3(203861_21572243)x1[0.5] 9p21.3p21.3(21579259_22075597)x1[0.8] 9p21.3p21.2(22081850_26050211)x1[0.5] 12q13.13q13.13(52294886_52388067)x3 15q22.2q22.2(61453278_61592715)x3 17p13.3p13.3(150732_2931676)x1[0.5] 17p13.3p13.3(2936307_3391011)x1[0.8] 17p13.2p11.2(3397639_21542019)x1[0.5]
88083	anemia, thrombo-cytopenia, AML	46,XX,t(6;9)(p23;q34)[8]/46,XX[1]	(DEKx3),(NUP214x3),(DEK con NUP214x2)[480/500]	6p22.3p22.3(18119418_18226579)x1[0.75] 9q34.13q34.13(131132438_131153811)x1[0.5]
88085	leukocytosis	46,Y,t(X;10;9;22)(q11;p13;q34;q11.2)[19]/47,idem,+8[1]	(ABL1x3),(BCRx3),(ABL1 con BCRx1)[405/500]	arr(1-22)x2,(X,Y)x1
EV88086	lymphoma	46,XY,t(2;11)(p25;q23)[1]/46,idem,-Y,+8[2]/46,XY[2]	(CDKN2Ax1,D9Z1x2)[98/200]/(3'IGHx2,dim5'IGHx2)[104/200], ish t(2;11)(KMT2A+;KMT2A-)[2]	9p21.3p21.3(21817336_21968663)x1[0.5] 9p21.3p21.3(21976746_22005383)x1[0.75] 9p21.3p21.3(22009308_22104160)x1[0.5]

Sample	Reason for Referral	Karyotype	Abnormal FISH	CMA (converted to GRCh38)
				11q22.1q22.1(99772315_100264159)x3[0.5] 11q22.3q23.3(103229656_115453069)x1[0.5] (8)x3[0.2] (Y)x0[0.5]
EV88088	AML	47,XY,+8,t(9;11)(p22;q23)[20]	(MLLT3x3),(KMT2Ax3),(MLLT3 con KMT2Ax2)[488/500]	9p22.1p22.1(19072482_19320487)x3 (8)x3 (Y)x2[0.5]
EV88089	granulocytic sarcoma	46,X,-Y,t(8;21)(q22;q22),+9,del(9)(q13q32)x2[5]/46,XY[15]	(RUNX1T1x3),(RUNX1x3),(RUNX1T1 con RUNX1x2)[31/500]/(ABL1x3,BCRx2)[18/500]	9q21.11q31.2(68351345_107847887)x1[0.1] (Y)x0[0.2]
EV88090	CMPD	46,XX,t(9;22)(q34;q11.2)[19]/46,XX[1]	(ABL1x3),(BCRx3),(ABL1 con BCRx2)[469/500]	9q34.11q34.12(130419555_130717717)x1[0.75] 22q11.23q11.23(23293899_23414891)x1[0.75]
EV88091	Acute leukemia, bilineage, T-cell and myeloid	46,XY,del(9)(q13q22)[9]/46,XY[11]	(TLX3x3),(BCL11Bx2),(TLX3 con BCL11Bx1)[424/500]	9p21.3p21.3(21828043_21996864)x1[0.75] 9q21.13q31.2(71857107_107192688)x1[0.5]
EV88094	AML	48,XY,t(1;3)(q32;p25),add(6)(p21.3),+13,inv(16)(p13.1q22), der(18)t(1;18)(q21;q21),+20[20]	(MYH11x3),(CBFBx3),(MYH11 con CBFBx2)[214/500]/ (D13S319,LAMP1)x3[93/200]/(D20S108,20qter)x3[91/200]	1q24.3q44(171641445_249000000)x3[0.3] 3p25.2p25.2(12582218_12675079)x1 7q34q34(142251693_143399068)x1[0.5] 11q22.1q25(100524067_135068576)x3[0.3] (13)x3[0.3] 18q22.2q23(69603144_80256240)x1[0.3] (20)x3[0.3]
EV88096	acute leukemia	46,XY,inv(2)(p11.2q13)[5]/46,XY[15]		2q14.2q21.1(120599518_130388693)x1[0.3] 8q24.13q24.21(125355953_129686186)x1[0.3]
EV88099	T-ALL	48,X,t(X;10)(q24;p13),+1,+4[6]/46,XX[14]	(STIL,TAL1)x3[125/200]/(MLLT10x3)[355/500]	(1)x3[0.6] (4)x3[0.6] 6q23.3q23.3(135091879_135409387)x3 16q22.1q22.1(66990865_67774694)x1[0.6]
EV88100	acute leukemia, MDS	46,XX,inv(3)(q21q26.2),del(5)(q13q33),add(17)(p13)[20]	(RPN1x3),(EVI1x3),(RPN1 con EVI1x2)[454/500]/(D5S630x2,EGR1x1)[198/200]/(MLLx3)[185/200]	5q14.2q33.2(82326551_153894589)x1[0.9] 11q23.1q25(112186582_135068576)x3[0.9] 17p13.3p13.1(150732_8158478)x1[0.8]

Sample	Reason for Referral	Karyotype	Abnormal FISH	CMA (converted to GRCh38)
EV88101	anemia, fever, bone pain	46,XX,del(11)(q21q23),t(11;15)(q11;q11.2),add(19)(q13.1)[2]/ 47,idem,t(5;9)(q11.2;p22),+21[1]/46,XX[17]	(CDKN2Ax0,D9Z1x2)[20/200]/(ETV6x2),(RUNX1x3),(ETV6 con RUNX1x2)[387/500]/(ETV6x2),(RUNX1x4),(ETV6 con RUNX1x2)[30/500]	2p11.2p11.2(88600364_88829551)x1 9p22.1p21.3(19589909_21800760)x1[0.3] 9p21.3p21.3(21807994_22492877)x0[0.5] 9p21.3p21.3(23971816_25466630)x1[0.2] 12p13.2p13.1(11618063_13133213)x1 17q11.1q11.2(27344679_30233516)x1[0.1] 17q11.2q12(30233516_37700251)x1[0.6] 17q12q23.2(37710931_62716038)x3[0.5] 17q23.2q24.1(62718712_66002488)x1[0.6] (21)x3[0.2]
EV88102	B-ALL	46,XX[20]	(CDKN2Ax1,D9Z1x2)[40/200]	2p14p11.2(64635266_88825975)x1[0.8] 6q14.1q22.1(78827961_115307126)x1 6q24.2q25.1(143704644_149826658)x1 9p22.1p21.3(19625174_21767405)x1[0.2] 9p21.3p21.3(21772930_22215463)x1[0.4] 12p13.2p13.2(11639155_11747039)x1 12q21.33q21.33(91813602_92157567)x1 16p13.3p13.3(3777822_3867646)x3 17p13.3p13.1(159683_8094096)x1[0.7] (X)x1[0.7]
EV88103	lymphoma	45,X,-X, dup(1)(q11q32), t(3;7)(q27;q22)[15],t(8;14;18)(q24.1;q32;q21.3),der(14)t(8;14;18)[2][cp20]	(MYCx3),(IGHx4),(MYC con IGHx2)[445/500]/(IGHx4),(BCL2x3),(IGH con BCL2x2)[424/500]	1q21.1q24.2(144009402_168910115)x3[0.75] 1q24.2q24.2(168915824_169888200)x1[0.25] 1q24.2q41(169892108_222802457)x3[0.75] 2p16.1p15(59929961_63403720)x3[0.9] 9p21.3p21.3(21604467_22258903)x1[1.6] 12q24.31q24.31(121574607_122813944)x1[0.75] 16q23.1q23.1(78701427_78967038)x1[0.75] 18p11.31p11.23(5594352_7718922)x1[0.2] 22q13.1q13.2(39447775_40832804)x1[0.8] (X)x1[0.8]

2C: MPseq analysis, SVAtools and the junction detection algorithms

2C-1: BMD-SV Pipeline Overview

DNA from the 29 representative clinical cytogenetic samples was tested by the BMD-SV pipeline. Bulk-extracted DNA libraries for each sample were prepared following the Illumina Nextera Mate-Pair Library Prep protocol (Illumina, San Diego, CA, USA). Sequencing was performed two samples per lane on the Illumina HiSeq2000, with 101-basepair (bp) read length. The sequencing data was mapped to GRCh38 by BIMA (Drucker et al. 2014). SV detection was performed by SVAtools. Figure 2C.1 provides sequencing statistics for each of the 29 samples. Average bridged coverage for the 29 samples was 60x. Figure 2C.2 provides a histogram of fragment lengths for a typical MPseq library. The algorithmic workflow for the BMD-SV pipeline is shown in Figure 2C.3.

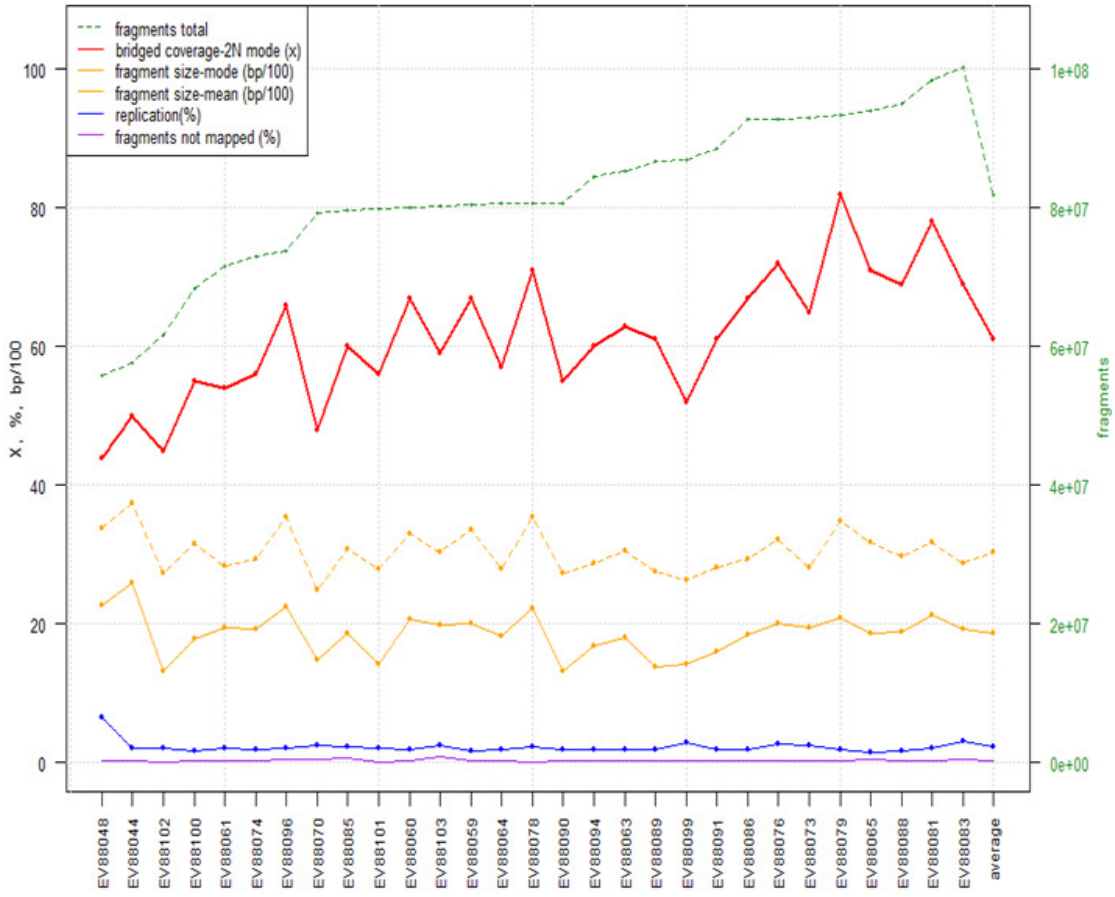


Figure 2C.1: MPseq quality statistics for the 29 samples.
Average bridged coverage was 60x. The high bridged coverage is a result of the large number of fragments (average ~80 million/ sample), low % replication, low % unmapped fragments, and the large fragment size (mode = ~2000 bp, mean =~3000bp).

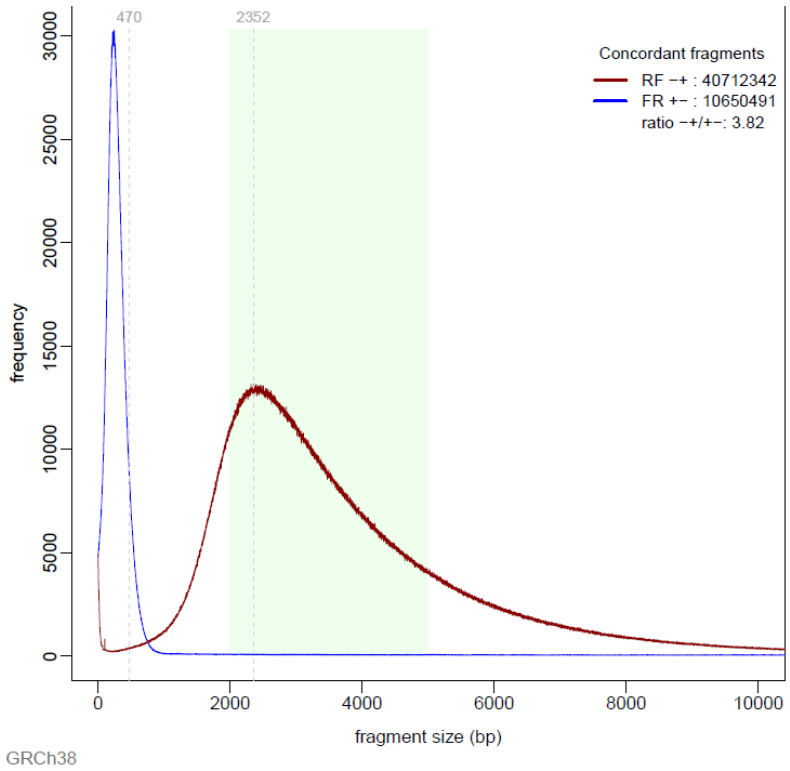


Figure 2C.2: Fragment length histogram of a typical MPseq library
Fragment length versus frequency. The red line peaking at 2352 bps indicates the frequency of reverse-forward mapping fragments, typically mate-pair fragments, given the reversal of orientation of the reads due to the circularized intermediate in the library preparation. The blue line indicates the frequency of forward-reverse mapping fragments, typically paired-end fragments. Paired-end fragments exist in our MPseq library preparations at a typical ratio of 3:1.

2C-2: Mapping - BIMA to Map MPseq libraries

BIMA was previously published by BMD (Drucker et al. 2014). BIMA converts the DNA alphabet to binary (the computer alphabet) using three encodings: A/G, C/G, and A/C, where the two listed bases are converted to 1 and the remaining two bases are converted to 0. BIMA converts each sequential 32 bases to a 32 bit binary number, which is then used as an index to a series of hash tables and look up tables. This fast lookup algorithm, combined with binary operations, enables BIMA to map 100 million mate-pair fragments to the 3.2 billion base reference genome faster than other algorithms (Drucker et al. 2014). The conversion of the reference genome and the reads from each fragment allows each base to be uniquely identified in binary. This conversion also allows for single nucleotide variants without loss of information.

BIMA maps both reads from a fragment concurrently to facilitate more accurate mapping, such as when one read maps to multiple positions, the other read can serve as an anchor to determine the correct position of the former. BIMA scores each mapping option and reports the position with the best score. When the same best score occurs for multiple mapping options, BIMA defaults to the position that occurs first by numerical order of the chromosome and then chromosome position. When BIMA detects a split read, two alignments will be reported. If BIMA cannot map a read, or the read maps to too many places, BIMA will map the read to chromosome 0, position 0. These fragments remain available for in-depth downstream processing. Since over 5% of the GRCh38 reference genome is not sequenced, at least 5% of all fragments should not map. On average, less than 5% of all fragments cannot be mapped by BIMA, thus there is some over-mapping by BIMA.

BIMA reports for each read the chromosome (chr), position (pos), and mapping metrics to facilitate junction detection by SVAtools$_{JD}$. BIMA nomenclature designates readA and readB of a read-pair in numerical order of the mapped position. Thus, chrA/posA for readA will always be the lower numbered chromosome and the lower position compared to chrB/posB for readB. BIMA can map to any reference genome of any length with any number of chromosomes. In this implementation, BIMA maps to GRCh38, which includes chromosomes X, Y, 1-22, plus the unlocalized, unplaced, and alternate sequences. For other applications, such as for detecting virus integration into the human genome (Gao et al. 2014), BIMA mapped to a reference genome created by concatenating GRCh38 with genomes from over 5000 viruses.

BIMA supports two file options for reporting mapping results: 1) .sva for importing into SVAtoolsJD and 2) SAM. A .sva file is generated for each chromosome in the reference genome. This sorting facilitates faster downstream processing than SAM files because: 1) SVAtools$_{JD}$ can parallel process data for each chromosome; and 2) related fragments are stored in smaller files and closer to each other, crucial for on demand visualization.

BIMA's flexibility allows for any insert size, mapping both paired-end and mate-pair fragments, important as both are present in a mate-pair library. Furthermore, BIMA is specially tuned to handle split reads which occur when the read crosses the breakpoint or crosses through the biotin junction of mate-pair libraries (up to 20% of mate-pair fragments).

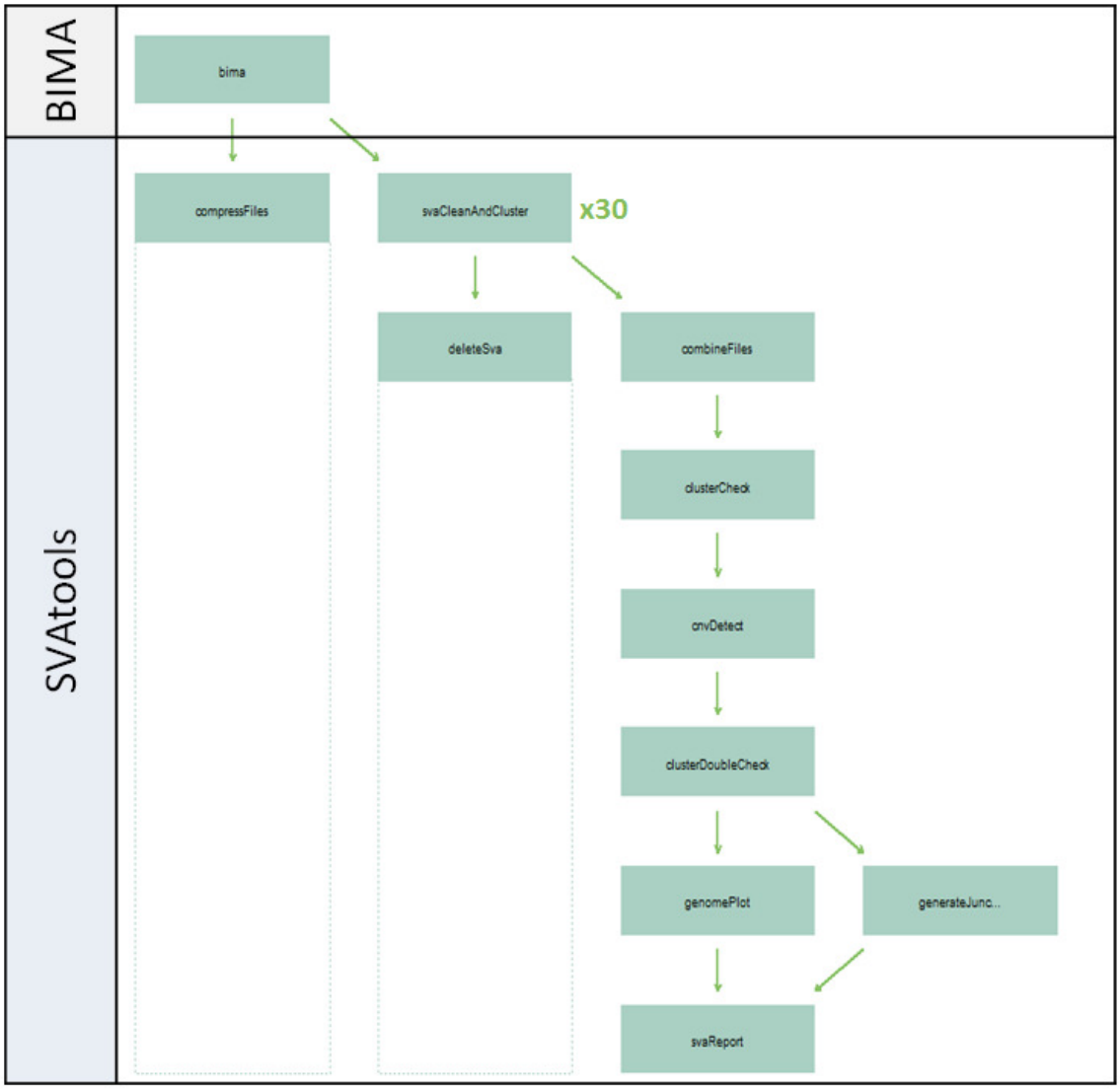


Figure 2C.3: BMD-SV algorithmic pipeline.
Each step can be executed independently, as a group, or sequentially from start to finish. BIMA, written in C, takes fastq files as input and maps to the requested reference genome, i.e. GRCh38. SVAtoolsJD, written and packaged in R, contains the remaining steps of the pipeline. The inputs are the .sva files produced by BIMA and the final outputs are .csvs, .pdfs, and .json files for listing and visualizing the reported SVs.

2C-3: SV detection – SVAtools to process BIMA output

SVAtools is a library written and packaged in the R software environment. SVAtools is capable of detecting SVs from NGS including: MPseq, traditional paired-end, custom capture (Jang et al. 2016) and amplicon sequencing. The SVAtools library operates in a series of steps (Figure 2C.3) that can be called individually, as a group, or in automatic succession. This flexibility allows “one-

button” operation of the entire pipeline, and quick feedback when adjusting configuration parameters to fine tune SV detection sensitivity or specificity.

The first algorithmic step, svaCleanAndCluster, operates in parallel on each of the 30 .sva files created by BIMA for GRCh38. Replicate fragments (also commonly called duplicates or non-unique read-pairs) are identified and excluded from any further analysis. The remaining fragments are sorted and indexed for faster downstream operations. For memory conservation the original, but now redundant, .sva files can be deleted. SVAtools provides methods for recreating the .sva files if needed.

To detect SVs, SVAtools combines three algorithmic approaches: read-pair, split-read, read depth/count (Tattini, D'Aurizio, and Magi 2015). SVAtools$_{JD}$ detects junctions by clustering the discordant read-pair and split-read fragments. Two or more discordant fragments supporting the same SV are identified as a “cluster” by the pipeline. Discordant fragments are those whose reads (readA and readB) map further apart than expected. SVAtools does not have an orientation requirement for distinguishing discordant from concordant fragments. Discordant fragments from both primary and split read mapping are clustered via a rapid custom clustering algorithm. The readA and readB positions are reduced from a 2-diminsional (1 position for each read in the fragment) to a 1-dimensional (1 position per fragment) array. Groups of fragments falling within a specified radius of each other along this 1-dimensional array are identified and clustered together. Details of each cluster, primarily fragment metrics such as total counts and size distributions, are concatenated and analyzed in the second step of SVAtools, combineFiles. Bridged coverage and base coverage calculations are also calculated, visualized, and stored.

The next step, clusterCheck has two main purposes. The first is to determine the breakpoints for each junction. The second is to remove false positive clusters by masking and filtering. The breakpoint resolution of MPseq is dependent on the bridged-coverage, and therefore dependent on fragment length and number of fragments. Typically the breakpoint resolution is guaranteed within half the fragment length. A histogram of the fragment lengths from a typical MP library is shown in Figure 2C.2. In this example, the peak (mode) fragment length is 2352bp. Therefore, the breakpoints reported for this sample are accurate within 1176 bps (2352 / 2 bps). But often the breakpoint reported by SVAtools$_{JD}$ is within a couple hundred bps of the true breakpoint. Furthermore, SVAtools$_{JD}$ can find the exact breakpoints if split reads are present. Figure 2A.2 demonstrates the resolution of the breakpoint position based on bridged coverage.

Masking and filtering reduces the false positive clusters to generate a reliable and meaningful list of junctions. Masking eliminates clusters detected in 'normal' samples free of cancer or constitutional diseases. Clusters may be detected in normal samples for several reasons. Normal sample clusters reveal: 1) likely benign variation present in the normal population but not in the reference genome; 2) artifacts incurred during the library prep; and 3) algorithmic artifacts, such as mapping errors due to errors or gaps in the reference genome, or repeat/homologous regions in the reference genome. These clusters are present in most samples sequenced by mate-pair and are therefore not diagnostically useful. The BMD-SV pipeline mask used in this study included clusters from 49 normal samples processed by the BMD-SV pipeline.

Filters remove poorly qualified clusters based on cluster quality metrics and provides a mechanism to adjust the sensitivity and specificity for junction detection. Strict thresholds will exclude many false positives, but perhaps remove true positives. Lenient thresholds will likewise allow many false positives. The primary filters include: minimum number of fragments in a cluster, minimum and maximum cluster span (distance between the outermost reads in a cluster), number of mismatches within reads in a cluster, homology score (sequence similarity) of reads in a cluster, size of intra-chromosomal rearrangements, and ratio of reads mapping to the positive vs negative strand. Default settings to maximize sensitivity and specificity have been determined via sequencing and mapping over 2000 mate-pair samples from over 30 different diseases. Typical filter settings were applied to the 29 samples in this analysis. To eliminate chimeras formed during the MPseq library preparation, all clusters must have at least 3 fragments. Clusters spanning more than 20,000 bps or less than 250 bps were also discarded. All clusters must have at least one read with less than 3 mismatches, and an average homology score of 1.65 or less. The strict homology score allows for some fragments within the cluster to map to multiple positions, but ensures the cluster as a whole was uniquely mapped. Clusters with high homology can be rescued based on other criteria in later steps.

The impact of masking versus filtering clusters for each sample is shown in Figure 2C.4. On average 32,800 clusters are detected for a sample, but only 14 are reported as junctions. $SVAtools_{JD}$ also provides methods to search for junctions manually. These methods help rule out false negative junctions, such as junctions that fail a filter or junctions with only 1 or 2 supporting fragments. These methods operate like FISH testing; the investigator must know which two genomic regions to interrogate and enables searching with much higher specificity than the standard pipeline.

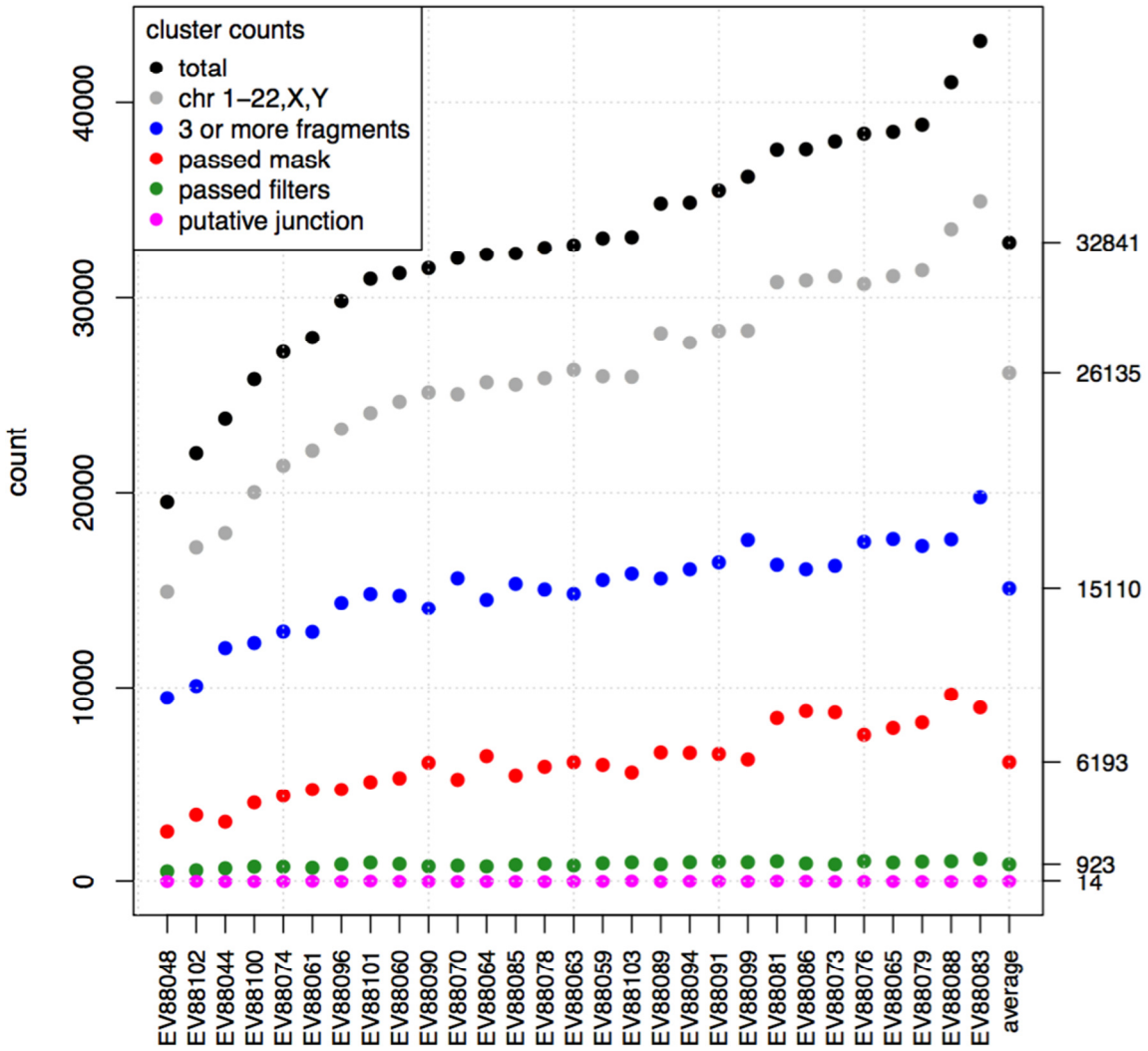


Figure 2C.4 Impact of masking vs filtering clusters on each of the 29 samples.
On average, each sample will produce nearly 33,300 clusters of 2 or more discordant fragments. Fragments with both reads mapping to just chromosomes 1-22, X, and Y account for 80% of the clusters, the other 20% have at least one read mapping to an unlocalized, unplaced, or alternate genomic sequence. Only 46% of the clusters have at least 3 fragments or more. Clusters with just 2 fragments are primarily due to false ligations (chimeras), formed during the MPseq library prep. Masking eliminates 81% of all clusters, and filtering removes 98% of all clusters. In practice, filtering primarily removes clusters with low fragment counts, while masking removes clusters with high fragment counts but are thought to represent artifacts from the pipeline or common population variants. On average, out of 32,841 clusters, 14 are reported by SVAtools$_{JD}$ as putative junctions.

The CNVdetect step assesses CNV using a read depth approach and the detected junctions to increase CNV sensitivity. The read depth approach uses the read count of concordant fragments,

~95% of all fragments in MPseq libraries, and assumes that after normalization they uniformly cover the reference genome sequence without bias. Regions that are found to have artificially high or low read depth are excluded. These regions are identified from the same 49 normal samples mentioned previously and are largely due to mapping artifacts. Normalization is performed by the normal sample with the highest copy number correlation to the patient sample. After normalization, a sliding window step detection algorithm is repeated for a range of window sizes from 100 kb – 1Mb to detect edges between regions with significant copy number difference. Breakpoints that are part of junctions found using the junction detection algorithms in SVAtools are added to the edge detection in order to supplement the statistically determined edges and improve sensitivity and resolution. The detected edges are used to segment the genomic data. Each segmented region is tested to determine which deviate significantly from the expected read depth and those with higher or lower read depth than expected are considered gains or losses, respectively. The algorithms can account for heterogeneous samples with numerous cell clones and are sensitive enough to easily detect CNVs when the cell population includes as little as 20% tumor. When less than 5% tumor, additional techniques are applied. Each CNV call is reported with a Normalized Read Depth score (NRD) to quantify the copy number level. Full details and analysis regarding CNV detection are provided in the forthcoming paper Smadbeck et al.

The next step in $SVAtools_{JD}$, clusterDoubleCheck, has two purposes: 1) review filtered clusters to unfilter those passing a second set of criteria; and 2) annotate and output the junctions to downloadable tables and graphics. Filtered clusters that are reviewed include clusters at non-terminal CNV edges. Such clusters passing a relaxed set of filters can be confidently rescued because we expect a junction at all non-terminal edges of CNVs.

2C-4: SV visualization – SVAtools to draw junction, region and genome plots

The remaining steps create output for analysis and distribution among team members, provided in several forms: .csv data tables and .pdfs to visualize the SVs. Tables list the breakpoints for each detected junction and the CNV start and stop locations. Genome plots provide a whole-genome snapshot of all detected junctions and CNVs. Figure 2C.5 provides two examples, a relatively stable and an unstable genome. Junction (Figure 2C.6A) and region plots (Figure 2C.2B) show gene level detail of each reported junction. The final step, svaReport, assembles the pipeline information and results into a single pdf for quick and portable review. The pipeline information includes essential sample metadata and quality control statistics.

A


1
2
3
4
5
6
7
8
9
10
11
12
X
22
21
20
19
18
17
16
15
14
13
Y
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
(Mb)

B

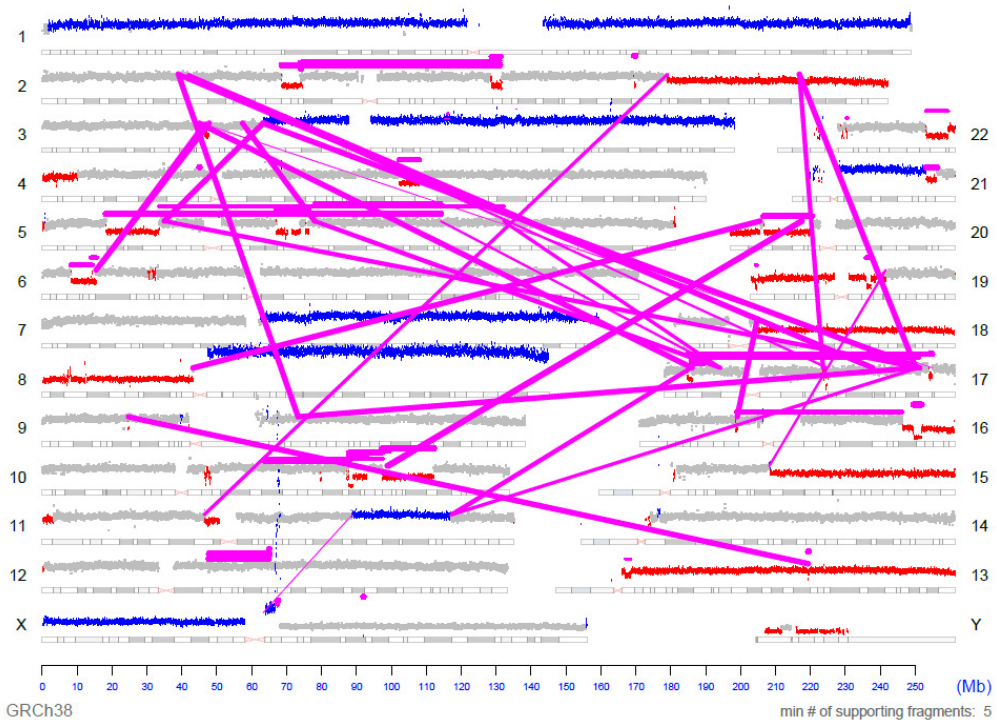
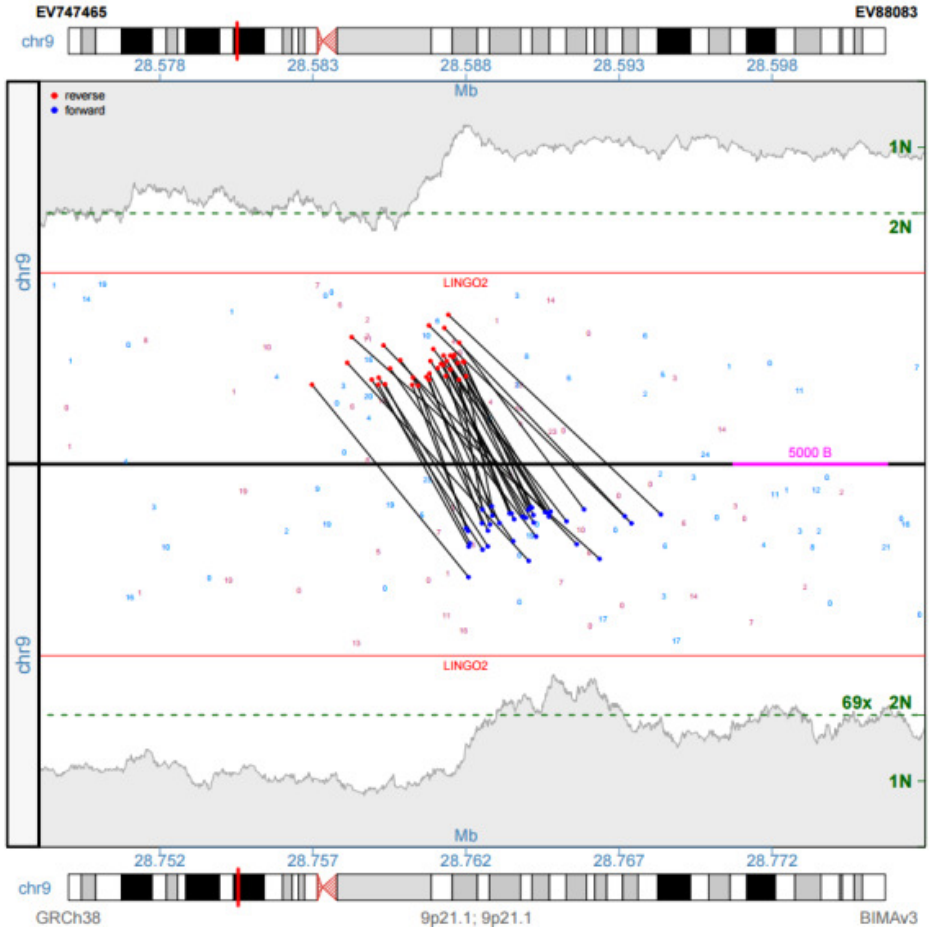

1
2
3
4
5
6
7
8
9
10
11
12
X
22
21
20
19
18
17
16
15
14
13
Y
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
(Mb)
GRCh38
min # of supporting fragments: 5

Figure 2C.5: Genome Plots

In a genome plot, all chromosomes are displayed and arranged in a U-shaped configuration for efficient use of space. Each magenta line represents a junction reported by SVAtools$_{JD}$. The endpoints of the line indicate the position of the two reported breakpoints, breakpoint A and B. The thickness of the line is proportional to the number of supporting fragments. Copy number is displayed by dot color and height above the idiogram for each chromosome. Neutral copy number is shown in gray, red for loss, and blue for gain. Areas with no dots, (centromeres and heterochromatin regions) remain unavailable in the GRCh38 reference genome and therefore fragments do not map to these areas. A) Genome plot with three detected junctions: a balanced inversion, a deletion CNV, and an unbalanced translocation. B) Genome plot of an unstable genome with numerous SVs involving over 30 junctions and many CNVs including whole arm and whole chromosome aneuploidy.

A


EV747465
EV88083
chr9
28.578
28.583
28.588
28.593
28.598
Mb
reverse
forward
1N
2N
LINGO2
5000 B
69x
28.752
28.757
28.762
28.767
28.772
GRCh38
9p21.1; 9p21.1
BIMAv3

B

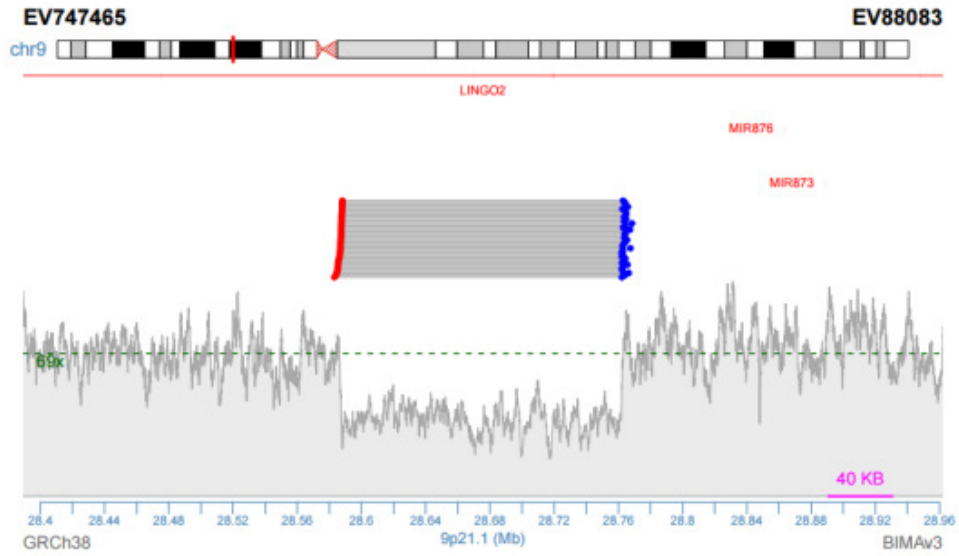

EV747465
EV88083
chr9
LINGO2
MIR876
MIR873
69x
40 KB
GRCh38
9p21.1 (Mb)
BIMAv3

Figure 2C.6: Junction and region plot illustrating a 200kb deletion on chromosome 9. A) Junction plots are a vertically stacked two panel graph, one panel for each breakpoint, breakpoint A illustrated above breakpoint B. A genomic region, 30kb in this example, centered on the reported breakpoint is shown along the x-axis. Corresponding idiograms indicate the chromosomal position of the region by a red bar. The lines show the fragments that span the junction and support the rearrangement. The position of each read of a supporting fragment is located by a dot and color coded by strand, red for reads mapping to the reverse strand, blue for forward strand. The bridged coverage for the region is illustrated by the shaded area. The green dotted line on the y-axis indicates the bridged coverage averaged across the entire genome (normalized to estimate 2N and 1N). Inspection of the bridged coverage will reveal possible copy number changes near the breakpoint. Genes within the region are displayed, indicating exon location and strand direction. Thus, the junction plot illustrates: 1) the genomic position and strand of each read supporting the junction: 2) local and genome-wide bridged coverage; and 3) nearby, possibly affected, genes. B) Region plots are drawn only for intra-chromosomal rearrangements, are similar to junction plots, except breakpoint A, breakpoint B, and the genomic segment in-between, are all shown in one panel.

Chapter 3: Sensitivity and junction detection performance of SVAtools$_{JD}$ compared to traditional cytogenetic techniques

3A: SV Detection Method Comparisons

SVAtools' junction detection sensitivity was evaluated from a test set of breakpoints, collected from the cytogenetic results of the 29 patients. The test set included 285 breakpoints from 87 rearrangements clinically reportable by at least one of the cytogenetic methods. Table 3A.1 is a heat map summarizing the detection of each rearrangement both quantitatively and qualitatively for each cytogenetic method and SVAtools$_{JD}$: quantitatively by the number of breakpoints detected in each rearrangement, and qualitatively based on ability to reconstruct the rearrangement. For example, CMA can detect the breakpoints involved in a copy number loss and gain, but cannot demonstrate that these imbalances are secondary to an unbalanced translocation. Table 3A.2 is a heat map summarizing the number of breakpoints detected per sample for each cytogenetic method and SVAtools$_{JD}$. The sensitivity of SVAtools$_{JD}$ compared to each cytogenetic method is shown in Table 3A.3

Table3A.1: Assessment of method performance for individual rearrangements. *

	A
Id	**Rearrangement**
17	deletion
49	deletion
64	deletion
13	homozygous deletion
16	deletion
19	deletion
20	deletion
21	tandem duplication
24	deletion
30	deletion
32	deletion
33	deletion
50	deletion
55	deletion
62	deletion
75	deletion
73	deletion
76	deletion
84	deletion
85	deletion
87	deletion
38	translocation with deletion at breakpoints
47	translocation with deletion at one breakpoint
6	nested deletion
41	nested deletion
52	unbalanced translocation
65	unbalanced translocation
9	complex: 4-way translocation with imbalances at breakpoints plus an inversion; 2chrs, 6 jcts
4	tandem duplication
14	complex: nested inversion, duplications; 1 chr; 3 jcts
36	tandem duplication
44	tandem duplication
54	complex: translocation with deletion at breakpoints; 2 chrs, 5 jcts
59	pericentric inversion with adjecent deletion
61	tandem duplication
70	complex: nested deletion, gain, deletion, 1 chr, 7 jcts
71	complex: nested deletion, 1 chr, 4 jcts
74	complex: deletion-inversion; 1 chr, 4 jcts
77	translocation with gain at breakpoints
82	tandem duplication
83	nested deletion
7	nested deletion
69	complex: translocation with deletion at one breakpoint, 2 chrs, 3 jcts
22	complex: translocation with deletions at both breakpoints; 2 chrs, 3 jcts
40	complex: translocation, inversion and copy number imbalance at both breakpoints; 2 chrs, 3 jcts
66	complex: translocation with molecular complexities at breakpoints, 2 chrs 5 jcts
3	balanced translocation
12	balanced translocation
15	3-way translocation
18	balanced translocation
27	balanced translocation
28	balanced translocation
43	balanced translocation
53	pericentric inversion
42	deletion
63	inversion
81	complex: translocation with deletion at some breakpoints; 3 chrs, 5 jcts
25	complex: translocation with imbalances at breakpoints; 2chrs, 5 jcts, part2
29	complex: 4-way translocation; 3 chrs, 4 jcts
48	balanced translocation
39	4-way translocation
60	balanced translocation
8	pericentric inversion
51	balanced translocation
67	balanced translocation
68	balanced translocation
80	translocation with adjacent deletion
34	nested deletion (terminal)
78	deletion with likely translocation
56	insertion
31	deletion
10	nested deletion
26	likely insertion
37	nested deletion (terminal)
79	complex: nested duplication; 1 chr, 3 jcts part 2
5	complex: translocation, deletion and inversion, 2 chrs, 3 jcts
11	isochromosome 7q
35	tandem duplication, possibly common SV
57	likely tandem duplication
58	complex: inversion followed by deletion
46	balanced translocation
1	isochromosome 17q
2	unbalanced translocation
23	deletion
86	deletion
72	nested deletion
45	homozygous deletion
	TOTALS

		B Breakpoints detected				C Rearrangement detected			
Id	break points	Karyotype	FISH	CMA	MPseq SVAtools$_{JD}$	karyotype	FISH	CMA	MPseq SVAtools$_{JD}$
17	2	0	2	2	2				
49	2	2	ND	2	2				
64	2	2	2	2	2				
13	2	0	ND	2	2				
16	2	0	ND	2	2				
19	2	0	ND	2	2				
20	2	0	ND	2	2				
21	2	0	ND	2	2				
24	2	0	ND	2	2				
30	2	0	ND	2	2				
32	2	0	ND	2	2				
33	2	0	ND	2	2				
50	2	0	ND	2	2				
55	2	0	ND	2	2				
62	2	0	ND	2	2				
75	2	0	ND	2	2				
73	2	0	ND	2	2				
76	2	0	ND	2	2				
84	2	0	ND	2	2				
85	2	0	ND	2	2				
87	2	0	ND	2	2				
38	4	4	4	4	4				
47	4	4	4	4	4				
6	4	0	2	4	4				
41	4	0	2	4	4				
52	2	2	ND	2	2				
65	2	1	1	2	2				
9	6	6	ND	6	6				
4	2	0	ND	2	2				
14	4	0	ND	4	4				
36	2	0	ND	2	2				
44	2	0	ND	2	2				
54	2	0	ND	2	2				
59	2	0	ND	2	2				
61	2	0	ND	2	2				
70	4	0	ND	4	4				
71	2	0	ND	2	2				
74	4	0	ND	4	4				
77	2	0	ND	2	2				
82	2	0	ND	2	2				
83	2	0	ND	2	2				
7	4	0	4	3	4				
69	6	0	6	2	6				
22	6	4	2	4	6				
40	8	4	2	4	8				
66	9	4	4	5	9				
3	4	4	4	0	4				
12	4	4	4	0	4				
15	6	6	6	0	6				
18	4	4	4	0	4				
27	4	4	4	0	4				
28	4	4	4	0	4				
43	4	4	4	0	4				
53	4	4	4	0	4				
42	2	0	2	0	2				
63	4	4	4	0	4				
81	8	8	8	0	8				
25	4	0	4	0	4				
29	6	0	6	0	6				
48	4	0	4	0	4				
39	8	8	6	0	8				
60	4	4	2	0	4				
8	4	4	ND	0	4				
51	4	4	ND	0	4				
67	4	4	ND	0	4				
68	2	2	ND	0	2				
80	4	4	ND	0	4				
34	3	0	ND	3	3				
78	1	0	ND	1	1				
56	2	2	ND	0	2				
31	2	0	ND	2	0				
10	6	0	2	6	4				
26	2	0	ND	2	1				
37	3	0	ND	3	2				
79	4	2	ND	4	3				
5	6	4	ND	4	4				
11	1	1	ND	1	0				
35	2	0	ND	2	0				
57	1	0	ND	1	0				
58	4	4	ND	2	2				
46	4	4	4	0	2				
1	1	1	ND	1	0				
2	6	6	6	3	0				
23	2	0	ND	2	0				
86	2	0	ND	2	0				
72	3	0	2	3	0				
45	4	4	ND	2	0				
	285	138	119	163	249				

***Assessment of method performance for each rearrangement. A) Each row represents a unique rearrangement evaluated across multiple methodologies. The 29 samples contained 87 unique rearrangements total. Two heat maps are shown to reveal the quantitative and qualitative detection performance of each method: B) Quantitative assessment: Breakpoints Detected. Each cell shows the number of breakpoints detected by each of the four test methods and is color coded to indicate the level of breakpoint detection: green, orange and red, for all, partial and no breakpoints detected respectively. Gray cells indicate breakpoints not tested by FISH; 48% (137/285) of breakpoints in this study. C) Qualitative assessment: Rearrangement Detected. Each cell indicates if the method fully described and reconstructed the rearrangement. For example CMA detects all the breakpoints in an unbalanced translocation but does not reveal the structure of the rearrangement.**

Table 3A.2: Assessment of method performance per sample *

Sample	Break points	karyo type	FISH	CMA	MPseq SVAtools$_{JD}$
EV88083	4	100%	100%	100%	100%
EV88090	4	100%	100%	100%	100%
EV88070	2	0%	100%	100%	100%
EV88064	4	0%	0%	100%	100%
EV88048	2	0%	0%	100%	100%
EV88100	8	88%	88%	50%	100%
EV88088	6	67%	67%	33%	100%
EV88101	27	37%	37%	48%	100%
EV88073	10	40%	40%	60%	100%
EV88060	8	50%	50%	38%	100%
EV88091	8	25%	50%	50%	100%
EV88086	14	29%	43%	57%	100%
EV88099	8	50%	25%	50%	100%
EV88065	8	75%	75%	25%	100%
EV88078	4	100%	100%	0%	100%
EV88079	4	100%	100%	0%	100%
EV88085	8	100%	75%	0%	100%
EV88061	12	50%	17%	100%	83%
EV88102	16	0%	13%	100%	81%
EV88074	10	40%	20%	80%	80%
EV88059	10	40%	20%	80%	80%
EV88081	24	0%	25%	75%	79%
EV88096	6	67%	0%	67%	67%
EV88076	6	0%	67%	33%	83%
EV88063	8	75%	50%	50%	75%
EV88094	17	71%	24%	41%	94%
EV88103	28	50%	29%	57%	89%
EV88044	11	100%	91%	36%	36%
EV88089	8	100%	50%	25%	25%
	285	48%	42%	57%	87%

*** Heat map of breakpoints detected per sample for each method. Each cell is color coded by percentage of breakpoints detected. Green = 100%, light green = 67-99%, yellow=34-66%, orange = 1-33%, red = 0%**

Table 3A.3: Sensitivity of MPseq with SVAtools$_{JD}$, compared to each cytogenetic method. *

			karyotype	FISH	CMA
		breakpoints	138	119	163
MPseq with SVAtools$_{JD}$	breakpoints	reported	120	109	137
		not reported	18	10	26
	Sensitivity	all	87%	92%	84%
		in >25% of cells	97%	100%	90%

*** based on detection of all breakpoints in each method and detection of breakpoints observed in greater than 25% of cells**

3A-1: Karyotype to MPseq with SVAtools$_{JD}$ comparison

Karyotype analysis revealed 36 rearrangements, with a total of 138 predicted breakpoints (Table 3A.3). Of the 38 rearrangements, only 7 were simple rearrangements with one junction each, such as deletions and unbalanced translocations. Of the 138 breakpoints, SVAtools$_{JD}$ reported 120 breakpoints for 87% sensitivity. Of the 18 breakpoints not detected by SVAtools$_{JD}$, 4 breakpoints mapped to the centromere, and the remaining 14 were observed in tumors with a low clonal proportion: 6 in 15%, 8 in 25%. When only breakpoints with greater than 25% of clonal proportion are considered, the SVAtools$_{JD}$ sensitivity compared to karyotype is 97%. Of note, the 6 breakpoints seen in 15% of cells were actually detected by SVAtools$_{JD}$ but were part of a complex t(15;17) translocation observed on multiple chromosome copies: 46,XX,t(15;17)(q22;q21)[17]/46,XX,der(15)t(15;17)(q22;q21),ider(17)(q10)t(15;17)(q22;q21)[3] Karyotype identified the presence of the low level translocation in addition to the balanced translocation in 17/20 cells. SVAtools$_{JD}$ clustered all supporting fragments together to report only one balanced translocation. Despite not fully characterizing all rearrangements in the clone, the disease-defining rearrangement for the sample was detected by SVAtools$_{JD}$.

3A-2: FISH to MPseq with SVAtools$_{JD}$ comparison

FISH analysis revealed 32 rearrangements, with a total of 119 predicted breakpoints (Table 3A.3). SVAtools$_{JD}$ reported 106/119 breakpoints for 91% sensitivity. Of the 10 breakpoints not detected by SVAtools$_{JD}$, 6 are the same breakpoints mention previously in the karyotype comparison (part of a complex translocation observed on multiple chromosome copies seen in a clonal proportion of 15%). The remaining 4 false negative breakpoints included 2 each from tumors with 6% and 25% clonal proportion. When only breakpoints found in 25% of cells or more are considered, the SVAtools$_{JD}$ sensitivity compared to FISH is 100%.

3A-3: CMA to MPseq with SVAtools$_{JD}$ comparison

CMA analysis revealed 64 rearrangements, with a total of 163 predicted breakpoints (Table 3A.3). SVAtools$_{JD}$ reported 137/163 breakpoints by junction detection for 84% sensitivity, increasing to 90% sensitivity (137/153) when only breakpoints found in 25% of cells or more are considered. Of the remaining 16 breakpoints not detected, 14 occurred at edges of a CNV detected by SVAtools' CNV detection methods, including 4 centromeric breakpoints. Thus similar to CMA, for these 14 breakpoints SVAtools could detect the CNV but not the junction. Part of a nested deletion detected by CMA, 9p21.3p21.3(21772930_22215463)x1[0.4], but not SVAtools$_{JD}$ accounts for the remaining false negative breakpoint. SVAtools$_{JD}$ did detect a junction with 2 supporting fragments but requires at least 3 fragments for reporting. FISH detected the adjacent deletion (CDKN2Ax1,D9Z1x2)[40/200] reported by CMA at 9p22.1p21.3(19625174_21767405)x1[0.2], but did not test, thus confirm, this deletion reported by CMA.

3B: Multi-method comparison: Karyotype, FISH, CMA, and MPseq with SVAtools$_{JD}$

Combining the full set of breakpoints for each cytogenetic method enables a comparison of each method's ability to detect breakpoints. Because each of the three cytogenetic methods has advantages and limitations, this comparison better illustrates the detection performance of SVAtools$_{JD}$ (Table 3B). For the three cytogenetic methods, a combined total of 285 breakpoints were detected from 87 rearrangements. SVAtools$_{JD}$ reported 87% (249/285) of breakpoints.

Table 3B: Multi-method comparison: karyotype, FISH, CMA and MPseq with SVAtools$_{JD}$. *

	karyotype	FISH	CMA	MPseq with SVAtools$_{JD}$
breakpoints				
reported	138	119	163	249
not reported	151	29 / 166	122	36
Total	285	148 / 285	285	285
Detection performance	49%	80% / 42%	57%	87%

***SVAtools$_{JD}$ achieved 87% detection performance, the highest of all methods, for the combined total of 285 breakpoints.**

Although FISH is known to be a very sensitive test, FISH was found to have the lowest detection performance 41% (119/285) (Table 3B) in the multi-method comparison. The low detection

performance reflects that FISH did not test 48% (137/285) of the breakpoints due to panel design or ordering practices. If the 137 breakpoints not tested are excluded from the total, FISH detection performance is 80% (119/148). FISH results did not always reveal the full nature of the rearrangement or report a CNV when the variant was outside the probe region. For example in sample EV88085, the FISH result (ABL1x3)(BCRx3),(ABL1 con BCRx1)[405/500], identifies more than 2 breaks but the investigator does not know where and does not know how many breaks. In comparison, both karyotype and SVAtools$_{JD}$ reported a four-way translocation, t(X;10;9;22)(q11;p13;q34;q11.2). FISH did detect the BCR-ABL1 fusion it was designed for, but not the ABL1 and MSN truncation, or the MSN->BCR fusion predicted by SVAtools$_{JD}$ by the other three junctions. Illustrating the other common limitation of FISH, in four samples from this study (including one example discussed below), FISH detected a heterozygous loss, but not the adjacent homozygous loss.

Karyotype detected 48% (138/285) of all breakpoints in the multi-method comparison (Table 3B). Often karyotype was returned normal due to poor metaphase quality. Over half of all breakpoints not reported by karyotype were because the rearrangement was too small to detect. The remaining breakpoints not reported were due to misinterpretation or under-appreciation of complex rearrangements. For example, the karyotype for sample EV88059 karyotype reported a balanced rearrangement, t(5;6)(q13;q23), but by SVAtools$_{JD}$ this is more complex, 5q13 and 6q23 are part of a 3-way translocation between chromosomes 5 and 6, with inversion and 2 losses (30Mb and 5 Mb) on chromosome 6. The losses were confirmed by CMA. There is a junction detected, but not reported by SVAtools$_{JD}$, connecting the left edges of each of the two chromosome 6 losses. The junction was not reported because it had only 2 supporting fragments; at least 3 fragments are required for reporting. SVAtools CNV detection did however report the edges of the CNV, and similar to FISH, since the region of interest is known, interrogation of the data was possible which revealed the low-level junction. Knowledge of this junction allows for complete reconstruction of the rearrangement. Sample EV88101 also illustrates this karyotype limitation. The karyotype reported a deletion, del(11)(q21q23), and an add(19)(q13.1). FISH ((KMT2Ax2)[200]) and CMA were normal, while SVAtools detected a balanced translocation, t(11;19)(q21;q13.11). This deletion was counted as a false positive for karyotype and a true negative for FISH, CMA, and SVAtools$_{JD}$.

CMA detected 57% (163/285) of all breakpoints in the multi-method comparison (Table 3B). Based on the rearrangement characterization by SVAtools$_{JD}$, 63% (102/163) of the CMA breakpoints were due to a rearrangement other than a simple deletion or tandem duplication.

Thus, while CMA could report the CNV, the nature of the rearrangement was not reported. Of the 122 breakpoints not reported by CMA, 92% (112/122) were due to undetected balanced rearrangements such as inversions and balanced translocation. Of the remaining 10 breakpoints not detected by CMA, 6 breakpoints were from a deletion or unbalanced translocation reported by FISH. The remaining 4 breakpoints were complex rearrangements detected by karyotype.

Chapter 4: Additional advantages of MPseq and SVAtools$_{JD}$ over traditional cytogenetic techniques

4A: Resolving Ambiguous, Cryptic and Complex Rearrangements with SVAtools$_{JD}$

In addition to the higher detection performance of SVAtools$_{JD}$ compared to standard cytogenetic methods, SVAtools$_{JD}$ can more completely characterize the rearranged genome by: 1) resolving the genomic configuration of complex rearrangements, and 2) reporting the breakpoints at a much higher resolution. To illustrate these points, five samples are described below. For each sample, the relevant cytogenetic test results for each method, the SVAtools$_{JD}$ results and the inferred molecular karyotype are listed. All positions are listed in GRCh38 coordinates. The RF, FR, FF, and RR columns refer to the number of fragments mapping to the corresponding strand orientation where R is reverse and F is forward, for position A and position B respectively.

In sample EV88100, SVAtools$_{JD}$ reports three non-complex junctions that required both karyotype and CMA to detect (Figure 4A.1), the inv(3), del(5), and the unbalanced translocation, der(17)t(11;17). SVAtools$_{JD}$ clarified the karyotype findings, demonstrating the add(17) as an unbalanced derivative and reported a possible VAMP2->BCO2 fusion. While FISH test demonstrated the inv(3), a standard FISH panel for MDS does not test for t(11;17) rearrangements. CMA demonstrated the del(5) and the deletion/duplication associated with the 7;11 translocation, however the derivative chromosome 17 could only be inferred, and the balanced inversion 3 was missed. Individually karyotype, FISH, and CMA provide a piece of the puzzle, but SVAtools$_{JD}$ provides a comprehensive understanding of the SVs in the neoplasm. Junction plots for these three rearrangements are in Figure 4A.2

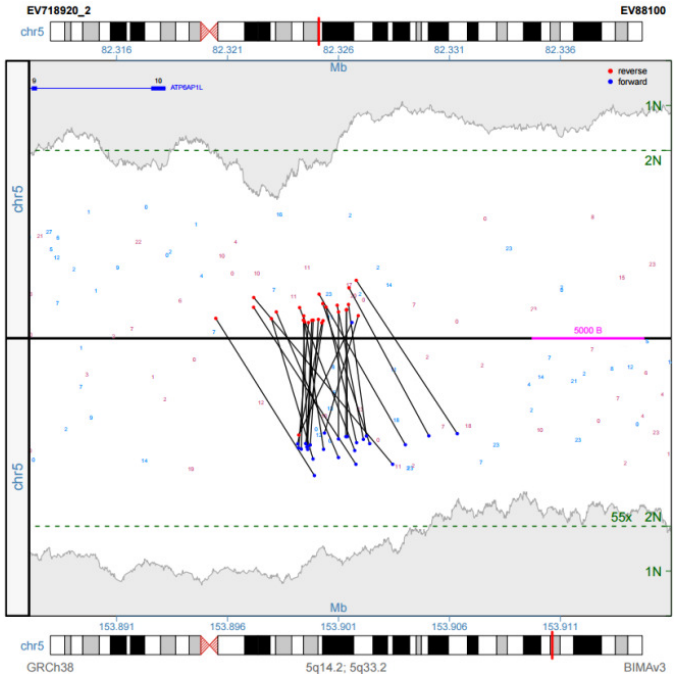
A

Sample	EV88100 RFR: acute leukemia, MDS								
Karyotype	46,XX,inv(3)(q21q26.2),del(5)(q13q33),add(17)(p13)[20]								
FISH	(RPN1x3),(EVI1x3),(RPN1 con EVI1x2)[454/500]/ (D5S630x2,EGR1x1)[198/200]/(MLLx3)[185/200]								
CMA	**Locus Start**	**Locus End**	**Position Min**	**Position Max**	**Size**	**Type**			
	11q23.1	11q25	112,186,582	135,068,576	22,881,994	Gain			
	17p13.3	17p13.1	150,732	8,158,478	8,007,746	Loss			
	5q14.2	5q33.2	82,326,551	153,894,589	71,568,038	Loss			
MPseq with SVAtools$_{JD}$	**Locus A**	**Locus B**	**Position A**	**Position B**	**Size**	**RF**	**FR**	**FF**	**RR**
	11q23.1	17p13.1	112,187,010	8,158,594	NA	0	0	12	1
	3q21.3	3q26.2	128,546,376	168,886,142	40,339,766	0	17	10	2
	5q14.2	5q33.2	82,326,442	153,901,421	71,574,979	26	1	0	0
Inferred Mol. Karyotype	seq inv(3)(q21q26.2),del(5)(q14.2q33.2),der(17)t(11;17)(q23.1;p13.1)								

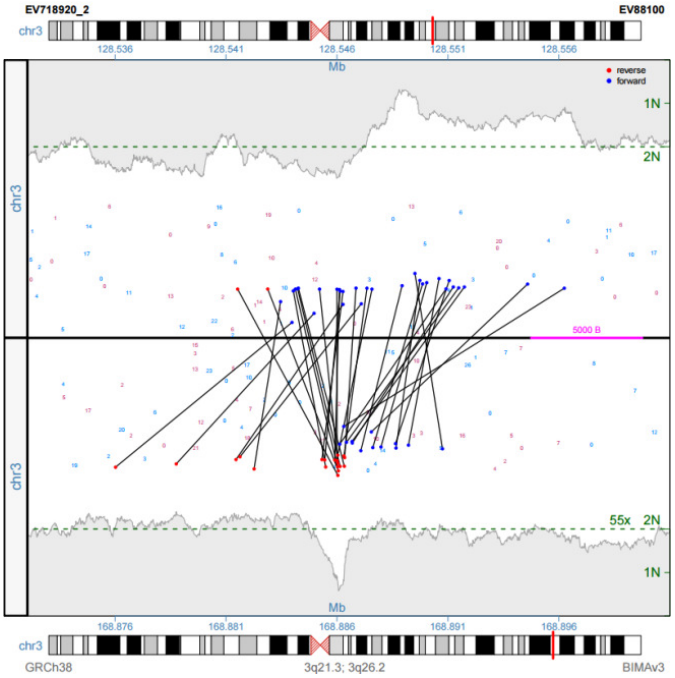
B

Figure4A.1: EV88100 a sample with three non-complex junctions.
A) Results of each SV detection method B) genome plot showing the copy number and junctions for each of the three rearrangements: del(5), inv(3) and t(11;17).

A


EV718920_2
EV88100
chr5
82.316
82.321
82.326
82.331
82.336
Mb
reverse
forward
1N
2N
5000 B
55x 2N
153.891
153.896
153.901
153.906
153.911
GRCh38
5q14.2; 5q33.2
BIMAv3

B


EV718920_2
EV88100
chr3
128.536
128.541
128.546
128.551
128.556
Mb
reverse
forward
1N
2N
5000 B
55x 2N
168.876
168.881
168.886
168.891
168.896
GRCh38
3q21.3; 3q26.2
BIMAv3

C

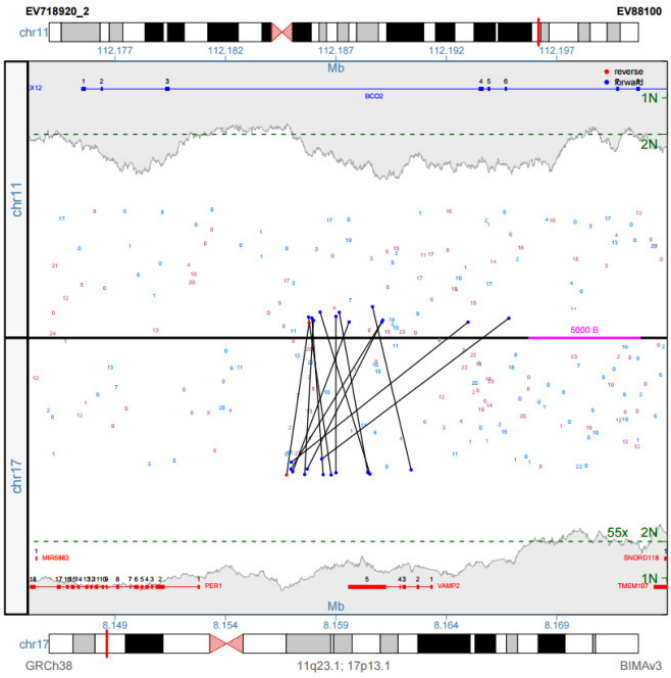


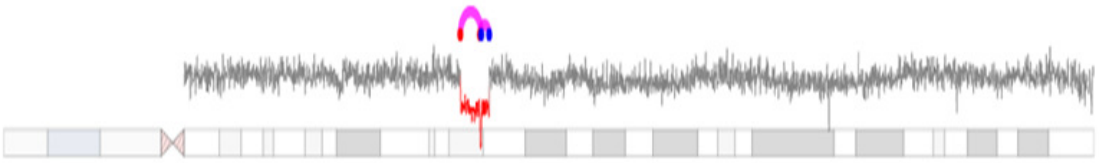
Figure 4A.2: EV88100 junction plots
A) del(5), B) inv(3) and C) t(11;17)

In sample EV88059, SVAtools$_{JD}$ reports two junctions on chromosome 13 (Figure 4A.3), resulting in a nested deletion: two overlapping deletions, 2,194 kb and 960 kb, with a 133kb homozygous loss including TRIM13, KCNRG, and the 3' end of DLEU2. This nested deletion illustrates a major limitation of FISH related to detection of deletion/duplication below the resolution of the probe. FISH reported one heterozygous loss, but missed the focal homozygous deletion. Karyotype was normal for chromosome 13. CMA revealed the homozygous loss and the two adjacent heterozygous deletions; however, the structural nature of the rearrangement could not be characterized. The junction plots (Figure 4A.4) used to reconstruct chromosome 13 (Figure 4A.5) are shown below.

A

Sample	EV88059 RFR: CLL								
Karyotype	46,XY,t(5;6)(q13;q23)[10]/47,XY,+12[2]/46,XY[8] (Normal 13)								
FISH	(D13S319x1),13q34(LAMP1x2)[112/200]								
CMA	**Locus Start**	**Locus End**	**Position Min**	**Position Max**	**Size**	**Type**			
	13q14.2	13q14.2	47,898,938	49,957,631	2,058,694	loss			
	13q14.2	13q14.2	49,971,104	50,097,568	126,464	loss			
	13q14.2	13q14.3	50,108,920	50,922,184	813,264	loss			
MPseq with SVAtools$_{JD}$	**Locus A**	**Locus B**	**Position A**	**Position B**	**Size**	**RF**	**FR**	**FF**	**RR**
	13q14.2	13q14.2	47,897,178	50,091,564	2,194,386	13	0	0	0
	13q14.2	13q14.3	49,958,440	50,918,770	960,330	13	0	0	0
Inferred Mol. Karyotype	seq del(13)(q14.2q14.2, del(13)(q14.2q14.3)								

B



C

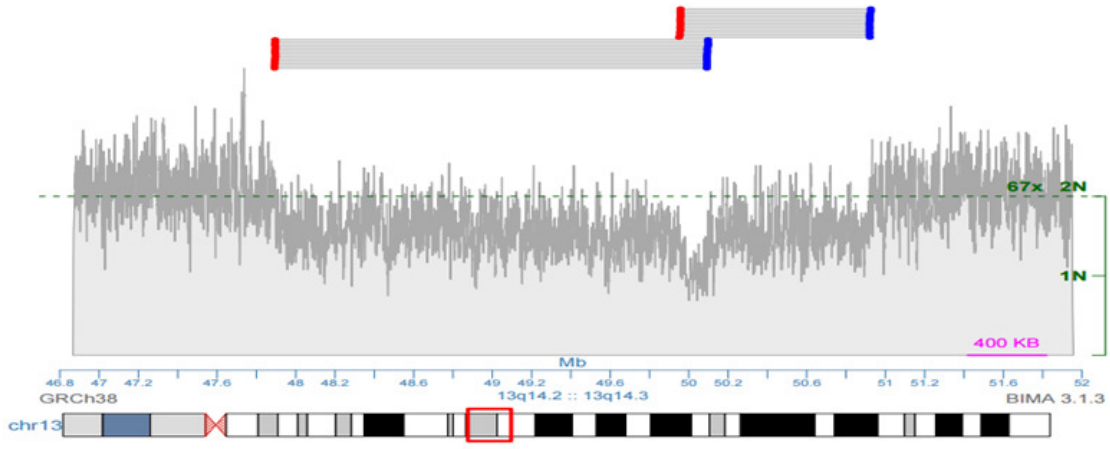
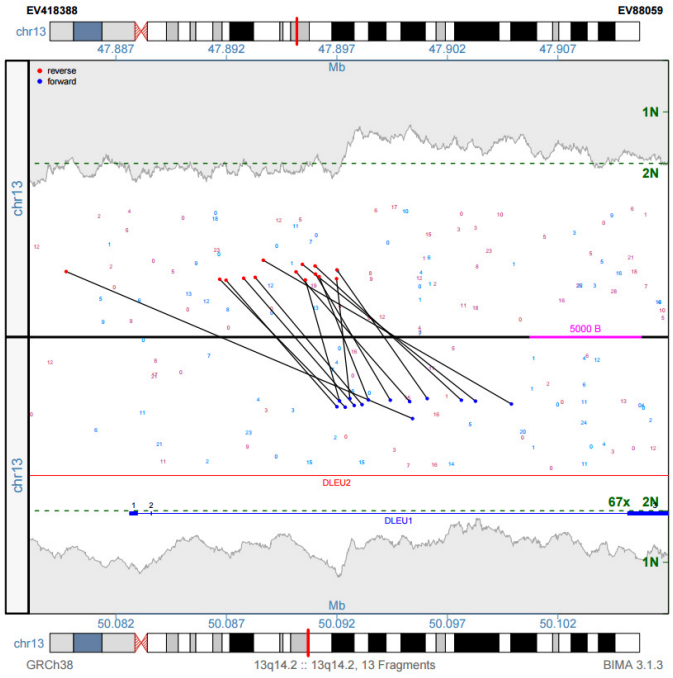


Figure 4A:3 EV88059, a sample with a nested deletion
A) Results of each SV detection method B) genome plot view of chromosome 13 showing the copy number and junctions C) region plot of chromosome 13 between 46.8-52 MB. The red dot-gray line-blue dot clusters represent the two deletion junctions reported.

A



B

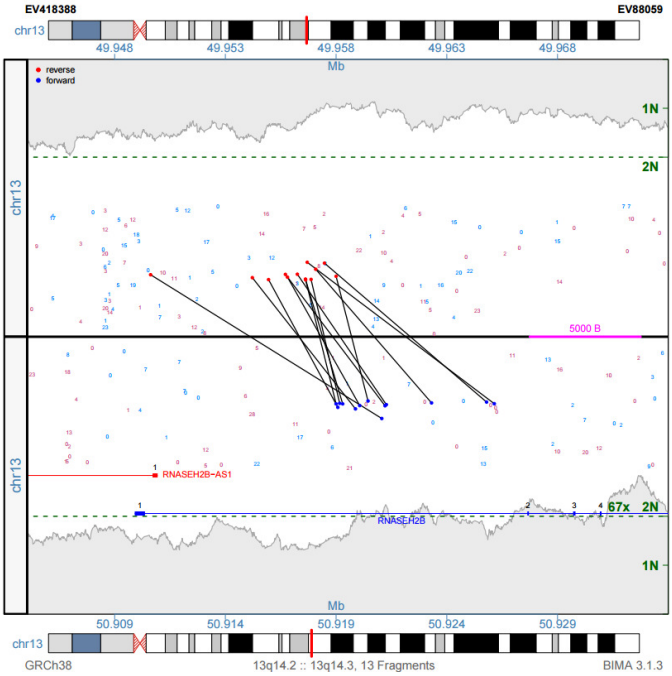


Figure 4A.4: EV88059 junction plots
A) del(13)(q14.2q14.2), B) del(13)(q14.2q14.2)

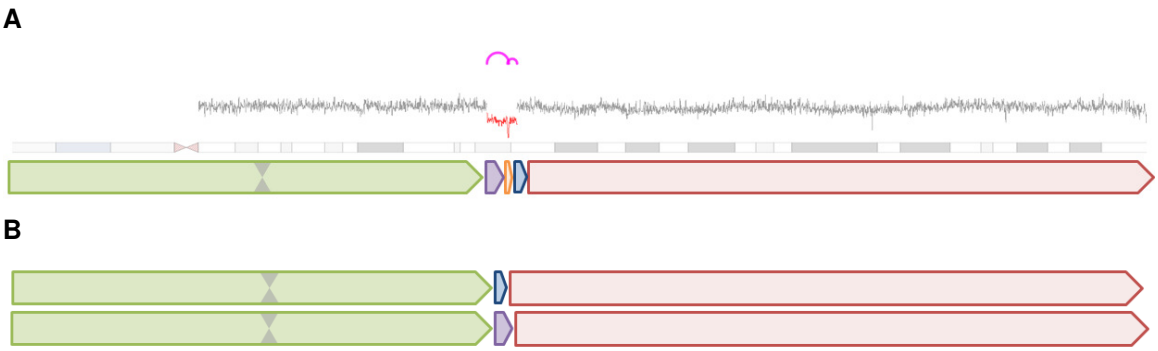


Figure 4A.5: EV88059 chromosome 13 reconstruction
A) chromosome 13 segmented by the CNV and junctions, as shown on the genome plot view B) both copies of the reconstructed chromosome 13, in both 49.958kb-50,091kb (the orange segment) is deleted, while in the top copy 50,091-50,919kb is deleted (the purple segment) and in the bottom copy 47,897-49,958kb (the blue segment) is deleted.

In sample EV88081, SVAtools$_{JD}$ clearly illustrates the presence of a 4-way rearrangement involving chromosomes 1, 4, and 11 (Figure 4A.6), which is missed by karyotype and CMA, and only implied by an atypical FISH result. The karyotype was reported normal, even though these would by cytogenetically visible rearrangements, due to the notorious difficulty in obtaining suitable metaphase preparations from B-ALL samples. CMA missed the 4-way rearrangement because it cannot detect balanced translocations. The FISH probes AFF1/KMT2A(MML) suggest a 3-way break, but reveal nothing about involvement of chromosome 1 in this rearrangement. Reconstructing the genomic rearrangement from the four junctions results in three altered chromosomes. The inv(1) is not a standard 2-junction inversion because the reunion of one inversion break occurs by the unbalanced translocation between chromosome 1 and 11. The inferred molecular karyotype is also written in long form for der(1) to fully describe the location of the inv(1) with respect to the der(1)t(1;11). The junction plots (Figure 4A.7) and the reconstructed chromosomes: der(1), der(4), and der(11) are shown below (Figure 4A.8).

A

Sample	EV88081 RFR: non-Hodgkin's, B-ALL								
Karyotype	46,XX[20]								
FISH	(AFF1x3),(KMT2Ax3),(AFF1 con KMT2Ax1)[167/500]								
CMA	**Locus Start**	**Locus End**	**Position Min**	**Position Max**	**Size**	**Type**			
	1					normal			
	4					normal			
	11					normal			
MPseq with SVAtools$_{JD}$	**Locus A**	**Locus B**	**Position A**	**Position B**	**Size**	**RF**	**FR**	**FF**	**RR**
	1p13.2	4q21.3	111,521,195	87,070,888	NA	0	0	0	11
	1q23.2	11q23.3	159,782,738	118,486,390	NA	18	1	0	0
	4q21.3	11q23.3	87,074,163	118,483,487	NA	0	20	0	0
	1p13.2	1q23.2	111,523,206	159,785,532	48,262,326	0	0	9	0
Inferred Mol. Karyotype	seq der(4)t(1;4)(p13.2;q21.3),der(11)t(4;11)(q21.3:q23.3),der(1)t(1;11)(q23.2;q23.3), inv(1)(p13.2q23.2)								
der(1) in long form	seq der(1)(1qter->1q23.2::1p13.2->1q23.2::11q23.3->11qter								

B

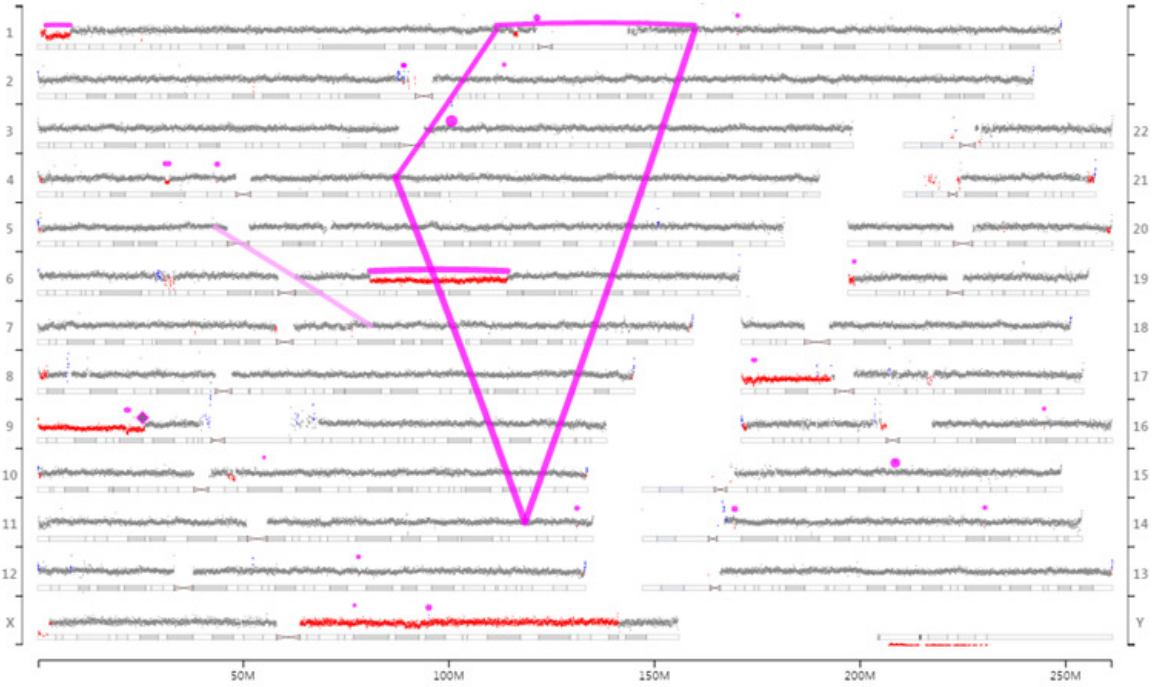
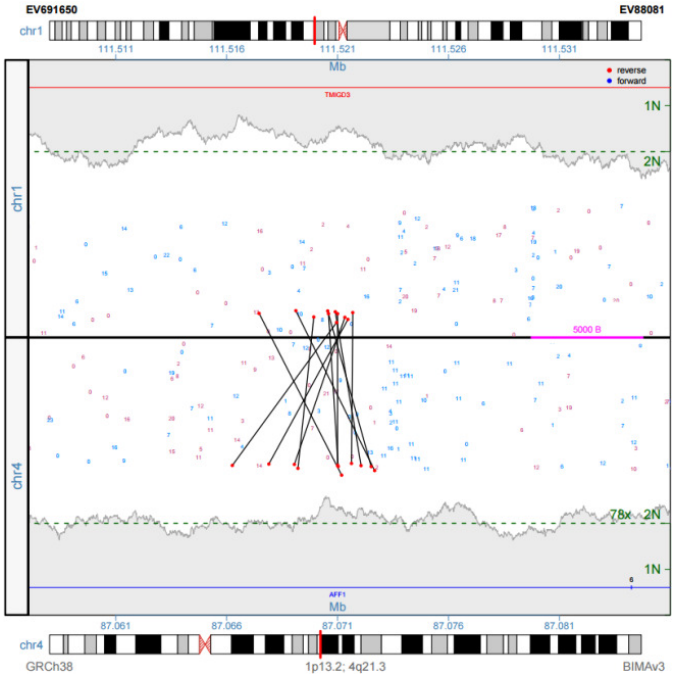
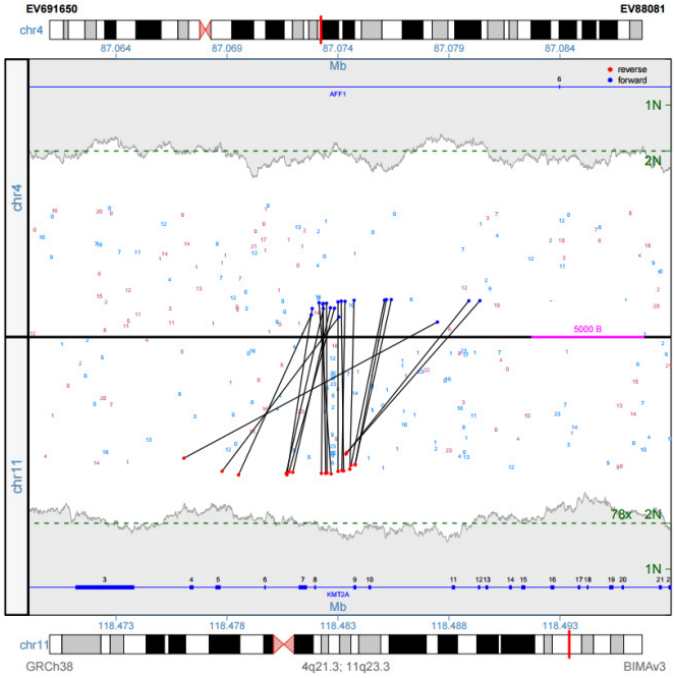


Figure 4A.6: EV88081 a sample with a complex four-way rearrangement.
A) Results of each SV detection method B) genome plot showing the copy number and junctions for the sample, including the four way rearrangement involving chromosomes 1, 4, and 11.

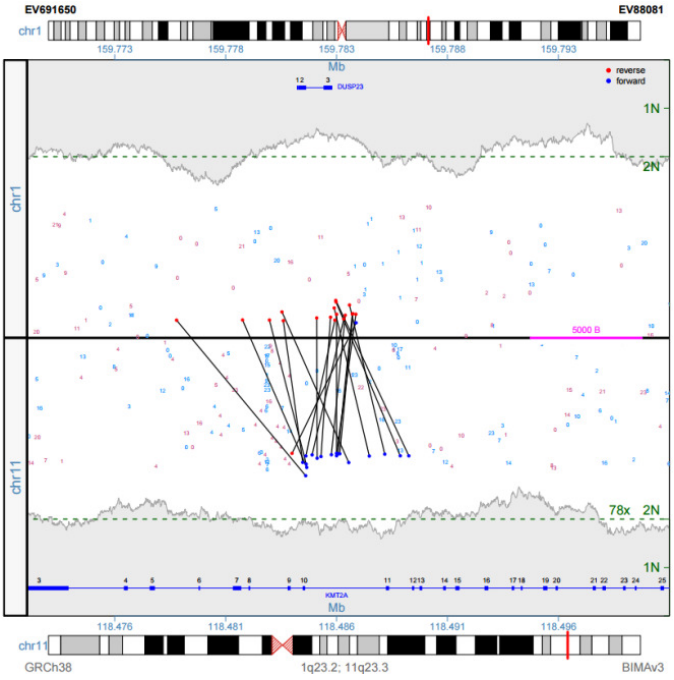
A


EV691650
EV88081
chr1
111,511
111,516
111,521
111,526
111,531
Mb
reverse
forward
TMEM63
1N
2N
chr1
5000 B
chr4
78x
2N
1N
AFF1
Mb
87.061
87.066
87.071
87.076
87.081
chr4
GRCh38
1p13.2; 4q21.3
BIMAv3

B


EV691650
EV88081
chr4
87.064
87.069
87.074
87.079
87.084
Mb
reverse
forward
AFF1
1N
2N
chr4
5000 B
chr11
78x
2N
1N
KMT2A
Mb
118.473
118.478
118.483
118.488
118.493
chr11
GRCh38
4q21.3; 11q23.3
BIMAv3

C



D

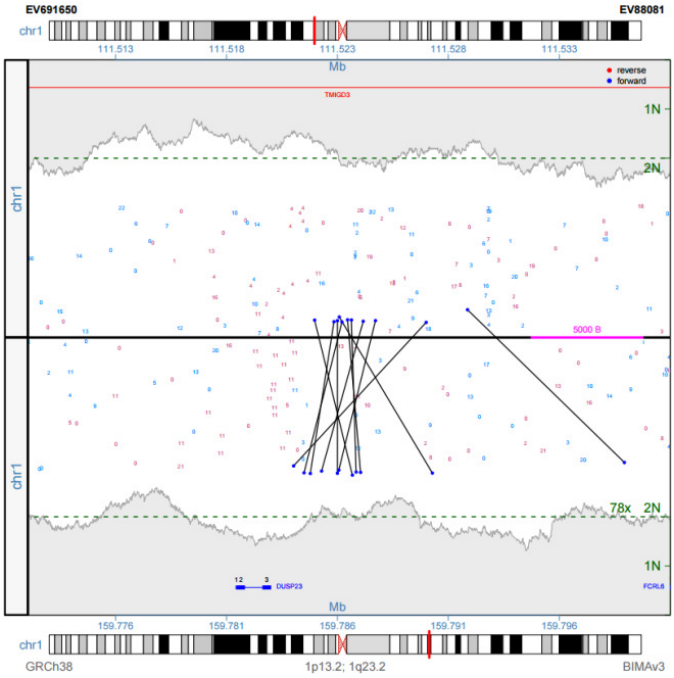


Figure 4A.7: EV88081 junction plots
A) der(4)t(1;4), B) der(11)t(4;11) C) der(1)t(1;11) D) der(1)inv(1)

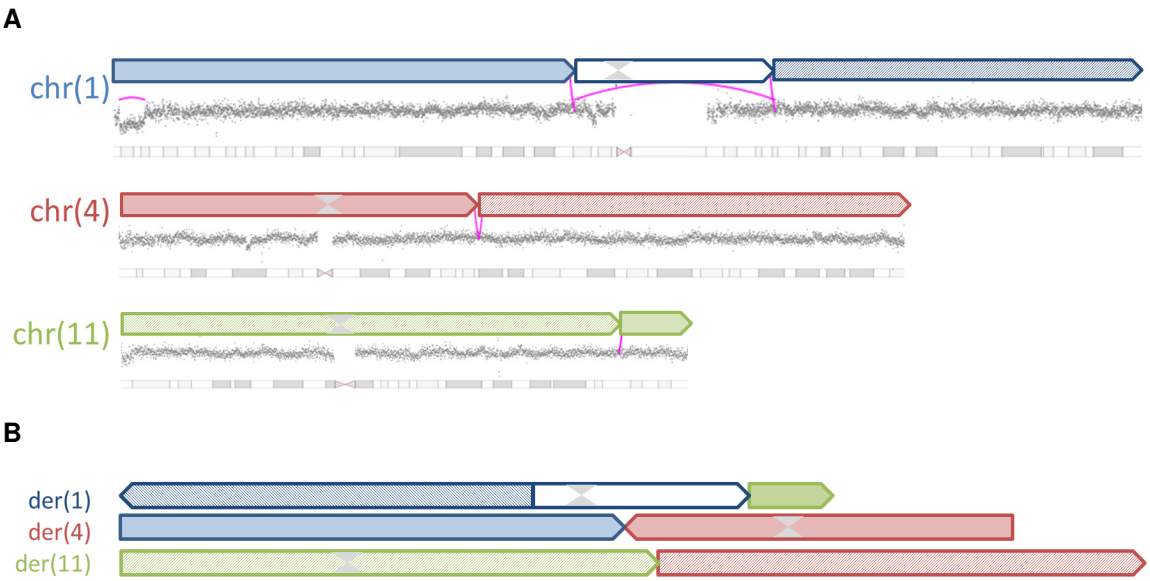


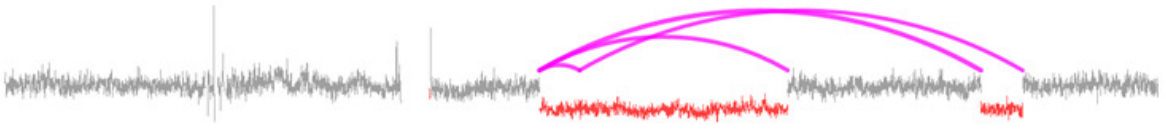
Figure 4A.8: EV88081 chromosome 1, 4, and 11 reconstruction.
A) chromosome 1, 4, and 11 each segmented by the CNV and junctions, as shown on the genome plot view B) the reconstructed derivative chromosomes for der(1), der(4) and der(11). The centromere location is indicated by the gray hourglass shape.

In sample EV88102, the karyotype is normal while CMA reports two apparently independent losses. SVAtools$_{JD}$ reports four junctions to indicate a complex rearrangement (Figure 4A.9). On chromosome 6, SVAtools$_{JD}$ reports 2 copy number losses and 4 junctions, indicating the segment in-between the two deletions is inverted. CMA is the only cytogenetic method that revealed any SV on chromosome 6; however, the structural nature of the rearrangement could not be characterized by CMA. The region plot provides a zoomed in view of the genomic region and illustrates the strand orientation for the breakpoints in all four junctions. Similar to sample EV88081, the inversions on chromosome 6 are not typical two-junction inversions. The reunion of one inversion break occurs with a breakpoint from the other inv(6). The long form of the inferred molecular karyotype is provided to show that two small, 17kb and 7kb, regions, from within the first deleted area are maintained but translocated within chromosome 6. The junction plots (Figure 4A.10) and the reconstructed chromosome (Figure 4A.11) are shown below.

A

Sample	EV88102 RFR: B-ALL								
Karyotype	46,XX[20]								
FISH	No test on chr6 (B-ALL Panel)								
CMA	**Locus Start**	**Locus End**	**Position Min**	**Position Max**	**Size**	**Type**			
	6q14.1	6q22.1	79,537,678	115,628,290	36,090,610	Loss			
	6q24.2	6q25.1	144,025,781	150,147,794	6,122,013	Loss			
MPseq with SVAtools$_{JD}$	**Locus A**	**Locus B**	**Position A**	**Position B**	**Size**	**RF**	**FR**	**FF**	**RR**
	6q14.1	6q14.3	78,834,079	84,715,197	5,881,118	0	0	1	15
	6q14.1	6q24.2	78,811,728	143,698,666	64,886,938	0	0	0	21
	6q14.1	6q22.1	78,816,584	115,311,472	36,494,888	0	0	15	1
	6q14.3	6q25.1	84,708,247	149,831,163	65,122,916	0	0	14	0
Inferred Mol. Karyotype	seq der(6)del(6)(q14.1q22.1);inv(6)(q22.1q24.2);del(6)(q24.2q25.1)								
der(6) in long form	seq der(6)(6pter→6q14.1::6q24.2→6q22.1::6q14.1→6q14.1::6q14.3→6q14.3::6q25.1→6qter								

B



C

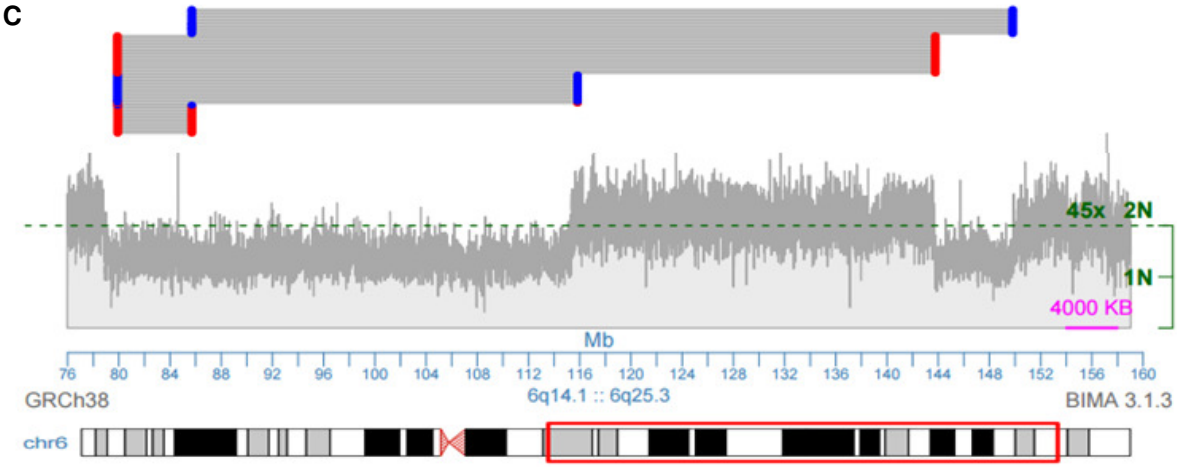
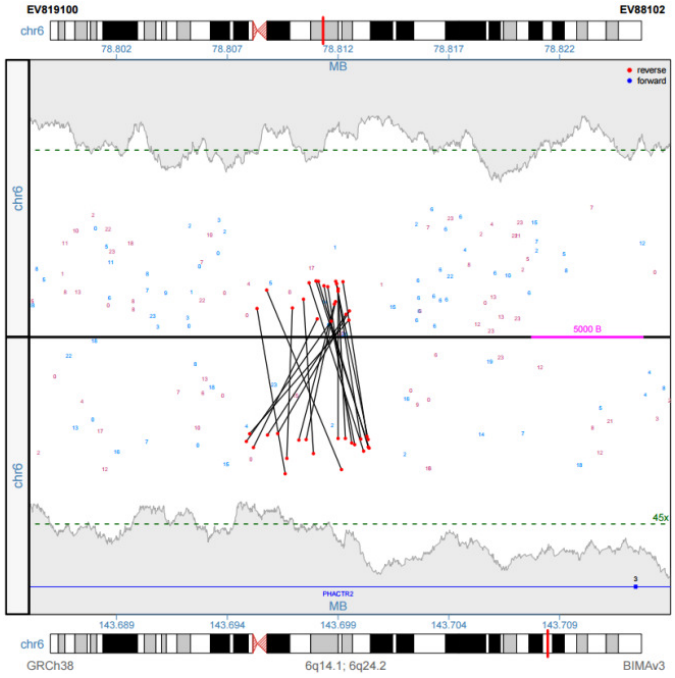
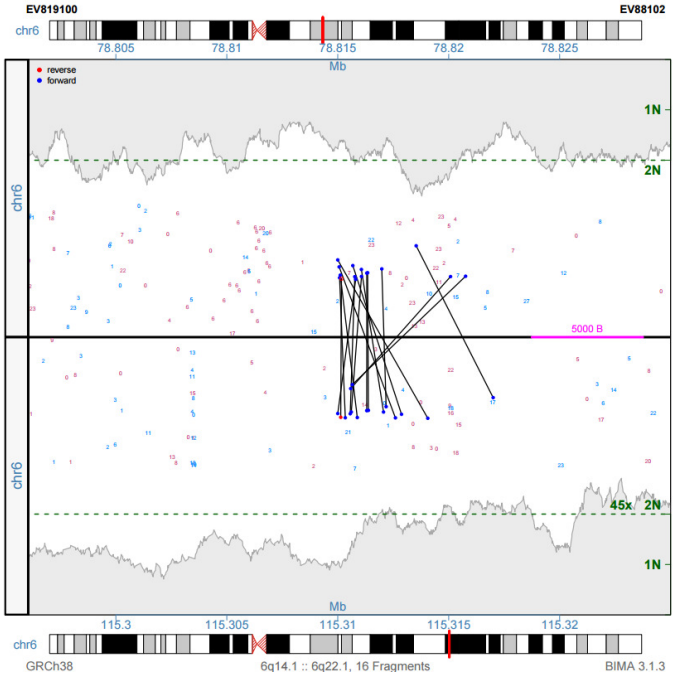


Figure 4A:9 EV88102, a sample with a complex inversion deletion rearrangement
A) Results of each SV detection method B) genome plot view of chromosome 6 showing the copy number and junctions C) region plot of chromosome 6 between 76-160 MB. The dot-line-dot clusters represent the supporting fragments for each of the four junctions reported.

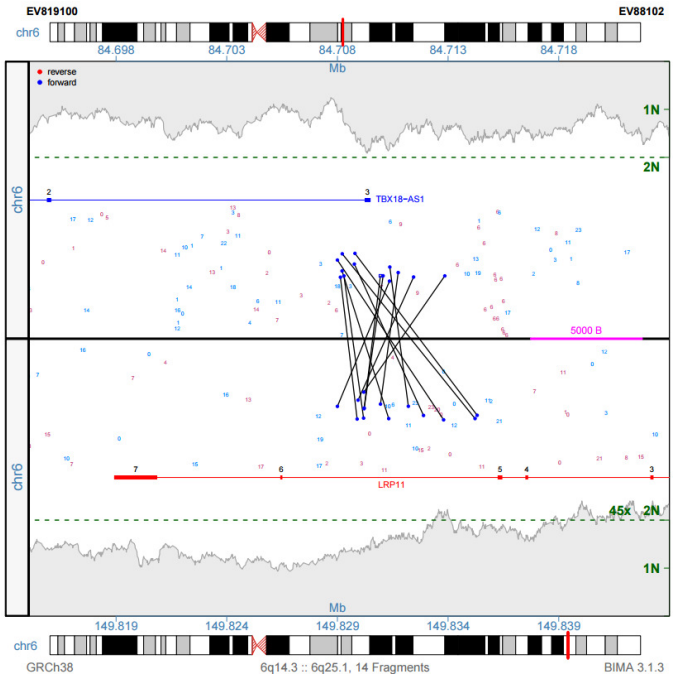
A


EV819100
EV88102
chr6
78.802
78.807
78.812
78.817
78.822
MB
reverse
forward
chr6
5000 B
chr6
45x
PHACTR2
MB
143.689
143.694
143.699
143.704
143.709
chr6
GRCh38
6q14.1; 6q24.2
BIMAv3

B


EV819100
EV88102
chr6
78.805
78.81
78.815
78.82
78.825
Mb
reverse
forward
1N
2N
chr6
5000 B
chr6
45x
2N
1N
Mb
115.3
115.305
115.31
115.315
115.32
chr6
GRCh38
6q14.1 :: 6q22.1, 16 Fragments
BIMA 3.1.3

C



D

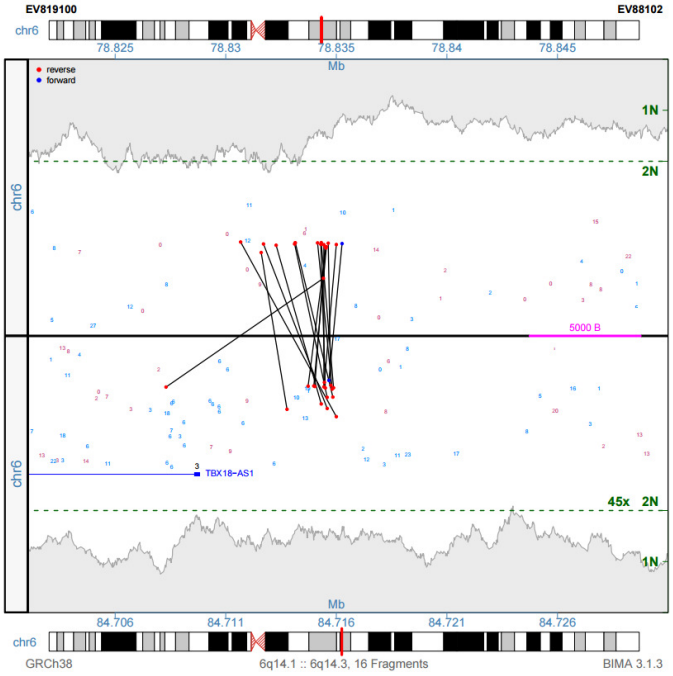
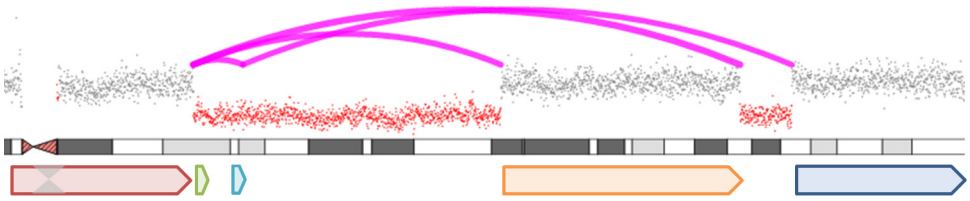


Figure 4A.10: EV88102 junction plots
A) der(6)inv(6;6)(q14.1q24.2), B) der(6)inv(6;6)(q14.1q22.1) C) der(6)inv(6;6)(q14.3;25.1) D) der(6)inv(6;6)(q14.1q14.3)

A



B

Figure 4A.11: EV88102 chromosome 6 reconstruction
A) Chromosome 6 p arm segmented by the CNV and junctions, as shown on the genome plot view B) the reconstructed derivative chromosome 6.

In the last example EV88099, FISH reveals there is a gain or disruption on chromosome 10 and karyotype reports a balanced translocation but not involving the same location as FISH. CMA was normal. SVAtools$_{JD}$ reports a balanced translocation, an inversion, and predicts a clinically relevant fusion gene (Figure 4A.12). The FISH result (MLLT10x3)[355/500] can be interpreted as a gain or disruption of the region, but requires additional testing to determine which scenario and any possible partners. The karyotype, t(X;10)(q24;p13), reports a translocation, but not at MLLT10. SVAtools$_{JD}$ reported two rearrangements: t(X;10)(p11.4;p12.31) and inv(X)(p11.22;q22.3). The balanced translocation predicts the fusion DDX3X->MMLT10. Neither FISH, karyotype, nor CMA reported this junction or predicted this fusion. FISH was not specific enough. The inversion likely complicated the ability for karyotype to identify the breakpoints for the translocation. CMA cannot detect balanced rearrangements. The junction plots (Figure 4A.13) and the reconstructed chromosome (Figure 4A.14) are shown below.

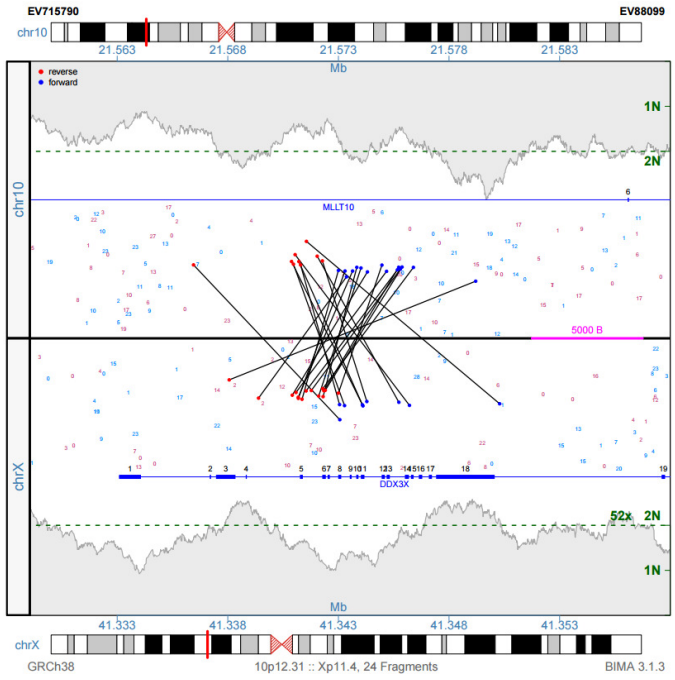
A

Sample	EV88099 RFR: T-ALL								
Karyotype	48,X,t(X;10)(q24;p13),+1,+4[6]/46,XX[14]								
FISH	(STIL,TAL1)x3[125/200]/(MLLT10x3)[355/500]								
CMA	**Locus Start**	**Locus End**	**Position Min**	**Position Max**	**Size**	**Type**			
	10					normal			
	X					normal			
MPseq with SVAtools$_{JD}$	**Locus A**	**Locus B**	**Position A**	**Position B**	**Size**	**RF**	**FR**	**FF**	**RR**
	10p12.31	Xp11.4	21,573,217	41,343,223	NA	9	15	0	0
	Xp11.22	Xq22.3	52,359,209	105,395,051	53,035,842	0	1	5	5
Inferred Mol. Karyotype	seq 48,X,t(X;10)(p11.4;p12.31),+1,+4,inv(X)(p11.22;q22.3)								

B

Figure 4A.12: EV88099 a sample with a balanced translocation.
A) Results of each SV detection method B) genome plot showing the copy number and junctions for the sample, including the balanced translocation t(X;10) and the inversion inv(X).

A



B

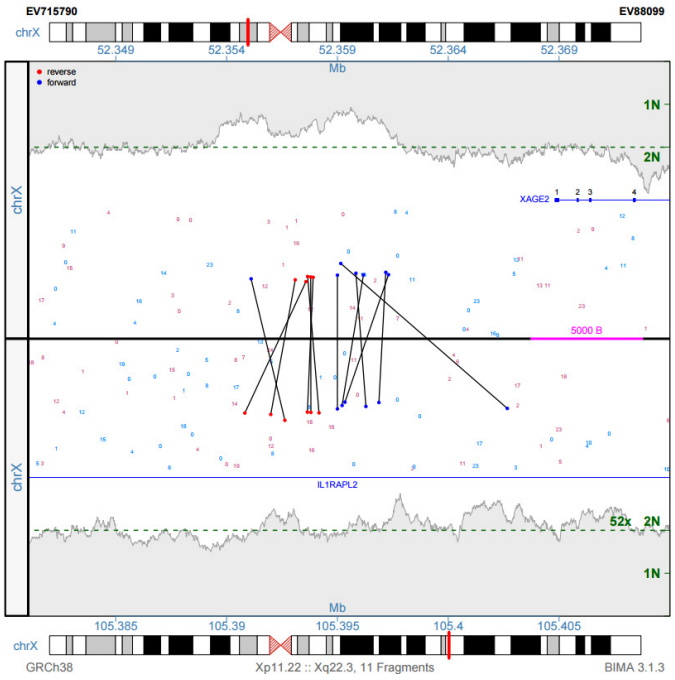


Figure 4A.13: EV88099 junction plots
A) t(X;10) B) inv(X)

Figure 4A.14: EV88099 chromosome 10 and X reconstructions
A) chromosome 10 and X each segmented by the CNV and junctions, as shown on the genome plot view B) the reconstructed derivative chromosomes for der(10), and der(X). The centromere location is indicated by the gray hourglass shape.

4B: Resolution of breakpoints

Fragments with split reads were detected in three of the 18 junctions from the five samples described above. The automated SVAtools pipeline includes split reads for junction detection, but not for reporting the exact breakpoint position. However, the split reads can be analyzed by manual methods to pinpoint the exact molecular breakpoint. Comparison of the actual breakpoint provided by split reads demonstrates that the estimated breakpoint positions reported were within 24-151 bps (Table 4B).

Table 4B: MPseq with SVAtools$_{JD}$ breakpoint resolution.*

			As reported by split reads		As reported by pipeline		difference	
Sample	**Chr A**	**Chr B**	**Position A**	**Position B**	**Position A**	**Position B**	**Pos. A**	**Pos. B**
EV88100	3q21.3	3q26.2	128,543,682	168,886,550	128,543,833	168,886,484	151	66
EV88100	5q14.2	5q33.2	82,327,433	153,899,622	82,327,333	153,899,576	100	46
EV88102	6q14.1	6q22.1	78,815,300	115,310,195	78,815,276	115,310,166	24	29

***Split reads have 1bp resolution revealing the molecular breakpoint. Listed are three junctions which each contained a split read. The difference in the breakpoint position reported by the pipeline versus the breakpoint position as determined by the split reads, ranges between 24-151 bps**

4C: Fusion genes and truncated genes

SVAtools$_{JD}$ can predict fusion and truncated genes based on the strand orientation of the genes at the junction and strand orientation of fragments crossing the junction. The high breakpoint resolution of MPseq also allows SVAtools$_{JD}$ to pinpoint which intron or exon is bisected. From the 18 junctions in the five examples described earlier, seven fusion genes and six truncated genes are predicted. (Table 4C)

Table 4C: SVAtools$_{JD}$ gene prediction. *

Sample	Locus A	Locus B	Position A	Position B	Size	Gene prediction	RF	FR	FF	RR
EV88100	11q23.1	17p13.1	112,185,770	8,156,561	NA	VAMP2 -> BCO2	0	0	12	1
EV88059	13q14.2	13q14.2	47,897,178	50,091,564	2,194,386	DLEU1 truncation DLEU2 truncation	13	0	0	0
EV88059	13q14.2	13q14.3	49,958,440	50,918,770	960,330	RNASEH2B truncation	13	0	0	0
EV88081	1p13.2	4q21.3	111,521,879	87,072,561	NA	AFF1 -> TMIGD3	0	0	0	11
EV88081	1q23.2	11q23.3	159,783,606	118,484,864	NA	DUSP23 -> KMT2A	18	1	0	0
EV88081	4q21.3	11q23.3	87,072,957	118,484,261	NA	KMT2A -> AFF1	0	20	0	0
EV88081	1p13.2	1q23.2	111,523,206	159,785,532	48,262,326	TMIGD3 truncation	0	0	9	0
EV88102	6q14.1	6q24.2	78,812,237	143,700,068	64,887,831	PHACTR2 truncation	0	0	0	21
EV88102	6q14.3	6q25.1	84,708,036	149,828,844	65,120,808	LRP11->TBX18-AS1	0	0	14	0
EV88099	10p12.31	Xp11.4	21,573,217	41,343,223	NA	DDX3X->MLLT10 MLLT10-> DDX3X	9	15	0	0
EV88099	Xp11.22	Xq22.3	52,359,209	105,395,051	53,035,842	IL1RAPL2 truncation	0	1	5	5

***SVAtools$_{JD}$ will detect and report possible gene fusions or truncations based on the rearrangement configuration. Gene fusions or truncations were detected in 13 of 18 junctions from the 5 examples. The "Gene prediction" column lists the gene(s) involved and the orientation of the fusion. All positions are listed in GRCh38 coordinates. The RF, FR, FF, and RR columns refer to the number of fragments mapping to the corresponding strand orientation where R is reverse and F is forward, for position A and position B respectively.**

Chapter 5: Discussion and Summary

Fragments produced by MPseq include both concordant (95%) and discordant (5%) fragments. SVAtools$_{JD}$ detects junctions by clustering the discordant fragments with read-pair and split read approaches, and detects CNVs from the concordant fragments with the read depth approach. Few SV callers include all three approaches: read-pair, split read, and read depth (Tattini, D'Aurizio, and Magi 2015; Lin et al. 2015). Because SVAtools detects both junctions and CNVs, SVAtools can integrate the results of both to further refine SV calls, improve sensitivity and breakpoint resolution. While the limits of detection are still being validated for MPseq, the current BMD-SV pipeline methods used in this study indicate detection of breakpoints observed in more than 25% of cells to be highly sensitive. With these criteria, SVAtools$_{JD}$ detected 97%, 100% and 90% of the breakpoints detected by karyotype, FISH, and CMA respectively. If SVAtools' CNV detection without a supporting junction is included, 93% (265/285) of breakpoints are reported by SVAtools. Additionally, when only breakpoints detected in at least 25% of cells are considered, SVAtools detected 99% (263/266) of the breakpoints.

While karyotype and FISH can both report balanced and unbalanced rearrangements, there are drawbacks to each. FISH is limited to the probed area. FISH is not whole genome and the investigator must know which genomic region to interrogate to perform a successful FISH test. Karyotype requires dividing cells and the standard of testing only 20 cells may not be statistically representative of the sample. Often karyotype is returned normal due to more metaphase quality and resolution is limited to the cytoband level. CMA is capable of providing higher breakpoint resolution than either karyotype or FISH; however CMA cannot detect balanced rearrangements or provide insight as to genomic structure when a copy number change is detected. Clinically, a gain or loss may not be significant, but a rearrangement involving these CNVs would be missed and may be significant. Identifying disrupted genes and regulatory elements is important for prognosis and treatment. Dosage effects of genes within gains and losses can be inferred by CMA, but the impact on a gene at the junction of a gain or losses cannot.

Junction detection, regardless of the algorithm employed, has some inherent limitations. Junction detection is limited to non-terminal rearrangements because terminal rearrangements only have one break and no junction, for example no junction will be reported if a gain at a terminal edge is a tandem duplication. However, terminal CNVs, including whole arm and whole chromosome CNVs can be detected by SVAtools' CNV detection methods. Junction detection will also fail to detect junctions in regions excluded from the reference genome including the heterochromatic

regions of chromosome 1, 9 and 16. Similar to CMA, losses and gains of the p-arm of the acrocentric chromosomes cannot be detected, while not clinically significant, rearrangements involving these locations would also be missed and may be significant. Junction detection also struggles in genomic regions with high sequence similarity. Fortunately, the long insert size of mate-pair fragments facilitates mapping of concordant reads in repeat areas, an advantage of MPseq over traditional paired-end sequencing. Thus read depth coverage often catches these missed junctions, however junctions occurring at the centromere are often filtered out due to high homology scores. Finally, junction detection can report low-level rearrangements, but like most algorithms, not without reporting false positives. For junctions detected by cytogenetic methods at very low levels, changes to the BMD-SV pipeline to improve coverage would facilitate increased junction detection sensitivity. Such changes could include sequencing at a higher depth, limiting normal cell contamination in the library preparation, and increasing the length of the mate-pair fragment insert size. Single nucleotide variant (SNV) detection is not possible with the current BMD-SV pipeline methods. Referring back to the example from section 2C, a depth of coverage suitable for detecting junctions, 93.75x bridged coverage, results in a depth of coverage not suitable for detecting single nucleotide variants (SNVs), 6.25x base coverage.

In clinical practice often two or more cytogenetic assays are performed to detect clinically relevant rearrangements and copy number changes. One test will be ordered and then reflexed to another, adding time and cost to the analysis. While karyotype detects a number of clinically-relevant rearrangements, CMA or FISH are needed to detect submicroscopic imbalances. For the CMA first approach, the pattern of gain and loss detected is highly suggestive and the structural rearrangement can be inferred (sample EV88100) but in complex cases reflexing to FISH or karyotype is often required. Even after performing karyotype, FISH and CMA, the rearrangements may not be full described (samples EV88059, EV88081, EV88102, EV88099). In each of the examples described, SVAtools$_{JD}$ fully characterized each rearrangement and reported disrupted and/or fusion genes. In addition, the breakpoint resolution for MPseq with SVAtools$_{JD}$ is half the fragment size, but SVAtools$_{JD}$ typically reports the breakpoint to less than a couple hundred bases and can detect the exact breakpoint manually using split reads. This representative set of clinical cases demonstrates the potential advantage of performing a single MPseq test with SVAtools$_{JD}$ over running multiple cytogenetic tests and eliminating the need for extensive reflex FISH. Future work will include full validation of the clinical performance; both sensitivity and specificity, of MPseq with SVAtools to more clearly define the clinical readiness for this assay to replace current gold standard cytogenetic techniques.

Bibliography

Alkan, C., B. P. Coe, and E. E. Eichler. 2011. 'Genome structural variation discovery and genotyping', *Nat Rev Genet*, 12: 363-76.

Boddicker, R. L., G. L. Razidlo, S. Dasari, Y. Zeng, G. Hu, R. A. Knudson, P. T. Greipp, J. I. Davila, S. H. Johnson, J. C. Porcher, J. B. Smadbeck, B. W. Eckloff, D. D. Billadeau, P. J. Kurtin, M. A. McNiven, B. K. Link, S. M. Ansell, J. R. Cerhan, Y. W. Asmann, G. Vasmatzis, and A. L. Feldman. 2016. 'Integrated mate-pair and RNA sequencing identifies novel, targetable gene fusions in peripheral T-cell lymphoma', *Blood*, 128: 1234-45.

Catic, Aida, Amina Kurtovic-Kozaric, Sarah H. Johnson, George Vasmatzis, Michael R. Pins, and Jillene Kogan. 2017. 'A novel cytogenetic and molecular characterization of renal metanephric adenoma: Identification of partner genes involved in translocation t(9;15)(p24;q24)', *Cancer Genetics*, 214-215: 9-15.

Chen, J. M., D. N. Cooper, C. Ferec, H. Kehrer-Sawatzki, and G. P. Patrinos. 2010. 'Genomic rearrangements in inherited disease and cancer', *Semin Cancer Biol*, 20: 222-33.

Drucker, T. M., S. H. Johnson, S. J. Murphy, K. W. Cradic, T. M. Therneau, and G. Vasmatzis. 2014. 'BIMA V3: an aligner customized for mate pair library sequencing', *Bioinformatics*, 30: 1627-9.

Feldman, A. L., A. Dogan, D. I. Smith, M. E. Law, S. M. Ansell, S. H. Johnson, J. C. Porcher, N. Ozsan, E. D. Wieben, B. W. Eckloff, and G. Vasmatzis. 2011. 'Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing', *Blood*, 117: 915-9.

Feuk, L., A. R. Carson, and S. W. Scherer. 2006. 'Structural variation in the human genome', *Nat Rev Genet*, 7: 85-97.

Gao, G., S. H. Johnson, J. L. Kasperbauer, B. W. Eckloff, N. M. Tombers, G. Vasmatzis, and D. I. Smith. 2014. 'Mate pair sequencing of oropharyngeal squamous cell carcinomas reveals that HPV integration occurs much less frequently than in cervical cancer', *J Clin Virol*, 59: 195-200.

Harris, F. R., I. V. Kovtun, J. Smadbeck, F. Multinu, A. Jatoi, F. Kosari, K. R. Kalli, S. J. Murphy, G. C. Halling, S. H. Johnson, M. C. Liu, A. Mariani, and G. Vasmatzis. 2016. 'Quantification of Somatic Chromosomal Rearrangements in Circulating Cell-Free DNA from Ovarian Cancers', *Sci Rep*, 6.

Jang, J. S., X. Wang, P. T. Vedell, J. Wen, J. Zhang, D. W. Ellison, J. M. Evans, S. H. Johnson, P. Yang, W. R. Sukov, A. M. Oliveira, G. Vasmatzis, Z. Sun, J. Jen, and E. S. Yi. 2016. 'Custom Gene Capture and Next-Generation Sequencing to Resolve Discordant ALK Status by FISH and IHC in Lung Adenocarcinoma', *J Thorac Oncol*, 11: 1891-900.

Kearney, H. M., S. T. South, D. J. Wolff, A. Lamb, A. Hamosh, and K. W. Rao. 2011. 'American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities', *Genet Med*, 13: 676-9.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder. 2007. 'Paired-end mapping reveals extensive structural variation in the human genome', *Science*, 318: 420-6.

Lin, K., S. Smit, G. Bonnema, G. Sanchez-Perez, and D. de Ridder. 2015. 'Making the difference: integrating structural variation detection tools', *Brief Bioinform*, 16: 852-64.

Medvedev, P., M. Stanciu, and M. Brudno. 2009. 'Computational methods for discovering structural variation with next-generation sequencing', *Nat Methods*, 6: S13-20.

Murphy, S. J., M. C. Aubry, F. R. Harris, G. C. Halling, S. H. Johnson, S. Terra, T. M. Drucker, M. K. Asiedu, B. R. Kipp, E. S. Yi, T. Peikert, P. Yang, G. Vasmatzis, and D. A. Wigle. 2014.

'Identification of independent primary tumors and intrapulmonary metastases using DNA rearrangements in non-small-cell lung cancer', *J Clin Oncol*, 32: 4050-8.
Murphy, S. J., J. C. Cheville, S. Zarei, S. H. Johnson, R. A. Sikkink, F. Kosari, A. L. Feldman, B. W. Eckloff, R. J. Karnes, and G. Vasmatzis. 2012. 'Mate pair sequencing of whole-genome-amplified DNA following laser capture microdissection of prostate cancer', *DNA Res*, 19: 395-406.
Newman, S., K. E. Hermetz, B. Weckselblatt, and M. K. Rudd. 2015. 'Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints', *Am J Hum Genet*, 96: 208-20.
Smadbeck, James B., Sarah H. Johnson, Stephanie A. Smoley, Athanasios G. Gaitatzes, Travis M. Drucker, Roman M. Zenka, Farhad Kosari, Stephen J. Murphy, Nicole Hoppman, Umut Aypar, William R. Sukov, Robert Jenkins, Hutton M. Kearney, Andrew L. Feldman, and George Vasmatzis. 'Copy Number Variant Analysis using Genome-Wide Mate-Pair Sequencing', *in submission*.
South, Sarah T. 2011. 'Chromosomal Structural Rearrangements: Detection and Elucidation of Mechanisms Using Cytogenomic Technologies', *Clinics in Laboratory Medicine*, 31: 513-24.
Stankiewicz, P., and J. R. Lupski. 2010. 'Structural variation in the human genome and its role in disease', *Annu Rev Med*, 61: 437-55.
Tattini, L., R. D'Aurizio, and A. Magi. 2015. 'Detection of Genomic Structural Variants from Next-Generation Sequencing Data', *Front Bioeng Biotechnol*, 3: 92.
Vasmatzis, G., S. H. Johnson, R. A. Knudson, R. P. Ketterling, E. Braggio, R. Fonseca, D. S. Viswanatha, M. E. Law, N. S. Kip, N. Ozsan, S. K. Grebe, L. A. Frederick, B. W. Eckloff, E. A. Thompson, M. E. Kadin, D. Milosevic, J. C. Porcher, Y. W. Asmann, D. I. Smith, I. V. Kovtun, S. M. Ansell, A. Dogan, and A. L. Feldman. 2012. 'Genome-wide analysis reveals recurrent structural abnormalities of TP63 and other p53-related genes in peripheral T-cell lymphomas', *Blood*, 120: 2280-9.
Weckselblatt, B., and M. K. Rudd. 2015. 'Human Structural Variation: Mechanisms of Chromosome Rearrangements', *Trends Genet*, 31: 587-99.
Weischenfeldt, J., O. Symmons, F. Spitz, and J. O. Korbel. 2013. 'Phenotypic impact of genomic structural variation: insights from and for human disease', *Nat Rev Genet*, 14: 125-38.