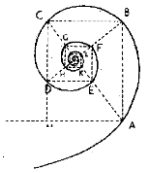
**Università degli Studi di Milano**

**Scuola di Dottorato in Medicina Molecolare**



Curriculum di Genomica, Proteomica e Tecnologie Correlate

Ciclo XXVI

Anno Accademico 2012-2013

**Dottorando**: Matteo BARCELLA

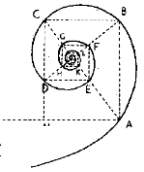
# MAPPING BCR-ABL1 FUSION POINTS IN CHRONIC MYELOID LEUKEMIA BY NEXT GENERATION SEQUENCING

**Direttore del Dottorato**: Prof. Mario Clerici

**Tutore**: Prof. ssa Cristina BARLASSINA

**UNIVERSITÀ DEGLI STUDI DI MILANO**

**SCUOLA DI DOTTORATO IN MEDICINA MOLECOLARE**



CICLO XXVI

Anno Accademico 2012/2013

TESI DI DOTTORATO DI RICERCA

**MED/04**

# MAPPING BCR-ABL1 FUSION POINTS IN CHRONIC MYELOID LEUKEMIA BY NEXT GENERATION SEQUENCING

Dottorando: Matteo BARCELLA

Matricola N° R09339

TUTORE: Prof. Cristina BARLASSINA

DIRETTORE DEL DOTTORATO: Prof. Mario CLERICI

## SOMMARIO

La leucemia mieloide cronica (CML) è una malattia mieloproliferativa citogeneticamente caratterizzata da una traslocazione reciproca tra il braccio lungo del cromosoma 9 e del cromosoma 22 t(9;22), che porta alla formazione della chimera BCR-ABL1. Questo gene codifica per una tirosina chinasi deregolata con attività oncogenica. Attualmente i pazienti affetti da CML sono sottoposti a trattamento con Imatinib, un inibitore della tirosina chinasi. La quantificazione del trascritto BCR-ABL1 mediante qRT-PCR viene utilizzata di routine nel follow-up del paziente in terapia per valutare la risposta farmacologica individuale. Tuttavia, questa tecnica non può identificare cellule leucemiche trascrizionalmente silenti che possono essere presenti nella malattia minima residua (MRD). Per poter monitorare la MRD è necessario quindi sviluppare un saggio di qPCR sulle sequenze genomiche di BCR-ABL1, che sono paziente-specifiche. Precedenti risultati del gruppo del Prof. Giovanni Porta (unpublished) hanno dimostrato che nel 30% dei tempi di follow-up valutati l'mRNA non viene rilevato mentre il test sul DNA risulta positivo, indicando la possibile presenza di cellule trascrizionalmente silenti. I punti di rottura cromosomici in questi pazienti sono stati caratterizzati tramite una laboriosa tecnica di long-range PCR e di clonaggio, non adatta per un'applicazione clinica. Al fine di superare questo limite tecnico, abbiamo sviluppato un saggio per isolare le regioni di ABL1 e BCR in cui mappano i punti di rottura canonici, che danno origine a differenti trascritti BCR-ABL1. Le regioni isolate sono state poi sequenziate con metodi di nuova generazione. Abbiamo identificato con successo i punti di fusione di BCR-ABL1 in 9 campioni su 10 raggiungendo una accuratezza al singolo nucleotide. Tutti i risultati sono stati validati con sequenziamento Sanger. Sui breakpoint individuati saranno sviluppati saggi di qPCR utili per monitorare la MRD. Questo progetto è stato sviluppato in collaborazione con il Prof. Porta dell'Università dell'Insubria.

## ABSTRACT

Chronic myeloid leukemia (CML) is a myeloproliferative disorder cytogenetically characterized by a reciprocal translocation between the long arms of chromosome 9 and 22 t(9;22) that leads to the formation of the BCR-ABL1 fusion gene, coding for a deregulated tyrosine kinase with oncogenic activity. In clinical routine, mRNA amount of the chimeric transcript is considered proportional to the leukemic clone and is used for the molecular monitoring of patients. However, qRT-PCR cannot identify transcriptionally silent leukemic cells that can be present in minimal residual disease (MRD). To monitor MRD it is necessary to develop a qPCR assay on DNA sequences spanning BCR-ABL1, that are patient specific. Previous results obtained by Prof. G. Porta's group (unpublished) have demonstrated that DNA detection is positive while mRNA is not in 30% of time points, indicating the presence of transcriptionally silent cells. Breakpoints in these patients were characterized by laborious long-range PCR and cloning not suitable for a clinical application. To overcome this limiting step we set up a DNA capturing assay to target all kind of breakpoints that give rise to different BCR-ABL1 transcripts. Captured regions were then sequenced with a next generation protocol. The idea was to use the identified patient specific breakpoints to setup qPCR assays to monitor MRD. We successfully identified BCR-ABL1 fusion points in 9 over 10 samples, with single nucleotide accuracy, by setting up a bioinformatics workflow specifically developed for this purpose. All findings were validated with Sanger sequencing. This project was performed in collaboration with Prof. Giovanni Porta of University of Insubria.

## INDEX

## INDEX OF THE FIGURES

## INDEX OF THE TABLES

## LIST OF SYMBOLS

| | |
|---|---|
| µ-BCR | Micro Breakpoint Cluster Region |
| A-EJ | Alternative non-homologous end joining |
| ALL | Acute lymphoid leukemia |
| alloHSCT | Allogenic hematopietic stem cell transplantation |
| AML | Acute myeloid leukemia |
| AP | Accelerated phase |
| ATP | Adenosine triphosphate |
| BCR | Breakpoint cluster region gene |
| bp | Nucleotide base pair |
| BP | Blastic phase |
| BQSR | Base quality score recalibration |
| BWA | Burrows wheeler aligner |
| CBA | Chromosome banding analysis |
| ChIP | Chromosome immune precipitation |
| CML | Chronic myeloid leukemia |
| C-NHEJ | Classic non-homologous end joining |
| CNV | Copy number variation |
| CP | Chronic phase |
| ddNTP | Di-deoxy nucleosite triphosphate |
| der(9) | Derivative chromosome 9 |
| dsDNA | Double strand DNA |
| EUTOS | EUropean Treatment Outcome Study |
| FISH | Fluorescent in-situ hybridization |
| gaIIx | Genome analyzer IIx |
| GASV | Geometric Analysis of Structural Variants |
| GATK | Genome analysis toolkit |
| %GC | Guanine cytosine percentage |
| gDNA | Genomic DNA |
| GTP | Guanoside triphosphate |
| GWAS | Genome wide association studies |
| HSC | Hematopoietic stem cells |
| I-FISH | Interphase – fluorescent in-situ hybridization |
| IFNα | Interpheron alpha |
| IGV | Integrative genome viewer |
| INDELs | Insertions / deletions |
| Kb | Kilobases |

| | |
|---|---|
| KDa | K Dalton |
| LDI PCR | Long-distance-inverse polymerase chain reaction |
| LRAI | Local Realignment Around Indels |
| m-BCR | Minor Breakpoint Cluster Region |
| MLR-PCR | Long-range multiplex polymerase chain reaction |
| MMEJ | Microhomology mediated end joining |
| MMR | Major molecular response |
| MPD | Myeloproliferative disease |
| MQ | Mapping quality |
| MRD | Minimal Residue Disease |
| NGS | Next Generation Sequencing |
| NHEJ | Non-homologous end joining |
| NLS | Nuclear localization signal |
| Npot | Nb_pairs_order_threshold |
| Npt | Nb_pairs_threshold |
| PB | Peripheral Blood |
| PCR | Polymerase Chain Reaction |
| Ph | Philadelphia chromosome |
| PIP2 | phosphatidylinositol-4,5-bisphosphate |
| qPCR | Quantitative PCR |
| qRT-PCR | Quantitative reverse transcriptase PCR |
| Rho-GAP | Rho GTPase activating protein |
| Rho-GEF | Rho guanine nucleotide exchange factor |
| RM | Repeat Masker |
| RT-PCR | reverse transcriptase PCR |
| SAM | Sequence alignment/map format |
| SEER | Surveillance, Epidemiology, End Results Program |
| SNPs | Single Nucleotide Polymorphism |
| SNV | Single Nucleotide Variant |
| SV | Structural Variant |
| TKI | Tyrosine Kinase Inhibitor |
| WBC | White Blood Count |
| WHO | World Health Organization |
| WM | Window Masker |

# 1. INTRODUCTION

## *1.1 Next-Generation Sequencing (NGS)*

Next-generation sequencing (NGS), refers to sequencing technologies that came after the automated Sanger method. Sanger method [1] is considered the first-generation sequencing technology and it was used to determine the DNA sequence (5375bp) of the bacteriophage phi X 174 in the 1977, the first genome ever sequenced [2]. Improvements to the method, including a semi-automation, were developed in early 80's by Smith, which introduced the four-color Sanger sequencing, using four fluorescently labeled dideoxynucleotides (ddNTP) for each base and enabling optical detection [3]. This method combined with capillary elettrophoresis leads to the development of the first fully automated DNA sequencing system, the ABI 370 developed by Applied Biosystem [4]. Additional improvements continued though the beginning of the new millennium, enabling parallelized sequencing of many DNA fragments up to 1 kilobases in length with an accuracy higher that 99,9% [5]. The development of the newer technologies was enhanced by the initiation of the Human Genome Project at the beginning of the 90'. The sequencing of the human genome was an exciting challenge between the public institution, the Human genome consortium, and the private Celera Genomics, headed by Craig Venter. This competition speeded up the production of the first draft of the human genome in 2001 [6] [7] and the final version in the 2003 [8]. The great investment on this project (3 billion US dollars) triggered the development of cheaper and higher-throughput techniques as well as bioinformatics algorithms and tools for the analysis of data produced by these machines. In 2005, 454 Life Sciences introduced the GS20, the first NGS platform on the market. The technique was validated by the combination of the single-emulsion PCR with the pyrosequencing of the Mycoplasma genitalia (580.069 bp) with a coverage of 96% and with an accuracy of 99.96% in a

single run of GS20 [9]. The principle of the pyrosequencing technique is based on the “sequencing by synthesis”, which depends on the detection of pyrophosphate release on nucleotide incorporation. Differently, Sanger method detects the nucleotide incorporation by the chain termination with ddNTP. The evolution of the NGS techniques consisted in the improvement of the detection of the next-added fluorescently labeled base by using CCD camera, performing a sort of snapshot of the growing DNA chain. This approach was applied to a large number of samples in parallel, attached either to a planar support or to beads, minimizing reaction volumes so as to lead to miniaturized sequencing systems. In the middle of 00s, several platforms emerged on the market among which Roche/454, Illumina, Applied Biosystem. By different approaches, each technology seeks to amplify single strands of a fragment library and to perform sequencing reactions on the amplified strands. In this work the Illumina platform was used and the sequencer adopted was the Genome Analyzer IIx (gaIIx).

### *1.2 Illumina Sequencing Technology*

Illumina sequencing technology is based on “sequencing by synthesis” method that utilizes reversible termination chemistry of nucleotide analogues [10]. As previously mentioned, there are two main core-objective in a typical NGS sequencing experiment: the amplification of a fragments library and sequencing of the amplified library. Library preparation includes DNA fractionation / fragmentation by nebulization or sonication, followed by enzymatic blunt ending and adapter ligation [11]. Adapters are key features for the amplification phase, which takes place on a solid support, the flow cell, and is performed by the Cluster Station, an automated device. The amplification is performed on the glass, oligonucleotide-covered surface of the flow cell, using solid-phase bridge PCR [12] [13]. The adaptors, which flank the DNA fragments, are bound to the oligonucleotide surface. A DNA polymerase creates copies of the template DNA fragment and the

immobilization ensures that all amplicons originating from the single molecule template are clustered together on the surface. Each cluster consists in hundreds of copies of the template. After cluster generation, the single strand amplicons are hybridized with sequencing primers at adapters flanking the DNA fragment [Figure 1]. At each cycle, a single base is incorporated with chemically modified nucleotides by using a modified DNA polymerase [10] [13]. After the four images have been captured at different channels, the sequencing cycle ends with a chemical cleavage of the fluorophore and the blocking group, enabling base incorporation at the next sequencing cycle. After n-cycles, which are setup by the user according to the purpose of the experiment, images of each cycles are analyzed and base calling is performed [10] [Figure 2]. This process allows transforming image-raw data to text-raw data files, which contains human readable reads of sequenced fragments. Hence, data are filtered in order to get rid of poor quality reads. The output of this data processing is a file in FASTQ format.

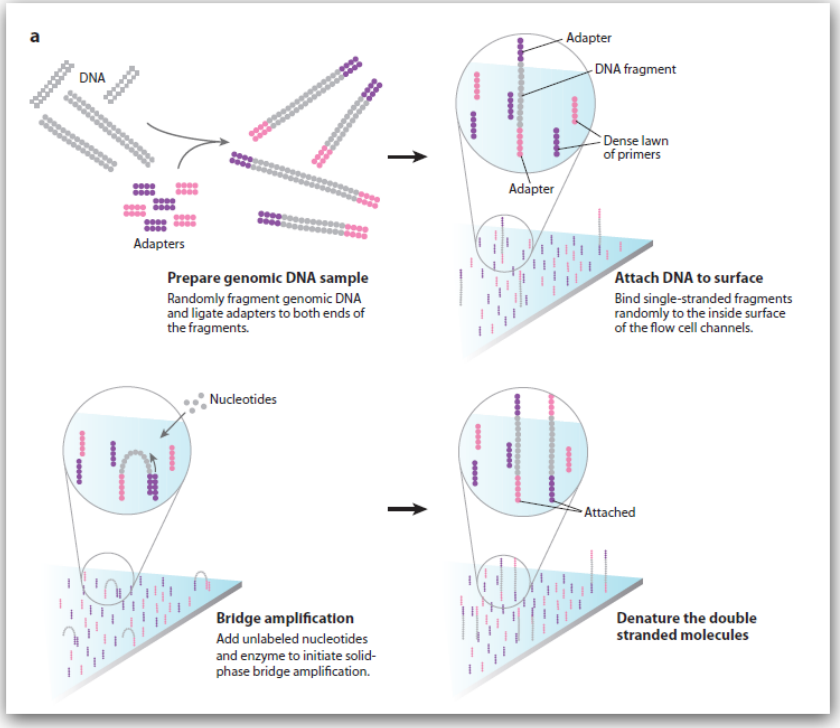



Figure 1. **Illumina sequencing (part I)**

*After DNA is randomly fragmented (by nebulization, sonication or by the use of ultrasounds) it is ligated to adapters to both the ends of each fragments. DNA with adapters is put on the flow cell surface covered by primers that bind the adapters sequences. Once DNA is attached as single strand fragments, unlabeled nucleotides and enzymes are added and solid phase bridge amplification can begin. This process creates clusters.*

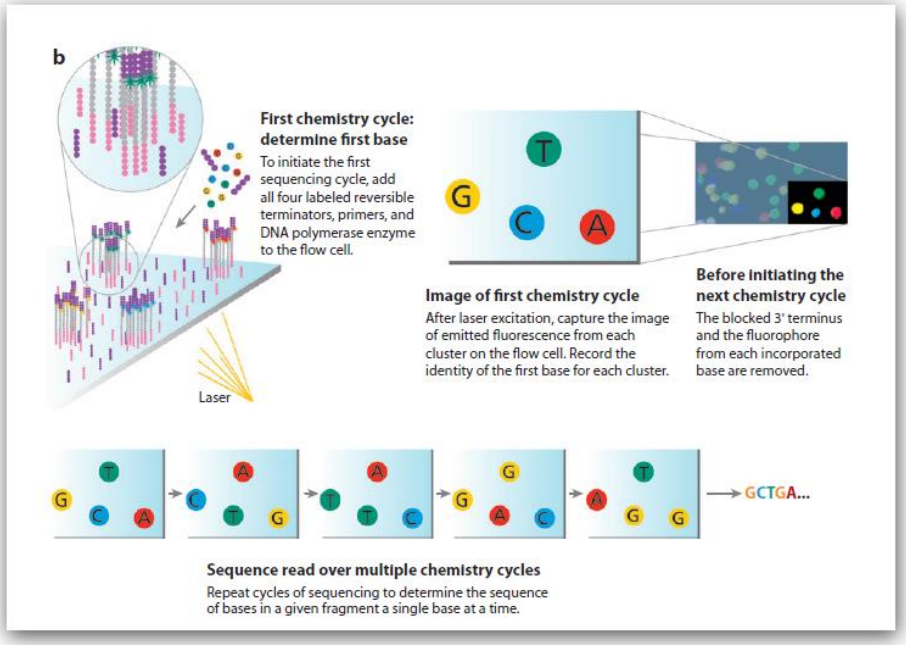



Figure 2. **Illumina sequencing (part II)**

*All four fluorescently labeled 3'-OH blocked nucleotides are added with DNA polymerase to the flow cell. Here the cluster strands are primed and extended by one nucleotide. Once the nucleotides are incorporated, the unused nucleotides with DNA polymerase are washed away and in the meanwhile a scan buffer is added to the flow cell. After the optic system scans each lane of the flow cell by imaging units called tiles. The fluorescent labels and the 3'-OH block group are then removed. The cluster strands are now ready for another cycle of fluorescent nucleotide incorporation. The number of cycles determines the length of reads produced.*

## *1.3 NGS approaches*

Originally, sequencing was a synonym for DNA sequencing. Nowadays NGS is routinely used to analyze DNA, RNA, proteins and to discover how these elements can interact each other's. There are mainly two big fields of sequencing: DNA and RNA sequencing. The first aims to study the genome and its elements, whereas the second to quantify mature transcripts, small RNA and how RNA-based molecules can interact with proteins and genome itself.

### *1.3.1 DNA Sequencing*

There are several approaches by which the genome can be studied. Nowadays the genome can be investigated in its whole length, or it is possible to analyze only coding-regions or examine specific regions of interest, by using whole genome, exome and target sequencing approach respectively. Actually, NGS platforms provide commercial solutions (reagents kits and software) to perform and analyze these experiments. They also provide custom design protocols directed to capture, enrich and sequence a specific regions of interest.

#### *1.3.1.1 Target Sequencing*

The rationale behind a target sequencing experiment is to study only the specific portion of the genome that is of specific interest. Targeted sequence enrichment refers to the set of technologies designed to isolate a specific genomic fraction (panels of genes or custom genomic regions) for subsequent NGS. It ultimately results in an enriched pool of target sequences such that there is an overall reduction in the genomic sequencing space, and hence, greater sequence coverage for each targeted region.

The reduction in sequencing space confers mainly two important advantages over sequencing the entire genome. First reducing the

sequencing space required per sample, multiplexing of samples becomes possible the overall costs per sample decrease as well and more samples can be sequenced at the same cost. Then, by targeting only the fraction of the genome that is either necessary or informative for the biological question being addressed, the complexity of the analysis is significantly reduced.

Several technologies exist for target enrichment, which can be classified by the enrichment method: hybridization based sequence capture, PCR-based amplification and molecular inversion probe-based amplification [14] [15]. Sequence capture permits the isolation of targeted loci from the background of the entire genome. The scale of the capture can range from several targeted loci to over a million target regions, making it suitable for both small-scale and large-scale projects. The design of a sequence capture experiments must be setup very carefully; indeed several key elements should be taken into account such as the mean depth of coverage for each target region, the probe specificity, the expected enrichment efficiency, the NGS platform in use and the genomic features of the target region [15] [16]. For instance, the mean coverage depth required for targeting a specific region relies upon the specificity of the probes used for that target and upon the presence of closely related sequences with which the target share some portions. Indeed, the genomic features of the target region should be analyzed very deeply in order to ensure the expected coverage both in terms of reads covering in depth and in percentage of target which can theoretically be captured. Two key features can negatively affect the mean coverage and the percentage of capturing. The first one is the GC content within the target region. It has been widely demonstrated that regions with a high GC content are more difficult to sequence [17]. The second is the presence of repeated regions. Repeated regions account for 56% of the genome [18], and are widely distributed

over the genome. Higher is the percentage of such regions in the target, higher will be the a-specific capturing. In order to bypass such hurdles there are several practice as increase the probes tailing in GC-rich regions and perform a cautious probe design which try to avoid the repeated sequences and in the meanwhile allows to maximize the theoretical percentage of capturing. Hence, is easy to understand how target sequencing can introduce a lot of bias depending on genomic region and on the experimental equipment (enrichment kits and NGS platform) used for the capture experiment.

Target sequencing is commonly used for identifying SNPs, INDELs, copy number variants and structural variants. Usually, considering the lower amount of DNA to sequence, target sequencing data are very accurate due to a deeper coverage achievable, even if discrete a-specific coverage is present through all the genome.

### *1.4 NGS applications on genomic data*

NGS is a high-throughput technology that produces a huge amount of data that must be handled, analyzed and interpreted. The progress of NGS technologies, has been followed by the development of computational methods and bioinformatics tools for analyzing data produced by NGS platforms. Nowadays plenty of different analyses can be performed on NGS data. The type of analysis depends on the approach and on the experimental design followed. Some analyses can be performed on multiple type of data, but it is recommended to perform specific-driven analyses suited for a specific biological issue. Data analysis is mainly focused on the identification of a particular feature/signature or profile of features, that are shared or that differ among a group of subjects/samples.

### *1.4.1 SNVs and INDELs*

The most common analysis in DNA sequencing field is the identification of single nucleotide variants (SNV) such as rare variants and single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs). This kind of analysis has driven the understanding of genomic variability until now. SNVs and INDELs can be analyzed starting from whole-genome data, exome data and target sequencing data. The identification of these genomic features is strictly dependent on the quality of the data and on the coverage of the genomic region to investigate. The depth of coverage, or simply the reads coverage, is a value which describes how many times a specific point / region of the genome has been sequenced and read. Higher is the coverage, more accurate is the SNVs and INDELs call. The clinical field is particularly interested in the identification of genomic variants which could have great impact on protein functionality or transcripts expression. Indeed the most common analysis is the SNVs and INDELs identification on exome data in which exons and their flanking regions are investigated. Several software are available to predict and evaluate how a specific SNVs or INDELs can impact on the protein functionality. Such information, along with gene and pathway annotation, gives the chance to rise hypothesis on disease mechanisms or causes. Up to now, a plethora of exome studies has been carried out, especially in the context of mendelian disease and family studies.

### *1.4.2 Structural variants*

Besides SNV and INDELs, there are other genomic features which are less studied due to the lack of bioinformatics tools and to the difficulties to assess and identify such variants. Among these features the are copy number variants (CNV) which spot duplication or loss events, inversions, translocations, both intra and inter chromosomal and tandem repeat regions. These genomic variants are called structural variants (SV)

because they can produce genomic rearrangements. SV are more difficult to detect and study than pointed-variants, like SNVs and small INDELs, due to the intrinsic complexity of the arrangement that they can promote. In fact, different classes of structural variant events can occur together in the same region and could be cell specific, making it difficult to study. These analysis can be theoretically performed both genome wide and in target regions (exome or custom target) even if specific requirements need to be fulfilled for ensure accuracy and reliability of results. For instance, CNV analysis should be performed at genome wide level and only with a high coverage in order to have a good accuracy.

In the next chapters, the attention will be focused on a specific class of SV: the translocations.

#### *1.4.2.1 Translocations*

Translocations are abnormal chromosomes regions that contain rearranged genetic material. In particular, a translocation can be intra or inter chromosomal depending on whether the transfer of genetic material occurs on the same chromosome or involves different chromosomes.

A translocation is called “Robertsonian” (ROB) when is characterized by the joining of long arms of two acrocentric chromosomes around a single centromeric region. The short arms also join to form a reciprocal product, which is usually lost in few cell divisions. Most of the times, ROB are non-deleterious because the essential genes are within long arms. When the ROB involve chromosomes 21 and 13 or 15, the heterozygous carrier is phenotypically normal because there are two copies of all major chromosome arms and hence two copies of all essential genes. However, the progeny of this carrier may inherit an unbalanced trisomy 21 or 13, causing Down syndrome and Patau syndrome respectively [19].

A generic translocation, when reciprocal can be unbalanced or balanced depending on whether or not it affects the copy number of any portion of

the genome. Balanced reciprocal translocations arise from the fusion of two double-strand breaks [20] [21]. Depending on the type of DNA break and on where it occurs, different repair mechanisms are active [21]. Many pathological translocations fall within the category of balanced reciprocal translocations among non-homologous chromosome. Among them, there are the Philadelphia chromosome (Ph) t(9,22) that causes the overexpression of ABL1 kinase in CML, the translocation t(8,14), found in 85% of Burkitt's lymphoma cases, that leads to MYC transcription factor overexpression, and the translocation t(11,22) that occurs in Ewing's sarcoma that brings to deregulation of the ETS transcription factor FLI1 [22]. Translocations that generate gene fusions arise from the joining of double strand DNA (dsDNA) breaks that occur at different sites on non-homologous chromosomes. In such case, dsDNA breaks are frequently joined by an endogenous DNA repair pathway called "non-homologous end joining" (NHEJ) [23]. In particular, there are two kind of NHEJ: classical and alternative. The classical pathway (C-NHEJ) consists in the binding of Ku70-80 heterodimer to broken DNA ends, then the recruitment of a complex of DNA protein kinase catalytic subunit and Artemis which binds to Ku70-Ku80-DNA complex and processes the DNA end through the nuclease activity of Artemis, and finally a complex with DNA ligases properties joins the DNA end [24]. The alternative pathway (A-EJ) also called microhomology-mediated end joining (MMEJ) is activate when the C-NHEJ is deficient [25] and occurs when in the regions flanking the DNA breaks there are homolog sequences. Several studies suggests that A-EJ, if active, produces an increased number of chromosome rearrangements compare to the C-NHEJ that causes a lower rate of translocation events.

#### *1.4.2.2 Methods for identifying structural variants*

SV identification techniques have made a step up the ladder with the advent of NGS technologies, which allow reaching an accuracy level that before was out-of-the-way. Indeed, before the introduction of NGS techniques, array-CGH was the gold-standard technique for the identification of SVs.

Array-CGH are cheap technologies, but are limited to the detection of “big size” SVs with a low accuracy in terms of genomic boundaries. Differently, NGS methods have been able to improve the accuracy, reaching the single base resolution of SV events. During the last 5 years, several bioinformatics tools for SV detection starting from NGS data have been developed [26]. Such methods belong to three different approaches [27] for SV identification: discordant pair approach [28], split-reads approach [29] and depth of coverage approach [30]. The discordant pair approach uses paired-end reads, which have a discordant insert size. The discordance can be explained as deviation from the expected insert size value, which can vary according to the target sequencing experimental design. BreakDancer [31] and SVDetect [32] are some of the software that use this approach. More details on the algorithm implemented in SVDetect tools will be shown in software section. The split-read approach is based on split-reads, which are reads that do not map correctly to the reference genome or that are unmapped. Such reads are very informative because they have a boundary portion that maps to the reference genome, whereas the flanking region, which is unmapped, spots the SV.

There are several algorithms for SV detection starting from partially mapped paired-end reads, as Pindel [33], SLOPE [34] and ClipCrop [35]. The first and second tools use orphaned reads, unmapped reads whose mate maps to reference genome, as putative reads for SV identification. The latter, ClipCrop, takes advantage from soft-clipping information

included in SAM format mapping file produced by Burrows-Wheeler alignment tools like BWA [36] . More details on the algorithm implemented in ClipCrop will be shown in software section.

The depth of coverage approach, uses the frequency of mapped reads or bases to each position of the reference genome (or a custom genomic interval) to call the SV. The rationale under this methods is pretty similar to the one on which array-CGH are build. Indeed, in case of duplication the number of mapped reads/bases to regions in the reference genome will increase, whereas in case of deletions it will decrease.

Every approach has its pros and cons, and specific applications are more suitable for some approaches compare to others. For instance depth of coverage approach is suited for the identification of deletions and insertions but not for inversions and translocations identification. Discordant pair and split-reads approaches performance very well in the identification of all types of SV, although the latter approach is more accurate (single nucleotide) than the first one [27].

### *1.5 Chronic Myeloid Leukemia*

Chronic myeloid leukemia (CML) is a clonal myeloproliferative disorder arising from an acquired alteration in hematopoietic stem cells (HSC), the Philadelphia chromosome (Ph) [37] [38].

In 95% of patients, the reciprocal balanced translocation between the long arms of chromosomes 9 and 22, t(9;22)(q34;q11) results in a shortened chromosome 22, called Ph, and an elongated chromosome 9, called derivative chromosome (der 9) [Figure 3]. This chromosomal rearrangement includes the juxtaposition of the Abelson gene (ABL1), on chromosome 9, with the Breakpoint Cluster Region (BCR) gene located on chromosome 22, leading to the formation of 3’ BCR-ABL 5’ fusion gene. This hybrid gene encodes the Bcr-Abl oncoprotein, a constitutively activated tyrosine kinase that plays a crucial role in CML pathogenesis.

CML accounts for 15% of all leukemia forms in adults [38] and for 0.2% of all cancers [22].

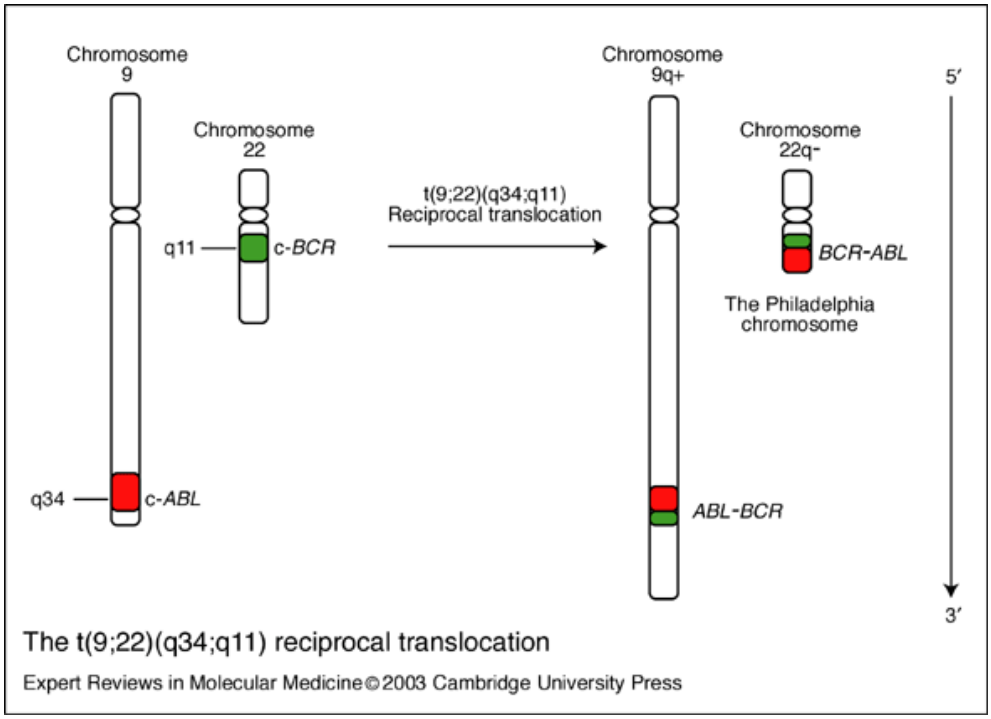



Figure 3. **t(9;22)(q34;q11) reciprocal translocation**
*The reciprocal translocation t(9;22)(q34;q11) occurs between the long arms of chromosome 9 and chromosome 22, producing a shortened (22q-) chromosome 22 called Philadelphia chromosome(Ph) and a longer (q9+) chromosome 9 named derivative 9 (der9). During this translocation, BCR and ABL1 genes are broken and two fusion products are formed: 5'→3' BCR-ABL1 on Ph and 3'→5' ABL1-BCR on der9.*

### *1.5.1 Epidemiology, etiology and risk factors*

The incidence of CML is 1-2 cases per 100.000 population. The median age of patients is 65 years and the median age of onset is about 45-55 years [38] according to SEER [39] .

The cause of CML is unclear. Some associations with genetic and environmental factors have been reported, but in most cases no causative

factors can be identified. There is little evidence linking genetic factors to CML. Offspring of parents with CML do not have a higher incidence of CML than the general population [40].

### *1.5.2 Course and symptoms*

CML arises in three phases: chronic, accelerated and blastic. CML in chronic phase (CP) is asymptomatic in 25% to 60% of cases disease is unmasked on routinely blood exams.

In symptomatic patients, typical symptoms are fatigue, weight loss and palpable splenomegaly [41] [42] in 30% to 70% of patients [43]. If untreated, CML in CP is associated with a median survival of 4 to 5 years [37]. The accelerated phase (AP) is a brief transitional phase, usually symptomatic. If treated with common therapy, the estimated 4-years survival rate exceeds 50% and the medial survival is 1-2 years. During the blastic phase (BP), cells fail to mature and resemble blasts found in acute leukemia (AML and ALL).

Patients in blastic phase show severe clinical conditions and have a poor prognosis, dying within 3 to 6 months [43].

### *1.5.3 Laboratory features*

In laboratory tests, CML exhibits an elevated white blood cells (WBC) count, frequently exceeding $1x10^{11}$/L [42]. The presence of leukocytosis (>1.2 -$1.5x10^{10}$/L) in the absence of infections should prompt a workup for CML. In particular, the absolute lymphocyte count is elevated at the expense of T lymphocytes. Another common laboratory feature of CML is the high platelet count at chronic phase (30% to 50% of patients) that drops (thrombocytopenia) when the accelerated phase occurs leading to anemia. The bone marrow is usually hypercellular [44] [42], with a cellularity of 75% to 90% [43]

### *1.5.4 Diagnosis*

The diagnosis is based on the detection of some hallmarks such as leukocytosis with basophilia and with immature granulocytes as myelocytes, premyelocytes and occasional myeloblasts, thrombocytosis and physical findings like splenomegaly. Blood counts and differential are fundamental for the disease phase evaluation. The diagnosis achieved by blood analysis must be confirmed by cytogenetic analysis showing the translocation (9;22)(q34;q11) and by reverse transcriptase polymerase chain reaction (RT-PCR) that detects BCR-ABL1 transcripts. The cytogenetic diagnosis is performed by chromosome banding analysis (CBA) of bone marrow cells at metaphase. If bone marrow is not available, karyotype analysis is replaced by the fluorescent in situ hybridization of blood cells at interphase (I-FISH) [45]. Using this technique it is possible to detect BCR-ABL$^+$ nuclei and Ph$^+$ variants [46]. RT-PCR is performed on RNA from blood or bone marrow and allows to detect the transcript type indicating the breakpoint region in BCR (major, minor, micro) and hence the BCR-ABL1 protein weight (p210, p230 and p190).

### *1.5.5 Staging and prognostic risk systems*

The determination of the stage of CML is the starting point to establish a correct therapy. A good prognosis leads to a better outcome and survival.
Staging of CML has been changed during the last 20 years, especially as long as the treatment with tyrosine kinase inhibitors (Imatinib) has been introduced as standard therapy. The staging system is based on several parameters by which it is possible to classify the disease phase in patient. The World Health Organization (WHO) in 2001 introduced standards for the evaluation and classification of CML phase [47]. 95% of patients are diagnosed in CP. The disease can develop to an accelerated phase (AP) and then to a blastic phase (BP). According to WHO criteria, AP is characterized by blasts (10-19%), basophils (>20%), platelets (<100x10$^9$/L

unrelated to therapy), whereas BP by a concentration of blasts >20% and the presence of extramedullary blasts or large clusters of blasts in bone marrow. In 2006, Cortes et al. [48] demonstrated that WHO standards in the CML staging need to be updated and proposed, taking into consideration previous treatments, new criteria for defining CML stages, improving the survival rate and the prognosis.

Although studies have demonstrated improvements in prognosis and staging, nowadays the advance in genomics analysis and treatments has introduced many potential parameters on which to develop new standards for the evaluation and staging of the disease. Nevertheless in clinics, persists some prognostic risk systems that were introduced in the pre-imatinib era, as the Sokal score classification [49] (1984) and the EURO [50] prognostic system (1998) created for taking into account IFNα-treated patients. These prognostic systems take into consideration several characteristics including age, spleen size, WBC and platelet count, and percentage of blasts, eosinophils and basophils in the peripheral blood. The Sokal score, the first system developed, consists in the evaluation of the hazard ratio (HR) by using the following formula:

$$HR = e^{(0.0116^* (age - 43.4) + 0.0345 (spleen\ size * 7.51) + 0.188 ((platelet / 700 \times 109/l)^* 2 - 0.563) + 0.0877 (blast\% - 2.1)}$$

This score defines three prognostic groups: low risk (<0.8), intermediate risk (0.8-1.2) and high risk (>1.2) [49]. Sokal score can be used as long as the patient is out of treatment. More recently, EUTOS risk score [51], a new prognostic system that takes into consideration imatinib-treated patients, has been introduced. EUTOS systems is partially detached from blood counts and other factors like age and spleen size [51].

### *1.5.6 Treatment and therapy*

CML was discovered in 60's and from the very beginning many efforts have been made to improve the clinical conditions of the patients.

The treatment of CML was historically based on busulfan, and on hydroxyurea, which is still used in for short pre-treatment phase in case of thrombocytosis and leukocytosis. In the 80-90's interferon-α (IFN-α) was the gold standard.At the end of 90's Imatinib, the first tyrosine-kinase inhibitor (TKI), was introduced [52] [53]. The advent of such drug has revolutionized the CML therapy. Today Imatinib is considered the gold standard for the first-line treatment of the disease. More recently, the second-generation of TKIs has been introduced in the treatment that include drugs like Dasatinib and Nilotinib which are currently used in second-line treatment in case of Imatinib failure or sub-optimal response [54] [55]. The choice of treatment is strictly dependent on the degree of cytogenetic response and on the detection of BCR-ABL mutations, which can negatively affect the drug action [56]. Nowadays there are three types of drugs that can inhibit the kinase activity with three different mechanism: stabilizing the inactive state (Imatinib, Nilotinib, Ponatinib), altering the active state (Dasatinib, Bosutinib) or by allosteric inhibition through the binding with regulatory domains (GNF-2, GNF-5) [57]. In particular, Imatinib acts by competing with ATP for the binding site on the kinase domain.

More recently, several new TKIs like Ponatinib and Bosutinib has been introduced in trials [54]. These drugs are specifically designed to maintain efficacy in the presence of specific point mutations, which can affect the functionality of first and second line TKIs [58]. The last line of treatment is based on allogenic hematopoietic stem cells transplantation (alloHSCT) which is carried out if both first and second line treatment fail or prior to the onset of blastic phase. The patients who achieve the blastic phase have to this day have an unfavorable prognosis.

### *1.5.7 Disease monitoring*

#### *1.5.7.1 Cytogenetic and RNA-based approaches*

The monitoring of CML can be cytogenetic or molecular. The cytogenetic monitoring is based on CBA and I-FISH. The cytogenetic response achieved by CBA analysis can be complete, partial, minor, minimal or none, depending on the percentage of $Ph^+$ metaphases detected. In case of bone marrow lack, cytogenetic response is evaluated using I-FISH. In this case the cytogenetic response can be defined complete when the percentage of BCR-ABL1$^+$ nuclei is less than 1% out of at least 200 nuclei at interphase [59]. The molecular monitoring is based on the quantification of BCR-ABL1 mRNA by using qRT-PCR. This method is nowadays considered the most sensitive tool for disease monitoring and it is particularly useful in minimal residual disease (MRD) monitoring. mRNA quantification is expressed as the ratio of BCR-ABL1 transcripts to ABL transcripts, that is expressed as % of BCR-ABL1 on a log scale according to the International Scale in order to guarantee comparability of results among different laboratories. According to this scale, BCR-ABL1 % equal to 10, 1, 0.1, 0.01, 0.0032 and 0.001 corresponds to a decrease of 2;3;4;4.5;5 logs below the standard baseline [59]. BCR-ABL1 expression below 3 logs corresponds to a Major Molecular Response (MMR). The molecular testing should be performed every 3 months until MMR is achieved, followed by controls every 3 - 6 months. If transcript levels increased 5 times or more in a single follow-up and MMR was lost, the monitoring should be repeated in a shorter time interval. They are early indicators of treatment failure. Differently if a very deep molecular response ($M^{4.0}$ – $M^5$) is achieved during TKI treatment, there are the prerequisite for therapy discontinuation within controlled trials [59] [45].

#### *1.5.7.2 MRD DNA-based monitoring*

MRD is currently assessed by the quantification of BCR-ABL1 mRNA through qRT-PCR. This method is the gold standard to monitor the disease and make the appropriate decisions on disease treatment. However, mRNA quantification although very sensitive, is not a direct observation of the number of leukemic cells and cannot deal with potentially transcriptionally silent CML cells. When the qRT-PCR shows the absence of BCR-ABL1 transcripts, two assumptions can be made: first, CML cells are not present in the sample or CML cells are present but are transcriptionally inactive. For this reason, a new sensitive method for MRD monitoring based on the quantification of BCR-ABL1 DNA using qPCR has been proposed [60]. This new method was applied by Mattarucchi and Porta [60] to 10 CML patients undergoing Imatinib treatment. The Authors compared BCR-ABL1 mRNA levels with the actual proportion of leukemic cells through quantitative amplification of DNA BCR-ABL1 fusion segments. They observed similar results with the two approaches, with most patients being positive to both mRNA and DNA after two years treatment. However in one of the two patients in which mRNA was undetectable, DNA was still present after 42 months. In this patient Imatinib should not therefore be discontinued. On the contrary, in the other patient both DNA and mRNA were negative suggesting this patient as a possible candidate for Imatinib withdrawl. The idea to identify quiescent leukemic cells carrying BCR-ABL oncogene but transcriptionally silent was not new. Indeed Kim et al [61] as well as Zhang et al [62] had previously developed a protocol for the quantification of leukemic cells by PCR techniques. Other studies supporting the better performance of DNA based approach to monitor MRD have also been performed by Bartley et al [63].

More recently, Porta (unpublished data) has demonstrated that DNA-based monitoring can spot the presence of BCR-ABL+ cells in ~ 30% of time points in which monitoring with qRT-PCR has a negative result [Figure 4].

The accuracy reached with DNA-based monitoring allowed to detect MRD better than with RNA qRT-PCR. The same Author could demonstrate that in some cases, when q-PCR assay is positive to CML cells and qRT-PCR is not, the minimal residue is due to the presence of quiescent stems cells (CD34+) that can turn to be active and promote the relapse. Figure 5 shows how the DNA-based disease monitoring is able to spot the presence of CML cells when cytogenetic and RNA-based approaches fail. However, the limit of this method was that breakpoints were resolved with laborious long-range PCR not suitable for a routine application.

The data reported above have been the basis for a collaboration between our research group and Prof.Giovanni Porta from Insubria University, for a further improvement of the DNA based method, to enhance the management of patients affected by CML. This has been the object and the aim of the present PhD project.

| | Patients | Number of Samples (N) | mRNA + (%, n/N) | mRNA - (%, n/N) | DNA + (%, n/N) | DNA - (%, n/N) | DNA+/mRNA - (%, n/N) |
|---|---|---|---|---|---|---|---|
| | 1 | 15 | 40 (6/15) | 60 (9/15) | 86.7 (13/15) | 13.3 (2/15) | **46.7 (7/15)** |
| | 2 | 7 | 71.4 (5/7) | 28.6 (2/7) | 100 (7/7) | 0 (0/7) | **28.6 (2/7)** |
| | 3 | 15 | 53.3 (8/15) | 46.7 (7/15) | 93.3 (14/15) | 6.7 (1/15) | **40 (6/15)** |
| | 4 | 17 | 58.8 (10/17) | 35.3 (6/17)* | 100 (17/17) | 0 (0/17) | **41.2 (7/17)** |
| | 6 | 26 | 100 (26/26) | 0 (0/26) | 100 (26/26) | 0 (0/26) | **0 (0/26)** |
| | 8 | 19 | 47.4 (9/19) | 52.6 (10/19) | 73.7 (14/19) | 26.3 (5/19) | **26.3 (5/19)** |
| | 9 | 17 | 47.1 (8/17) | 52.9 (9/17) | 94.1 (16/17) | 5.9 (1/17) | **47.1 (8/17)** |
| | 10 | 12 | 41.7 (5/12) | 58.3 (7/12) | 100 (12/12) | 0 (0/12) | **58.3 (7/12)** |
| Tot | 8 | 128 | **60.2 (77/128)** | **39.1 (50/128)** | **93 (119/128)** | **7 (9/128)** | **32.8 (42/128)** |

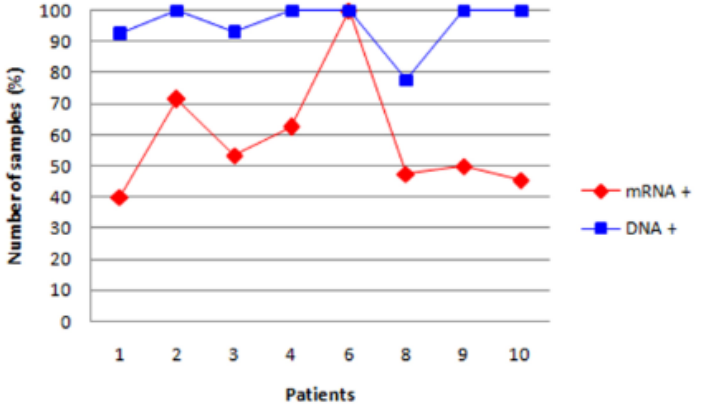



Figure 4. **DNA vs mRNA MRD monitoring comparison**

*On the top, MRD monitoring with DNA is able to identify leukemic cells in ~33 % of the total number of time points in a follow up study on 8 samples treated with Imatinib (IM) (Porta et al [unpublished]). In patient 6 no differences were found between DNA based monitoring and mRNA based monitoring due to unresponsive to IM treatment.*
*On the bottom, the % of DNA and mRNA detection across samples.*

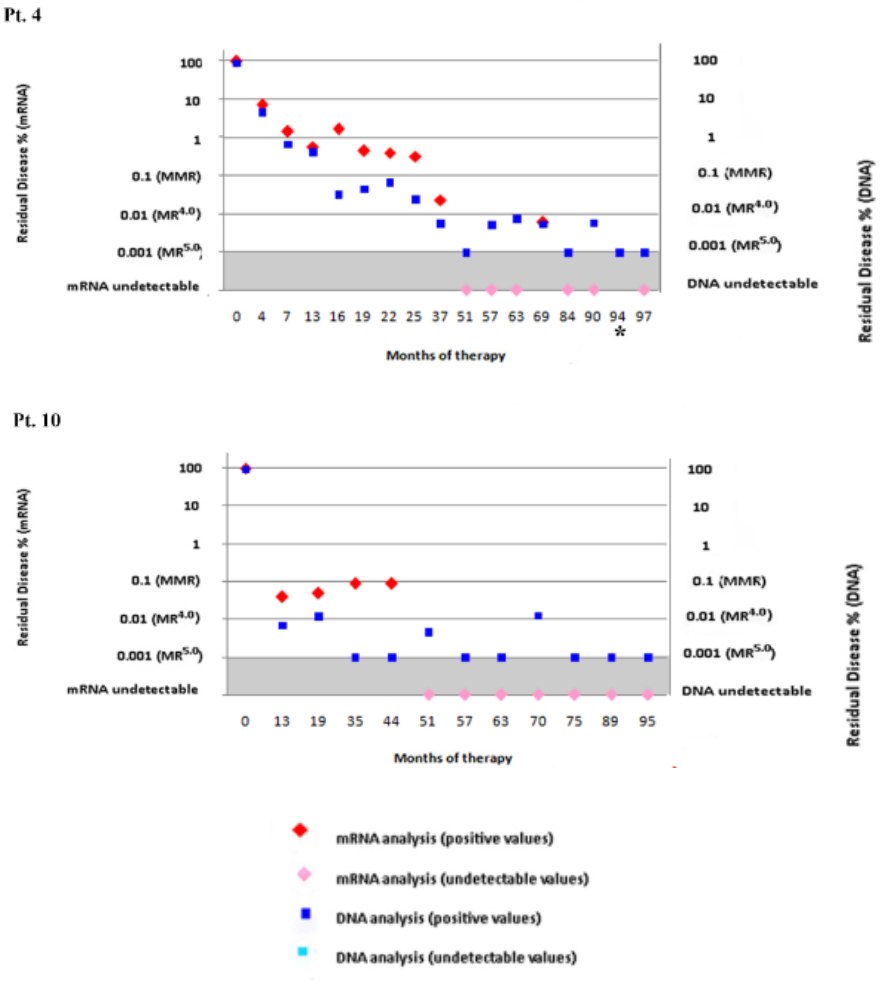



Figure 5. **MRD qPCR vs qRT-PCR**

*MDR Monitoring in sample 4 and 10. In sample 4 (Pt.4) from 40$^{th}$ month to 97$^{th}$ month, DNA is always detected, whereas mRNA is always undetectable except on 69$^{th}$ month. In patient 10 DNA analysis of MRD is always positive, whereas mRNA is always negative.*

## *1.6 BCR, ABL1 and BCR-ABL1*

The hallmark of CML is a t(9;22)(q34;q11) that gives rise to the formation of Ph in 95% of patients [37]. At molecular level, t(9,22) is characterized by the fusion of the proto-oncogene ABL1, located on 9q34.1 and the breakpoint cluster region BCR at 22q11.2 leading to the formation of a fusion gene BCR-ABL1 which is translated into an oncogenic fusion protein [37] [64]. In the next paragraphs BCR, ABL1 and BCR-ABL1 fusion gene and related proteins will be described in terms of sequence, structure and functions.

### *1.6.1 BCR*

#### *1.6.1.1 Gene and Transcripts*

BCR is located on the long arm of chromosome 22 in the 11.23 cytogenetic band and is the acronym for *Breakpoint Cluster Region*, due to the presence of clusters of breakpoints in which the chromosome 22 can be broken during the BCR-ABL1 gene fusion. BCR is located on chromosome 22 from coordinate 23,522,552 to 23,660,224 (137,673bp) on the reference genome (Hg19) [65] [66]. It is transcribed in 16 different transcripts (splice variants) whose only two encode for protein products [65]. The main transcript is 7082bp long and includes 23 exons encoding a protein of 1271aa, whereas the alternative transcripts is 4708bp long includes 22 exons and encodes an isoform of 1227aa according to Ensemble and UniProtKB [65]. The region spanned by BCR is GC rich (52.6763 ± 21.931 %) [67] and presents a percentage of repeated sequences of 32.47% according to Repeat Masker [18]. The BCR gene is expressed ubiquitously [68] but highest mRNA levels are found in brain and hematopoietic cells [69].

The expression of the gene occurs mainly in the early stages of myeloid differentiation and decrease significantly as cells mature to polymorphonuclear leukocytes [69].

#### *1.6.1.2 Protein*

Bcr protein has a very complex organization including several functional domains and motifs:

- Bcr-Abl1 oligomerization domain at N-terminus essential for the oncogenic activity of the Bcr-Abl1 tirosine kinase;
- Serine-threonine kinase domain located in the first exon, by which the protein can phosphorylate itself or several substrates like casein and histones. This domain includes several Src homology-2 (SH2) binding domains, which are necessary for the binding with Abl1 [64];
- Rho-GEF domain, with Rho guanyl-nucleotide exchange factor activity;
- Rho-GAP domain for RAC1 and CDC42 GTPase activation;
- Pleckstrin homology domain [57] involved in the intracellular signaling;
- C2 structural domain involved in targeting proteins at cell membrane.

Bcr protein comes up as a homotetramer of 148 -160 KDa and resides in both cytoplasmic and nuclear compartments [64] [37] [64]. In normal cells, BCR is a GTPase-activating protein for RAC1 and CDC42, promoting the exchange of RAC or CDC42-bound GDP by GTP, thereby activating them [70]. BCR is involved in the regulation of cell cycle, in the signal transduction; it acts as a positive regulator of phagocytosis and negative regulator of inflammatory response. Moreover, it plays a role in the organization and biogenesis of actin cytoskeleton [68] [37] as well as brain

development. Indeed, sequence variants of the Bcr protein may be associated with bipolar disorder. However, the function of wild type Bcr in cells remains unclear.

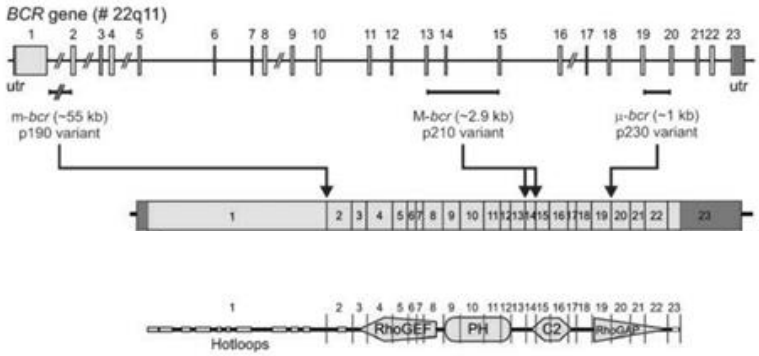



Figure 6. **BCR annotation**

*The figure represents the BCR gene and its products (mRNA and protein) in terms of sequence annotation. BCR gene is characterized by 23 exons. The clusters of breakpoints are located mainly in three regions (μ-BCR, M-bcr and m-BCR). BCR is transcribed in two different isoforms differing for the presence of a 23rd exon in the longer form. BCR protein has a complex structure including oligomerization domain, a kinase domain with SH binding domains, a RHO-GEF domain with guanyl-nucleotide exchange factor activity, the pleckstrin domain (PH) that mediates intracellular signaling and C2 domain for targeting proteins at cell membrane.*

### *1.6.2 ABL 1*

#### *1.6.2.1 Gene and Transcripts*

The cellular ABL1 gene is the human homologue of the viral ABL (v-ABL) oncogene carried by the Abelson murine leukemia virus. Viral ABL1 originates from cellular ABL1 (c-ABL), probably incorporating the mammalian ABL1 gene at some point during the evolution.

ABL1 is a gene located on chromosome 9 at the following genomic coordinates (hg19): 133,589,333 - 133,763,062 (171,74Kb) [66]. Abl1 contains two alternative 5' exons spliced to a common set of 3' exons to yield the two major ABL1 RNA transcripts. These transcripts initiate in

different promoter regions and give rise to proteins that vary in their N-terminus. The canonical transcript according to [71] is 5766bp in length and it encodes an 1130aa protein, whereas the secondary transcript is 3824bp long and encodes for a protein of 1149 aa. The secondary transcript has a longer first intron compared to the canonical transcript that has a short first intron. The remnant sequence (from exon 2 to exon 11) is the same for both transcripts. The region spanned by ABL1 has a standard GC content (44.1186 ± 21.931 %) but presents a high average percentage of repeated sequences (54.29%) according to repeat masker [18] [67]. In malignant hemopathies, ABL1 gene is mainly fused with BCR, but further fusion partners (ETV6, RCSD1, SFPQ, NUP214, EML1) has been discovered [72]. These fusions partners have a coiled-coil or helix-loop-helix domain that promote the oligomerization of ABL1 and enhancement of its kinase activity [72].

#### *1.6.2.2 Protein*

ABL1 has two isoforms (Abl1a, Abl1b) that differ only for 19 amino acids, due to a tiny difference between the two 1$^{st}$ alternative exons.

| Exon | Amino acid Sequence |
|---|---|
| **1a** | MLEICLKLVGCKSKKGLSSSSSCYLE |
| **1b** | M**G**QQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEH |

Table 1. **Alternative 1st exon in ABL1**
*ABL1 has two alternative 1st exons that differ only for 19 amino acids that encodes for a myristoylable residue (glycine) at position 2 in the 1b exon. Such residue is involved in the auto-inhibition process of ABL1.*

The sequence annotation is the same for both isoforms, for less than a myristoylable glycine at position 2 located in the exon 1b.

The Abl1b protein consists of a myristoylated (Myr) N-terminal region, followed by SH3 and SH2 domains, the SH2/kinase linker, the tyrosine kinase domain, and a long last exon region with a C-terminal F-acting

binding domain (F-ABD) [57] [73] [74]. In the carboxy-terminal region, besides that ABD, ABL1 contains proline-rich (PXXP) SH3 binding sites, three nuclear localization signals (NLS), one nuclear exporting signal (NES) and a DNA binding domain (DNA BD). The tyrosine kinase domain along with SH3 and SH2 domains are assembled in order to maintain an auto-inhibitory structure in which SH3 and SH2 domains function as a 'clamp' that block the tyrosine kinase in the inactive state [75]. The crystal structure of the Abl1 core reveals that the myristic acid group binds a hydrophobic pocket in the C-terminal lobe of the tyrosine kinase [76]. This binding promotes the stability of the tyrosine kinase domain by keeping SH3-SH2 grip in place. Indeed mutations in the myristoyl signal sequence [77] and mutations that affect the interaction of SH2 domains to the C-lobe of the protein results in the up regulation of the kinase activity. Moreover, deletions or mutations of the SH3 domain, as well as mutations in the SH3-SH2 linker can promote up-regulation of the kinase activity [76] [46]. The regulation of tyrosine kinase activity is exerted not only by a complex network of intermolecular interactions but also by all the elements able to stabilize or destabilize the auto inhibitory conformation. For instance phosphatidylinositol-4,5-bisphosphate (PIP2), peroxiredoxin 1 (PRDX1), the tumour suppressor protein FUS1 can negatively regulate the ABL kinase activity [73], whereas Ras and Rab interactor 1 (RIN1), which interacts with both the ABL SH3 and the SH2 domains is able to up-regulate the activity [76].

Moreover, the phosphorylation of specific residues (Y245, Y412) has been demonstrated to stabilize the active-state of the protein kinase [73].

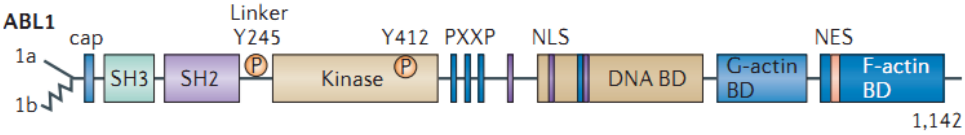



Figure 7. **Protein sequence annotation**

*ABL1 protein presents several domains and motifs. The Src homology domain 2,3 (SH2, SH3), an N-CAP that in Abl1b contains the myristoyl group, the SH2-SH3 linker, the tyrosine kinase domain (SH1) and the Y412 residue located in the activated loop autophosphorilation site. Proline rich elements (PXXXP) by which ABL1 can interact with other molecules, the NLS and NES nuclear signals for exerting inter and extra-cellular signaling respectively. In the C-terminal the G and F-actin domain necessary for cytoskeleton interactions. Figure adapted from* [57].

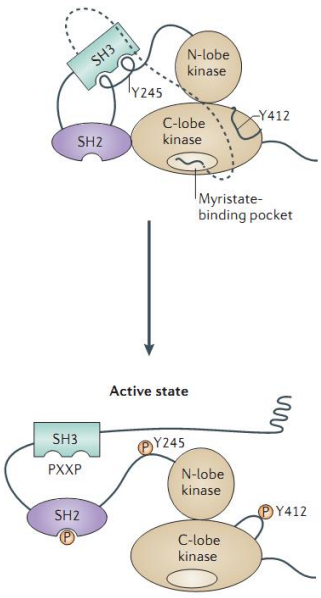



Figure 8. **ABL1 tirosine kinase activity regulation**

*The activity of ABL1 tirosine kinase is strictly dependent on the regulation of the stability of a complex network of interactions among the ABL1 functional domains. ABL1 is physiologically auto inhibited The SH3 domain binds to the linker sequence connecting the SH2 and the kinase domains, and the SH2 domain interacts with C-terminal lobe of the kinase (SH1) domain forming an SH3–SH2 clamp structure. The stability of this complex is enhanced by the binding of myristoylated residue at the N-CAP (1b isoform) with the C-lobe kinase. When these molecular interactions are broken or simply affected, the autoinhibited state is loss and the tyrosine kinase become active. Figure adapted from (64).*

ABL1 is a ubiquitously expressed and highly conserved proto-oncogenic tyrosine kinase. The protein is present both in the nucleus and in the cytoplasm of cells. It has been implicated in regulation of cell proliferation, differentiation, apoptosis, cell adhesion and stress response [37] [57]. The function of ABL1 and the pathways activated by it are different depending on the tissue in which is expressed. More insights on the ABL1 activated pathways will be described in the next chapters.

### *1.6.3 BCR-ABL 1*

#### *1.6.3.1 Mechanisms of BCR-ABL1 oncogene origin*

Chronic myeloid leukemia is characterized by the expression of the BCR-ABL1 oncogene. In 90-95% of patients, the fusion gene arises from t(9;22) that leads to Ph formation. In the remnant 5%, the fusion gene can arise from Ph variants where the Ph is further rearranged with other chromosomes (Masked Ph), or from cryptic insertions of ABL1 region into BCR gene (or vice versa) [46]. The occurrence of these complex rearrangements is very difficult to explain temporary, indeed they could occur by serial translocations events or by a single complex simultaneous event. In a cytogenetic study with FISH conducted on 450 CML patients [46], about 9.5% of patients present chromosome rearrangements involving at least one (90.7%) or more (9.3%) chromosomes, in addition to the chromosomes 9 and 22. Among these, the 83% show classic Ph with masked der(9), whereas the remnant 17% a classic (der9) with a masked Ph. Cases with masked der(9) showed additional genetic material coming from different chromosomes with prevalence of 4,6 and 12. Only a small percentage (11%) of such cases the 5'ABL1-BCR 3'gene is retained on der(9). In the majority of cases the 5'ABL1/3'BCR is lost for deletions occurring at der(9) or is found in others chromosomes. In cases with "masked Ph" the 5' BCR-ABL 3' fusion gene has been found on der(9) or other chromosomes. In 42% of cases with multiple chromosomes

rearrangements, micro deletions events at breakpoints are detected [46]. This study shows how complex are the rearrangements to which the cell can undergo and how difficult can be the identification and understanding of them.

#### *1.6.3.2 Fusion gene and transcripts*

The fusion of BCR and ABL1 occurs in a head-to-tail way, with the 5' end of ABL1 joined to the 3' of BCR. The fusion gene can be translated in three types of proteins depending on the different breakpoint cluster region involved in the translocation.

In BCR gene, there are three clusters regions in which the translocation can take place: major (M-BCR) [78], minor (m-BCR) [79] and micro (µ-BCR) [80]. M-BCR region is an area of 5 spanning from exon 12 to exon 16, m-BCR region range from the 1$^{st}$ to the 2$^{nd}$ exon and the µ-BCR region spanning from exon 19 to exon 20 [37] [64]. In ABL1 gene, the breakpoint can occurs over an area ranging from 3' end of upstream gene EXOSC2 to the exon 2 of ABL1, but is mainly located between two alternative 1$^{st}$ exons in ABL1 [81]. Hence, BCR-ABL fusion gene can give rise to three proteins: p190, p210 or p230 according as the breakpoint in BCR is located in m-BCR, M-BCR and µ-BCR respectively. Taking into account that ABL1 has two alternative transcripts and that the BCR breakpoint in M-BCR can occur over five exons (b1-b5), different fusion transcripts can be produced. In particular, regarding the M-BCR, breakpoints mostly take place in the region from exon 13 to exon 15 and the resulting fusion transcripts (b2a2 and b3a2) give rise to two 210-kDa tyrosine kinase proteins (p210 $^{BCR-ABL}$).

ABL1 breakpoints do not affect protein weight because the impact of ABL1 alternative transcripts (ABL1a,b) on the BCR-ABL protein length is 19 amino acids only. The majority of patients (95%) exhibit mRNA transcripts with a b3a2 or a b2a2 junction and in 5-10% of cases both transcripts are present because of alternative splicing [82] [83] [84] [85].

It has been widely demonstrated that the frequency of b2a2 and b3a2 transcripts depends from the population under analysis. Generally, European and Eastern population have higher frequency of b3a2 transcript, whereas in Western countries and Africa populations, the frequency of b2a2 is higher [82].

The co-expression of both transcripts can be explained as the result of alternative splicing rather than by the presence of two different clones from which the CML can spread. Very few information are available concerning hybrid transcripts that arise from µ-BCR and m-BCR [82].

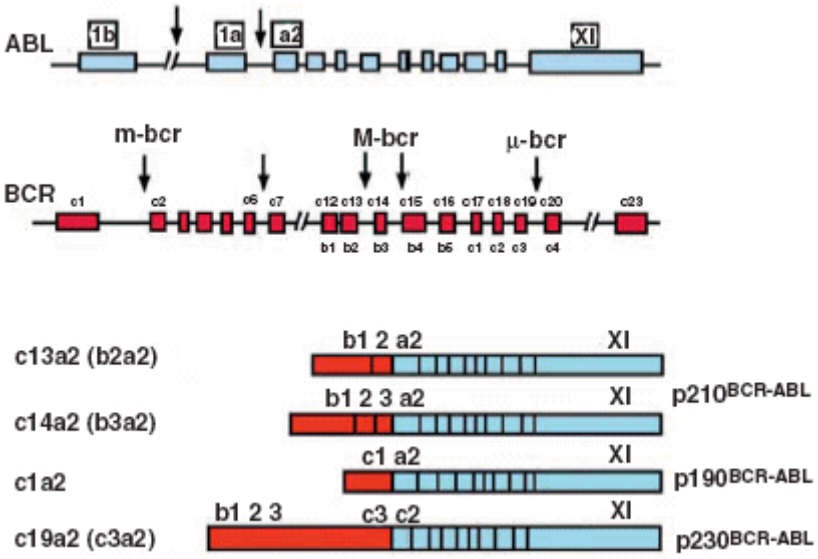



Figure 9. **BCR-ABL1 transcripts**

*On the top, ABL1 and BCR genes with exon-intron structure. Black arrows highlight the principal regions in which breakpoints occur in ABL1 and BCR. On the bottom the principal BCR-ABL1 transcripts that can arise from BCR-ABL1 fusion.*

### *1.6.3.3 Protein and function*

As said above, BCR-ABL1 oncogene can give rise to 3 distinct fusion proteins of molecular mass 190, 210, and 230 kD, which contain the same portion of the c-Abl tyrosine kinase in the COOH terminus but include different amounts of Bcr sequence at the $NH_2$ terminus [86].

The protein $p^{210}$ is present in 95% percent of CML patients; $p^{190}$ occurs rarely in CML and in 20-30% adults and 5% of childhood B-cell acute lymphocytic leukemia (B-ALL), whereas $p^{230}$ is detected in neutrophilic CML [87]. All three forms induce a similar myeloproliferative syndrome [88]. It has been widely demonstrated, by using both in-vitro and in-vivo biological models, that BCR-ABL1 oncoprotein is a fundamental requisite for promoting CML pathogenesis. In vitro experiments on human $CD34^+$ cells have shown that overexpression of BCR-ABL1 oncoprotein leads to increase of proliferation in response to growth factors, reduction of adhesion to fibronectin and increase of growth factor independent survival [57] [76].

The properties of BCR-ABL1 expressing cells mimics those found in CML progenitors cells isolated from patients [70] [57]. Studies performed by using a murine stem-cell retroviral vector to express the BCR-ABL oncogene in hematopoietic progenitor cells, have demonstrated that the expression of the transgene in mice BM was able to induce a myeloproliferative disorder (MPD) that resemble CML with 100% efficacy [57]. These animal models show CML specific features such as the increase of peripheral-blood (PB) cells (granulocytes) number, extramedullary hematopoiesis in liver, typical of BP, and splenomegaly widely present in CML patients. In vivo models, using mice expressing point mutation in the ATP-binding site of ABL1, do not develop CML demonstrating that kinase activity is essential for promoting BCR-ABL leukaemogenesis in vivo [88]. Hence, the deregulation of the tyrosine kinase activity is the key target for promoting CML and its features. By evidences collected from several studied performed both in vitro and in vivo, it has demonstrated that ABL1 kinase activity is essential but not sufficient for leading to CML-like MPD [88]. Several motifs and domain such as SH3, SH2 and the N-terminal coiled coil oligomerization domain of BCR are necessary to promote MPD [88] [57]. Mutations and deletions in these

domains negatively affect the kinase activity [57]. The oncogenic activity of BCR-ABL1 is exerted by the phosphorylation of several substrates that activate multiple signaling pathways. In particular, BCR-ABL1 promotes the development of a malignant phenotype by altering cells adhesion properties, activating the mitogenic signaling and in the meanwhile inhibiting the apoptosis [75] [69] [37]. The cellular adhesion is impaired by the regulation of integrins expression along with the phosphorylation of Crkl that regulates the cellular motility and the integrin-mediated cell adhesion [70]. The activation of mitogenic signaling along with the inhibition of apoptosis is hold by the phosphorylation of several molecules in RAS, STAT3 and 5, PI3K-AKT, NF-κB and JUN pathways [57] [70] [37] [89]. Uncontrolled signaling through these pathways induces abnormal proliferation and neoplastic expansion in addition to ability to escape from the negative regulation of cell proliferation by stromal cells.

Others researches have demonstrated that BCR-ABL1 is able to decrease the apoptotic response to mutagenic stimuli, providing an additional survival advantage to the leukemic clone [90].

## 2. AIM OF THE WORK

The fusion gene BCR-ABL1, that arises from the t(9:22) translocation, is the hallmark of Chronic Myeloid Leukemia (CML) and the specific target of pharmacological treatment. The detection of BCR-ABL1 transcript is used to follow up patients under treatment and to monitor treatment efficacy. However, this RNA based technique can lead to errors in quantifying tumor cells under expressing the fusion gene [60]. Thus, a DNA based marker of the translocation could overcome this limit and will facilitate patient management by confirming the absence of leukemic cells.

The primary aim of this project was to apply Next Generation Sequencing technology to identify BCR-ABL1 breakpoints in CML.

The patient specific breakpoints will be used to setup qPCR assays to monitor Minimal Residual Disease (MRD).

# 3. METHODS AND SOFTWARE

The project as a whole included both an experimental and a bioinformatics part. From now onwards the discussion will focus primarily on the bioinformatics analysis carried on by myself. Considering that the work of this thesis was pivoted around bioinformatics analysis, we decided to split the *"materials and methods"* section in *"methods and software"* section to emphasize the bioinformatics tools used in the data analysis.

## *3.1 Samples*

The samples analyzed included K562 cells (CCL-243) and 9 samples belonging to CML patients. K562 cells (CCL-243) were purchased from ATCC (http://www.lgcstandards-atcc.org) and cultured according to ATCC instruction. The nine human samples came from peripheral blood or blood marrow specimen. Samples were from patients with CML diagnosed at chronic phase. Patient's clinical conditions and relative laboratory data are shown in the table below Table 2. Informed consent for the use of cells for research was obtained in accordance with the Declaration of Helsinki and with approval of the Ethics Committee of the Insubria University and Hospital of Bergamo, Italy.

| Sample | Age | Molecular transcript | Anamnesis | Cytogenetic | | |
|---|---|---|---|---|---|---|
| | | | | CBA | nuc ish | ish |
| **CL.1** | --- | p210 | - | | | |
| **PT.2** | N.A | p210 | N.A | N.A | N.A | N.A |
| **PT.3** | N.A | p210 | N.A | 46,XY,t(9;22)(q34;q11)[19] | N.A | N.A |
| **PT.4** | 55 | p190 | Asthenic. Pain in the hands and feet. Pallor. | 46,XY [23] | N.A | N.A |
| **PT.5** | 85 | p230 | N.A | N.A | N.A | N.A |
| **PT.6** | 69 | p210 (b3a2) | Asthenia, and splenomegaly. Previous ischemic heart disease | N.A | N.A | N.A |
| **PT.7** | 13 | p210 (b3a2) | N.A | 46,XX,t(9;22)(q34;q11)[18] | (ABL,BCR)x2[400] | N.A |
| **PT.8** | 42 | p210 (b3a2) | Neutrophilic leukocytosis, heterozygous beta-thalassemia, insulin-dipendent by the age of 15, diabetic retinopathy | 46,XX,t(9;22)(q34;q11)[23] | N.A | N.A |
| **PT.9** | 41 | p210 (b2a2) | Leukocytosis | 46,XY[17] | (ABLx3,BCRx2)(ABL con BCRx1)[300] | t(X;9;22)(p11;q34;q11)(ABL+,BCR+;ABL+;BCR-)[10] |
| **PT.10** | 60 | p210 (b3a2) | Hypertensive Hypercholesterolemia | 46,XX,t(9;22)(q34;q11)[20] | N.A | N.A |

Table 2. **Sample features**

*Several features including the molecular transcript, the clinical condition and the cytogenetic of patients at diagnosis are described. Cytogenetic formula are formatted according to the International System for Human Cytogenetic Nomenclature (ISCN). In squared brackets, the number of metaphases (for CBA) and nuclei (FISH) at interphase [nuc-ish] or metaphase [ish] analyzed.*

## *3.2 Methods*

The methods section will focus on the chronological description of the procedures implemented in the different phases of the project.

### *3.2.1 Cytogenetic analysis*

Conventional cytogenetic analyses, carried out routinely, were subjected to quality control according to ISO 9001:2000 accreditation of the laboratory, and samples that did not conform to standards were rejected. Cytogenetic analysis was performed in laboratories of analysis at Bergamo Hospital (Italy) and Varese Hospital (Italy) by using three methods: CBA, FISH at metaphase (ish) and FISH at interphase (nuc-ish). All the findings are listed in **Table 2**.

### *3.2.2 Probes design for target enrichment of BCR-ABL1 region*

In order to analyze only the regions of interest where the t(9,22) occurs, we developed a customized sequencing target enrichment protocol. The rationale was to select the best and most selective subset of probes to both maximize the capturing of the area of interest and minimize the “non-specific” capturing. During the probe design, as mentioned in the introduction, it is important to take into consideration the genomic context in which the target region is located. Starting from BCR-ABL1 breakpoints regions reviewed in literature [81] [64] [60] [91] four target regions were chosen, in particular three regions for BCR (major, minor and micro breakpoints regions) and one region for ABL1 Table 3.

| Region | Chr | Interval | Length (kb) | %GC | %RR |
|---|---|---|---|---|---|
| **M-BCR** | 22 | 23,629,346 - 23,638,343 | 8.897 | 55.25 ± 20.82 | 18 |
| **m-BCR** | 22 | 23,523,148 - 23,596,167 | 73.020 | 50.44 ± 22.37 | 40.12 |
| **µ-BCR** | 22 | 23,652,511 - 23,660,223 | 7.713 | 56.64 ± 20 | 3.8 |
| **ABL1** | 9 | 133,577,268 - 133,730,483 | 153.215 | 43.51 ± 23.36 | 58 |

Table 3. **Targeted regions**
*The table shows in each column: the type of targeted region, chromosome, genomic interval, region length, %of GC and % of repeated regions according to repeat masker* [18].

Probes design was carried out using eArray software provided by Agilent (https://earray.chem.agilent.com/earray/).

The experimental design of the probes was performed considering the two main factors that can affect capturing, which are the GC percentage and the presence of repeated sequences. Hence, different sets of probes were tested to determine those that could capture the higher percentage of the regions of interest (ROI).

The probes design was based on tuning of two parameters: the tiling of the probes, which represent the number of times that the probes cover the target region, and the number of bases that overlap the repeated sequences (RS). In theory, in order to avoid unspecific capturing, repeated regions should be totally masked. However, in order to have the right balance between specificity and capturing rate of ROI, we also designed probes that span 20-40 bp of RS. Repeated sequences were masked with two programs: repeat-masker (RM) [18] and window-masker (WM) [92]. Tests were then performed by using genomic coordinates of repeated sequences detected by RM and those detected by both RM and WM.

For BCR target regions, it was established that the best sets of probes were those with a 5x tiling and with the overlapping on the repeated sequences detected by both RM and WM sets to 20 bp for the M-BCR [Table 5] and to 40 bp for m-BCR [Table 6] and the µ-BCR region [Table 7].

Concerning to the ABL1 region, the best set of probes was found to be the one with tiling at 4x and with 40bp overlapping with the coordinates of repeated sequences common to RM and WM [Table 8].

Using the top probe-set designed for M-BCR, we performed a test to assess its ability to capture 27 breakpoints in M-BCR previously identified by the Porta group. As can be seen from the table [Table 26], only two breakpoints could not be captured because they fell in a repeat sequences 299 bp long (AluSx1 [chr9: 23633112-23633411]).

This probe-set was also used for capturing. No further tests were performed on other three target regions. The tests performed to validarte the probe sets showed that the design of the probes is particularly affected by repeated sequences [Table 4]. Indeed ABL1 and m-BCR, that are characterized by the highest amount of repeated sequences, 58% and 40% respectively, could theoretically be captured for 63.5% and 74.8% of their length respectively [Table 8, Table 6]. Conversely, M-BCR and µ-BCR have a lower amount of repeats regions (18% and 3.8% respectively) and could be theoretically completely captured [Table 5, Table 7]. Using as masking set the coordinates of repeated sequences common to both methods (RM and WM) the capturing efficiency was higher than that obtained by using as masking set the coordinates of repeated sequences detected by RM only. This phenomenon is related to number and extension of the repeated elements. Higher is the extension of the genomic portion to mask lesser will be the percentage of capturing. The difference between the percentage of capturing using the coordinates of repeated sequences detected by RM and that obtained with coordinates detected by both software (RM+WM) resulted similar in all regions ranging from 11 to 16 and correlated with the size of the region itself [Table 4]. Longer is the region to target higher will be the chance to have different sets of repeated regions detected by the two different software (RM and WM). The only exception is the µ-BCR that

is both short in length and poor of repeated sequences, which are equally detected by both software [Table 4].

The main three-breakpoint cluster regions within BCR gene (M-bcr, m-bcr and µ-bcr) on chromosome 22 and the region from the exon 8 of EXOSC2 gene to exon 3 of ABL1 gene on chromosome 9 were captured. The total size of captured region was ~250 kb. Agilent RNA probes, of 120-bp, were with 4x tile across ABL1 and 5x across BCR.

| Region | Masking | Overlap | Tiling | % capture | Δ capture (Both –RM) |
|---|---|---|---|---|---|
| M-BCR | both | 20 | 5x | 93.0 | 11.1 |
| M-BCR | RM | 20 | 5x | 81.9 | |
| m-BCR | both | 20 | 5x | 71.361 | 12.26 |
| m-BCR | RM | 20 | 5x | 59.102 | |
| µ-BCR | both | 20 | 5x | 96.331 | 0 |
| µ-BCR | RM | 20 | 5x | 96.331 | |
| ABL1 | both | 20 | 5x | 55.975 | 16.46 |
| ABL1 | RM | 20 | 5x | 39.515 | |

Table 4. **Theoretical capturing with RM and RM ∩ WM masking**
*The table summarizes the theoretical percentage of capturing achievable by using probe sets with tiling 5X and overlap with repeated sequences of 20bp. The % of capture is collected both using probe set masked with RM and masked with coordinates common to WM and RM. In the last column the difference of % capturing between WM ∩ RM and RM.*

| M-BCR region (chr22:23629346 - 23638343) | | | | | | |
|---|---|---|---|---|---|---|
| **Test name** | **Tiling Frequency** | **Avoid Overlap** | **Total Number of Baits Info** | **Total Bases Covered by Baits** | **Target Length** | **% Target Base Pairs Covered** |
| BCR_2x_RM_over20 | 2x | 20 | 111 | 7073 | 8998 | 78.6 |
| BCR_5x_RM_over40 | 5x | 40 | 283 | 7367 | 8998 | 81.9 |
| BCR_5x_RM_over20 | 5x | 20 | 279 | 7367 | 8998 | 81.9 |
| BCR_4x_RM_over20 | 4x | 20 | 225 | 7373 | 8998 | 81.9 |
| BCR_3x_RM_over20 | 3x | 20 | 172 | 7423 | 8998 | 82.5 |
| BCR_3x_RM_over40 | 3x | 40 | 175 | 7463 | 8998 | 82.9 |
| BCR_3x_both_over20 | 3x | 20 | 197 | 8333 | 8998 | 92.6 |
| BCR_4x_both_over20 | 4x | 20 | 260 | 8333 | 8998 | 92.6 |
| BCR_3x_both_over40 | 3x | 40 | 201 | 8343 | 8998 | 92.7 |
| **BCR_5x_both_over20** | **5x** | **20** | **325** | **8364** | **8998** | **93.0** |
| BCR_5x_noRep | 5x | 1 | 371 | 8998 | 8998 | 100.0 |
| BCR_3x_noRep | 3x | 1 | 223 | 8998 | 8998 | 100.0 |

Table 5. **M-BCR probe sets evaluation**

*The table shows in each column: probe set name, tiling frequency (number of baits covering each base), base pairs overlapping repeated sequences, number of baits included in the probe set, base pairs covered by baits, length of the targeted sequence., % of target covered by the baits. On the top the region name with genomic coordinates. The best probe set is highlighted with bold character.*

| m-BCR region (chr22: 23,523,148 - 23,596,167) | | | | | | |
|---|---|---|---|---|---|---|
| **Test name** | **Tiling Frequency** | **Avoid Overlap** | **Total Number of Baits Info** | **Total Bases Covered by Baits** | **Target Length** | **% Target Base Pairs Covered** |
| BCR_minor_3x_RM_over20 | 3x | 20 | 966 | 42940 | 73020 | 58.806 |
| BCR_minor_5x_RM_over20 | 5x | 20 | 1575 | 43156 | 73020 | 59.102 |
| BCR_minor_4x_RM_over20 | 4x | 20 | 1273 | 43300 | 73020 | 59.299 |
| BCR_minor_5x_RM_over40 | 5x | 40 | 1594 | 43804 | 73020 | 59.989 |
| BCR_minor_4x_RM_over40 | 4x | 40 | 1289 | 43870 | 73020 | 60.079 |
| BCR_minor_3x_RM_over40 | 3x | 40 | 989 | 44180 | 73020 | 60.504 |
| BCR_minor_3x_both_over20 | 3x | 20 | 1148 | 51380 | 73020 | 70.364 |
| BCR_minor_4x_both_over20 | 4x | 20 | 1516 | 52026 | 73020 | 71.249 |
| BCR_minor_5x_both_over20 | 5x | 20 | 1880 | 52108 | 73020 | 71.361 |
| BCR_minor_3x_both_over40 | 3x | 40 | 1234 | 54585 | 73020 | 74.753 |
| **BCR_minor_5x_both_over40** | **5x** | **40** | **2007** | **54632** | **73020** | **74.818** |
| BCR_minor_4x_both_over40 | 4x | 40 | 1620 | 54742 | 73020 | 74.968 |

Table 6. **m-BCR probes set evaluation**

*The table shows in each column: probe set name, tiling frequency (number of baits covering each base), base pairs overlapping repeated sequences, number of baits included in the probe set, base pairs covered by baits, length of the targeted sequence., % of target covered by the baits. On the top the region name with genomic coordinates. The best probe set is highlighted with bold character.*

| μ-BCR region (chr22: 23,652511-23,660,223) | | | | | | |
|---|---|---|---|---|---|---|
| **Test name** | **Tiling Frequency** | **Avoid Overlap** | **Total Number of Baits Info** | **Total Bases Covered by Baits** | **Target Length** | **% Target Base Pairs Covered** |
| BCR_micro_3x_RM_over20 | 3x | 20 | 182 | 7430 | 7713 | 96.331 |
| BCR_micro_3x_both_over20 | 3x | 20 | 182 | 7430 | 7713 | 96.331 |
| BCR_micro_3x_RM_over40 | 3x | 40 | 182 | 7430 | 7713 | 96.331 |
| BCR_micro_3x_both_over40 | 3x | 40 | 182 | 7430 | 7713 | 96.331 |
| BCR_micro_5x_RM_over20 | 5x | 20 | 302 | 7430 | 7713 | 96.331 |
| **BCR_micro_5x_both_over20** | **5x** | **20** | **302** | **7430** | **7713** | **96.331** |
| BCR_micro_5x_RM_over40 | 5x | 40 | 302 | 7430 | 7713 | 96.331 |
| BCR_micro_5x_both_over40 | 5x | 40 | 302 | 7430 | 7713 | 96.331 |
| BCR_micro_4x_RM_over20 | 4x | 20 | 243 | 7445 | 7713 | 96.525 |
| BCR_micro_4x_RM_over40 | 4x | 40 | 243 | 7445 | 7713 | 96.525 |
| BCR_micro_4x_both_over40 | 4x | 40 | 243 | 7445 | 7713 | 96.525 |
| BCR_micro_4x_both_over20 | 4x | 20 | 243 | 7445 | 7713 | 96.525 |

Table 7. **μ-BCR probes sets evaluation**

*The table shows in each column: probe set name, tiling frequency (number of baits covering each base), base pairs overlapping repeated sequences, number of baits included in the probe set, base pairs covered by baits, length of the targeted sequence., % of target covered by the baits. On the top the region name with genomic coordinates. The best probe set is highlighted with bold character.*

| ABL1 region (chr9:133577268-133730483) | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Test name** | **Tiling Frequency** | **Avoid Overlap** | **Total Number of Baits Info** | **Baits Removed Due to Avoid Overlap** | **Total Bases Covered by Baits** | **Target Length** | **% Target Base Pairs Covered** |
| ABL1_3x_RM_over20 | 3x | 20 | 1266 | 553 | 59328 | 153216 | 38.722 |
| ABL1_5x_RM_over20 | 5x | 20 | 2051 | 818 | 60543 | 153216 | 39.515 |
| ABL1_4x_RM_over20 | 4x | 20 | 1662 | 690 | 60645 | 153216 | 39.581 |
| ABL1_5x_RM_over40 | 5x | 40 | 2146 | 712 | 63687 | 153216 | 41.567 |
| ABL1_4x_RM_over40 | 4x | 40 | 1749 | 591 | 63795 | 153216 | 41.637 |
| ABL1_3x_RM_over40 | 3x | 40 | 1355 | 452 | 64335 | 153216 | 41.99 |
| ABL1_3x_both_over20 | 3x | 20 | 1748 | 1173 | 83582 | 153216 | 54.552 |
| ABL1_5x_both_over20 | 5x | 20 | 2834 | 1702 | 85763 | 153216 | 55.975 |
| ABL1_4x_both_over20 | 4x | 20 | 2302 | 1426 | 85955 | 153216 | 56.101 |
| ABL1_5x_both_over40 | 5x | 40 | 3239 | 946 | 96899 | 153216 | 63.243 |
| ABL1_3x_both_over40 | 3x | 40 | 2032 | 635 | 96980 | 153216 | 63.3 |
| **ABL1_4x_both_over40** | **4x** | **40** | **2650** | **766** | **97332** | **153216** | **63.526** |

Table 8. **ABL1 probes sets evaluation**

*The table shows in each column: probe set name, tiling frequency (number of baits covering each base), base pairs overlapping repeated sequences, number of baits included in the probe set, base pairs covered by baits, length of the targeted sequence., % of target covered by the baits. On the top the region name with genomic coordinates. The best probe set is highlighted with bold character.*

### *3.2.3 Sequencing Library preparation*

Genomic DNA (gDNA) of each patient was extracted from peripheral blood (PB) or bone marrow specimen (BM) by using standard methods and from K562 as well. SureSelect XT Target Enrichment kit (Agilent Technologies, Santa Clara, CA, USA) was used for DNA library preparation and capturing of the target regions.

In brief, 4 µg of each gDNA were sheared on Covaris Instrument (Covaris, Woburn, MA, USA) and a fragment size distribution ranging from 100 to 1000 bp was verified on Bioanalyzer 2100 (Agilent Technologies). The DNA fragments were end-repaired, ligated to Illumina adapters (Illumina, San Diego, CA, USA) and the fragments ranging from 400 bp to 700 bp in length were selected on 2% agarose gel. After 10 cycles of PCR, each amplified adapter-ligated library was run on Bioanalyzer 2100 for quantification. 500 ng of each library was treated with the capture biotinylated RNA probes for 24 h at 65°C. The RNA/DNA hybrid target regions were then selected onto the streptavidin-coated magnetic beads DynabeadsMyOne Streptavidin T1 (Invitrogen Life Technologies, Carlsbad, CA, USA). Finally, each targeted DNA library was PCR enriched and during this step index tags were also added to each construct.

Two pools of libraries with different index tags, one with 5 and the other with 4 libraries, were setup and each pool was then sequenced in four lanes of a flow cell on GAIIx, an illumina platform. Sequencing was performed with a paired-end protocol to produce sequences of 114 bp in length.

### *3.2.4 Data analysis Pipeline*

BCR-ABL1 breakpoints identification was achieved by setting up an automatic bioinformatics pipeline. The rationale was to look for reads pairs that did not map with an expected insert size, in particular pairs in which one read maps in chromosome 22 (in one of the BCR breakpoint regions), and its mate maps in chromosome 9 in ABL1. The bioinformatics pipeline includes five core modules [Figure 10]:

1) Quality control checks on raw reads
2) Pre-alignment data processing
3) Reads alignment to the human reference genome (Hg19)
4) Post-alignment data processing
5) BCR-ABL1 breakpoints identification

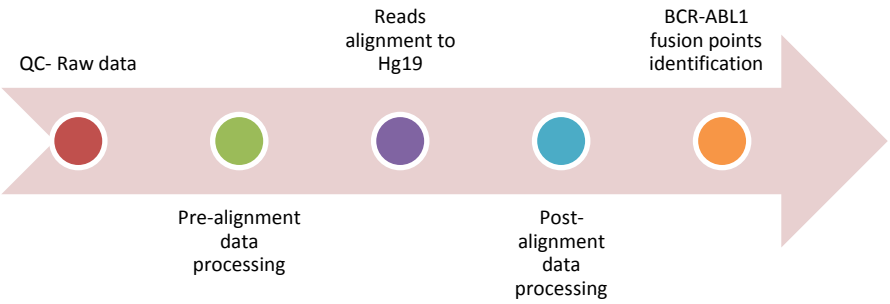



Figure 10. **Data Analysis Workflow**
*The chart briefly depicts the data analysis workflow adopted in this study. Raw data were checked for quality and appropriate procedures were carried out to ensure the best quality data to align to human reference genome (hg19). Aligned data were than processed to increase the data quality. Finally we performed BCR-ABL1 fusion point identification using a two-steps procedure: identification of the breakpoints regions with discordant pair approach (SVDetect) and refinement to single nucleotides breakpoint using split-read approach (ClipCrop + BLAT).*

#### *3.2.4.1 Quality control checks on raw reads*

The quality control was performed with FastQC (v.0.10.0) [93] evaluating several metrics such as "*per sequence quality score"* and "*per base sequence quality"*. The first metric allows detecting subsets of sequences that have a low score and the second allows spotting base quality drops during sequencing run. More details about these metrics and statistics will be discussed in the software section.

#### *3.2.4.2 Pre-alignment data processing*

Raw data were trimmed using Prinseq [94]. Trimming is a read-cutting operation for increasing the quality of the reads that have to be aligned to a reference genome. In general, trimming is a data pre-processing practice that allows to work on very high quality reads only. In the current project, most of the targeted area is characterized by a large amount of repeated sequences. In this case, it has been proved that with a high quality set of reads the percentage of reads mapped to the reference is increased [95] [96]. In our pipeline, a dynamic trimming was setup to ensure both high quality and low loss of nucleotides. In particular, the trimming was setup for cutting low quality bases from 3' end using the following parameters: -m 30 -l 90 –t 20. The t value defines the maximum base quality score for trimming nucleotides; the l value is the minimum read length allowed and the m value defines the minimum mean base quality over the whole read.
Filters were applied in this order: t, l and m. The trimming proceeded from 3' to 5' until the next-cutting nucleotide had a quality score greater than 20. At this point if the read length was greater than 90, the m value was then checked. If both m and l parameters were not satisfied the read was discarded. After the pre-processing step a subset of raw reads was produced, in which the minimum reads length was 90bp and the minimum average base quality was 30. Scores are expressed in Phred format [97], which will be discussed later in the software section.

#### *3.2.4.3 Reads Alignment and data processing*

The alignment to the reference genome (Hg19) was carried on by using the Burrows-Wheeler Aligner (BWA v.0.6.2) [98] with default parameters. The choice to use default parameters arise from several tests tuning mainly two parameters such as the n, l and k that account for maximum mismatch bases allowed, seed length and maximum differences occurring in the seed respectively. These tests did not show any improvement in mapping. Several tries to align reads using multi-hits aligner like mrFAST [99] to Hg19 were also carried on. Results obtained by this latter approach were not satisfactory and are not presented here.

We used as reference sequence the human reference genome (Hg19) retrieved by UCSC repository [100].

Aligned reads were then processed by Genome Analysis Toolkit (GATK) [101] in order to ensure high quality mapped reads for breakpoint identification. The most used and recommended procedures were applied: reads duplicates marking (DM), local realignment around INDELs (LRAI) and base quality score recalibration (BQSR) [102].

#### *3.2.4.4 Breakpoints identification*

The identification of breakpoints coordinates on Hg19 was carried out in two phases. In the first phase we identified the region in which the breakpoints occur by using the discordant pair approach, whereas in the second phase we refined the breakpoints coordinates through a split-read approach.

##### 3.2.4.4.1 Breakpoints detection

We used SVDetect [32] for detection and ClipCrop [35] and BLAT [103] for refinement. Among the methods developed to identify SV from NGS, SVdetect use the *discordant pair* approach, whereas ClipCrop use the *split read* approach. SVdetect, as described in detail in the software section, is able to find structural variations such as translocations, inversions,

deletions and small duplications, by using paired-end reads that do dot map with concordant/normal insert size. The insert size is the distance between the paired-end adapters (forward and reverse). This distance should be theoretically equal to the length of the DNA fragments captured by the probes. The insert size is the difference between the start coordinates of reverse and forward reads. Hence, considering that the library size without adapters was ~ 500bp in length [Figure 11] and both adapters used were 60bp in length, the expected insert size was ~380bp.

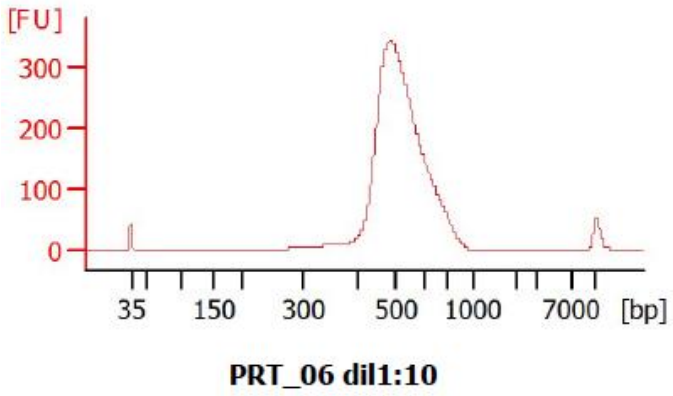



**Figure 11. Library size - Bioanalyzer**
*The figure shows the distribution of the sequenced fragments' lengths. On the X-axis the length of the fragments (bp) whereas on Y-axis the frequency of fragments expressed as fluorescent signal (FU). The peak is at 500 bp.*

The "normal" insert size was calculated as the mean insert size (μ) of normally mapped paired-end reads [32]. In order to select a subset of paired-end reads, which do not map with a typical range of insert size, a sigma threshold ($\sigma_t$) parameter is setup. This parameter defines how many standard deviations from (μ) can be accepted to consider the insert size as normal.

Hence, reads which had an insert size outside the interval ( $\mu - \sigma_t * \sigma < I < \mu + \sigma_t * \sigma$), were used for SV identification.

In particular, two different $\sigma_t$ ($i\sigma_t$ , $d\sigma_t$) were set for INDELs and duplication detection respectively. More details regarding the computational mechanisms, which underlie the SVs identification, will be described in the software section.

| Sample | μ | σ |
|---|---|---|
| **K562** | 348 | 91 |
| **02** | 372 | 82 |
| **03** | 364 | 79 |
| **04** | 340 | 81 |
| **05** | 340 | 85 |
| **06** | 352 | 103 |
| **07** | 349 | 77 |
| **08** | 346 | 84 |
| **09** | 342 | 81 |
| **10** | 330 | 79 |

Table 9. **Insert size and sigma value**
*The table shows the average insert size (μ) and the value σ corresponding to a single s.d from the normal distribution of the insert size length.*

The SVdetect output was primary inspected for inter-chromosomal translocations that occur between chromosome 22 and 9. A secondary analysis was performed for identifying intra-chromosomal structural variants such as deletions, insertions and inversions, but results will not covered in this work. Both inter and intra SV detected were filtered by tuning the nb_pairs_threshold to a value of 20. The best breakpoints, in terms of reads pairs and final score, spotting both BCR-ABL1 and ABL1-BCR fusion products were selected for coordinate's refinement. The parameters set in the configuration file are listed in (Table 10).

| Parameter | Value |
|---|---|
| **Split link file** | 1 |
| **Strand filtering** | 1 |
| **Order filtering** | 1 |
| **Insert size filtering** | 1 |
| **Nb pairs threshold** | 20 |
| **Indel sigma threshold** | 2 |
| **Dup sigma threshold** | 3 |
| **Final score threshold** | 0.85 (INTER) / 0.95 (INTRA) |
| **Mu length (μ)** | Sample dependent [table CC] |
| **Sigma length(σ)** | Is calculated as 1 st.dev from μ |
| **Window size** | Suggest value for detecting translocation is equal to: $2\mu + 2\sqrt{2} * \sigma$ |
| **Step length** | ½ to ¼ of windows size value is the value suggested. |

Table 10. **SvDetect Parameters. Put description**

*The table summarizes the parameters used in SVDetect analysis. From top to bottom: first 4th rows set output characteristics and filtering, Nb pairs threshold describe the number of discordant pair reads supporting the SV, Indel and Sup sigma threshold establish how many s.d consider to use read in indel or dup calling. The final score threshold set a score to filter output results, $\mu$ and $\sigma$ described in* Table 9*, window size is the size of the sliding window to cluster PEM reads and create links. In the last row the step length that define how to slide across windows.*

#### *3.2.4.4.2 Breakpoints coordinates refinement*

The breakpoints coordinates refinement was carried out using ClipCrop, a tool for detecting structural variations with soft-clipping information. Soft-clipped sequences are partially unmatched fragments in a mapped read.

BWA aligner is able to re-map part of a read whose alignment failed in its whole length. If the read is mapped partially, the information of the partial mapping is stored into SAM format as clipping information [104].

ClipCrop can calculate the putative breakpoints using the boundary position between mapped sequence and the soft-clipped sequence in the clipped read. Then it remaps soft-clipped sequences around the detected putative breakpoints and infers which type of SV is really occurring in that region.

This procedure is natively implemented in ClipCrop by using BWA or Shrimp [105] as aligner for the re-mapping phase of clipped portions of split-reads.

In our pipeline instead of using BWA or Shrimp, we used BLAT for the remapping of the clipped reads. Details concerning the algorithm implemented in ClipCrop will be described in the relative software section.

BCR-ABL1 breakpoints identified by ClipCrop were filtered out for records that did not fall near (± 400bp) the boundaries coordinates of the breakpoint region identified by SVDetect. Therefore, the best-refined coordinates for each patient breakpoint were selected considering the direction of soft-clipped sequence in clipped reads. In particular, the coordinates of the breakpoints spotted by right-clipped reads in BCR and the coordinates of the breakpoints identified by left-clipped reads in ABL1 determined the BCR-ABL1 breakpoints, whereas the contrary for ABL1-BCR. More details can be found in the related software section. The setup parameters are listed in table [Table 11].

| PARAMETER | VALUE |
|---|---|
| **BAM FILE** | Sample_02.bam |
| **FASTA FILE** | Hg19.fasta |
| **BP FILTER PROCESSES** | 12 |
| **MAX BREAKPOINT DIFF** | 2 |
| **MIN BP CLUSTER SIZE** | 8 |
| **MIN MEAN BASE QUAL** | 10 |
| **MIN SEQ LENGTH** | 10 |
| **BASES AROUND BREAK** | 1000 |
| **MAX SV DIFF** | 10 |
| **MIN SV CLUSTER SIZE** | 8 |
| **MAPPER** | bwa |
| **MAPPER THREADS** | 12 |

Table 11. **ClipCrop parameters**
*The table shows parameters setup in ClipCrop analysis.*

### *3.2.5 Sanger sequencing validation*

Sanger sequencing validation was carried out in outsourcing. Chromatograms, received as .abl1 files, were visualized by CromasLite [http://technelysium.com.au/] and extracted in FASTA format. FASTA files from each patient were then aligned with BLAST [106] and BLAT [103] to the reference genome in order to validate the breakpoints coordinates identified during the analysis.

### *3.2.6 BCR-ABL1 and ABL1-BCR breakpoint analysis*

The description was carried out highlighting:

- The distance between the breakpoints and the repeated sequences in surrounding regions.
- The localization at gene level (distance to nearest exon)
- The type of DNA break (inferring)

Moreover, a specific and deeper analysis was performed for sample 7 and in order to shed light on the complex rearrangement found in it.

## 3.3 Bioinformatics Tools

### *3.3.1 Data processing Tools*

The following section will include the description of the software used in data processing from raw reads to breakpoints identification.

#### *3.3.1.1 QC on raw data – FastQC and PrinseQ*

The quality control on raw data produced by GAIIx was performed by FastQC. This program outputs a series of metrics and statistics starting from sequences files in FASTA (Sanger) or FASTQ (Illumina) format. The output can be exported both in text and charts-like format. The metrics and statistics produced are:

1. Basic Statistics
2. Per base sequence quality
3. Per sequence quality scores
4. Per base sequence content
5. Per base GC content
6. Per sequence GC content
7. Per base N content
8. Sequence length distribution
9. Sequence level duplication
10. Overrepresented sequences
11. Kmer content

The most important metrics that describe the quality of the data are the "*per base sequence quality"* and "*per sequence quality scores"* metrics. The former describes how the quality of the base changes during sequencing cycles, whereas the latter expresses the distribution of the average base quality score along the reads. Base quality score usually decreases while proceeding in the sequencing cycles. This is due to the drop of efficiency that affect the DNA polymerase during sequencing. The Illumina base quality score ranges from 2 to 40 according to Phred score [97]. Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P. Q is equal to:

$$Q = -10*\text{Log}_{10}P \text{ , so } P = 10^{(-Q/10)}$$

Hence, a quality score of 40 means an error every 10000 base call, so a percentage of error of 0.01% or base call accuracy of 99,99%. Following this formula, a Phred score of 20 is linked to a base call accuracy of 99%. A Phred score of 20 is considered an appropriate cut off value for considering a base reliable or not.

The "*per sequence quality scores"* metric allows to view how the "basecall goodness" is distributed in the pool of reads produced. A good run shows a single peak of frequency around a mean Phred quality score of 36-40.
Per base sequence content, kmer content and overrepresented sequences are good metrics for identifying possible contaminations or the presence of sequencing adapters and indexes.
Other metrics such as the duplication levels and the sequence length distribution are strictly associated to the nature of the region sequenced and to the experimental (PCR-amplification cycles) and data processing procedures, such as the trimming, adopted.
Trimming was performed by PrinseQ (v0.20.1), a tools for NGS data quality control. The program has many features for processing raw data (trimming, filtering and reformatting) and generates many statistics. The only additional features implemented in PrinseQ compared to fastQC are the reads' complexity analysis and metrics that concern the composition of tag sequences and poly-A/T tails. The trimming procedure can be carried on in several ways as removing the 3' or 5' ends by a fixed number of bases or remove entirely a reads that has an average base quality below a fixed threshold or remove a percentage of the reads, and so on. The trimming procedure implemented in this work has been described in chapter 3.2.4.2.

#### *3.3.1.2 GATK, Samtools and PicardTools*

The Genome Analysis Toolkit (GATK) is a suite of bioinformatics tools for handling and analyzing NGS data. GATK is structured in walkers, single computational modules, which can analyze NGS data in several ways., Differently by other competitor software, GATK includes different modules for the fine processing of mapping data. In particular there are two modules exclusively implemented in this suite: the local realignment around indels (LRAI) and the base quality score recalibration (BQSR). LRAI is a module that realign mapped reads in complex regions in which the mapping could

be more difficult for the presence of gaps like small insertion and deletions. The BSQR is a procedure that reassign the basecall score taking into consideration the genomic context in which the reads map [102]. GATK include also two different modules for SNPs and INDELs call: GenotypeCaller and HaplotypeCaller. The first one is more driven to SNPcall, whereas the second perform better in INDELs call. Details on the algorithms beyond such modules are described in [102]. GATK comprises also a modules for coverage statistics (DepthOfCoverage) that was used for the production of the coverage statistics [101].

Samtools [104] has been mainly used for handling mapping files (sorting, indexing and file format conversion).

PicardTools [116] was used for marking of duplicates. This procedure consists in the removing of duplicate reads that usually comes when the number of PCR cycles is too high. The percentage of duplicates can vary depending on the target size and on the library preparation. Theoretically smaller is the target region, greater will be the probability to have duplicates, because the chance to produce the same fragments during the library preparation is higher.

### *3.3.2 Ph Breakpoints identification Tools*

The identification of the breakpoints was made using two software: SVDetect and ClipCrop. It was decided to use these software for several reasons: SVDetect is able to identify a number of events greater than GASV [107] and BreakDancer [108] and has the peculiarity to generate input files to be used with Circos [109]. ClipCrop compared to competing software in the category of split-read methods (Pindel and BreakDancer) has proven to be more accurate [35]. Moreover, intermediates files produces by ClipCrop were used as input files for custom scripts, for instance to align split-reads with BLAT.

#### *3.3.2.1 SVDetect*

SVDetect is a bioinformatics tools developed by Zeitouni et al. [32], to identify genomic structural variations from paired-ends reads and mate-pair from NGS data produced by ILLUMINA and SOLID platforms. SVDetect falls into the category of software that rely on discordant-pair based methods.

Compared to similar software like BreakDancer and GASV it can detect duplications, discriminate balanced and unbalanced events, support multisample SV analysis and last but not least, provide a graphical output for SV viewing by using Circos.

The workflow analysis implemented in SVDetect can be summarize in four steps:

1. anomalous reads selection
2. creation of links through window-sliding approach
3. filtering and clustering of putative links
4. detection of SV type based on links features

The first steps select all paired-end reads that can be potentially support a SV event. In particular, paired-end reads, which map with incorrect orientation and with a non-typical insert size, are selected.

At this step the mean insert size of normally mapped paired-end reads ($\mu$) and the calculated standard deviation (s.d) value from the distribution of normally mapped paired-end reads ($\sigma$) are collected.

The second step is to split the reference genome into overlapping windows of fixed size and using a windows-slide approach to identify groups of pairs, which shared the same genomic region. If a read, by its ends, can anchor two windows, a link between the pair of windows can be formed. Every link that connect two genomic fragments (windows size length) is then annotated with several features like chromosomal location, number of reads pairs that cover these fragment, orientation and order of paired-end

reads. Starting from these features, the SVDetect algorithm performs two operations: links filtering and clusters call. The links filtering is based on a collection of parameters, which can be set by the user. There are both boolean and floating parameters. Strand, order and insert size filtering parameters are boolean, and can be set 1 or 0 depending on the choice to filter or not on strand, order and insert size features. These parameters are strictly dependent on μ and σ parameters. Among the floating parameters, there are the nb_pairs_threshold (npt), the indel_sigma_threshold (ist) and the nb_pairs_order_threshold (npot). The first parameter determines the minimum number of pairs within the cluster in order to call the cluster itself. The second one establishes the number of s.d from sigma for insert size filtering and for the INDELs calling (if the mean insert size of the cluster is greater than μ + ist * σ a deletion is called; whereas if it is lower than μ + ist * σ an insertion is called). The last parameter sets the minimum number of pairs in a subgroup of paired-end reads for balanced events.

The last step consists in the detection of SV type starting from features of paired-ends reads within the clusters. In particular, nine structural variants can be detected: deletions [Figure 12], insertions [Figure 13], duplications, inversions, duplicated inversions, balanced [Figure 14] and unbalanced [Figure 15] translocations and balanced/unbalanced inverted translocations. The description of each event by visualization of discordant-pair mapping is provided inside the figures captions. The sets of features, which allow discriminating these events, are summarized in Table 12. Translocation events can be intra-chromosomal or inter-chromosomal, if the pairs of the paired-end reads maps on the same or on different chromosomes respectively.

| SV type | Mapping chromosome | Strand sense of cluster | Mean insert size of cluster | Balancing |
|---|---|---|---|---|
| **Deletion** | Same | Normal | > (μ + ist* σ) | - |
| **Insertion** | Same | Normal | < (μ + ist* σ) | - |
| **Inversion** | Same | Reverse | - | Balanced |
| **Duplication** | Same | Reverse | > (μ + dst* σ) | - |
| **Duplicated inversions** | Same | Reverse | < (μ + dst* σ) | - |
| **Unbalanced translocations** | Different | Normal | | Unbalanced |
| **Balanced translocations** | Different | Normal | | Balanced |
| **Unbalanced inverted translocations** | Different | Reverse | | Unbalanced |
| **Balanced inverted translocations** | Different | Reverse | | Balanced |

Table 12. **Structural events detectable by SVDetect**
*The table shows the structural events detectable by SVDetect including different features. The 2nd column describe if the SV is intra or inter chromosomal, the 3rd one the strand sense of the cluster that identify the event, the 4th one describes the mean insert size and the last column shows the balancing of the event.*

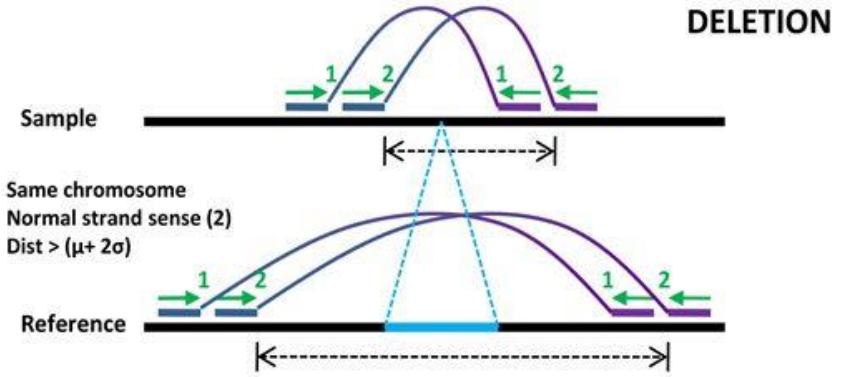



Figure 12. **SVDetect deletion event**

*The figure shows how paired-end mapping reads (PEMs) are used by SVDetect to detect a deletion event. In particular a deletion event is called when the insert size distance is greater than (μ +2σ) and pairs are on normal strand sense.*

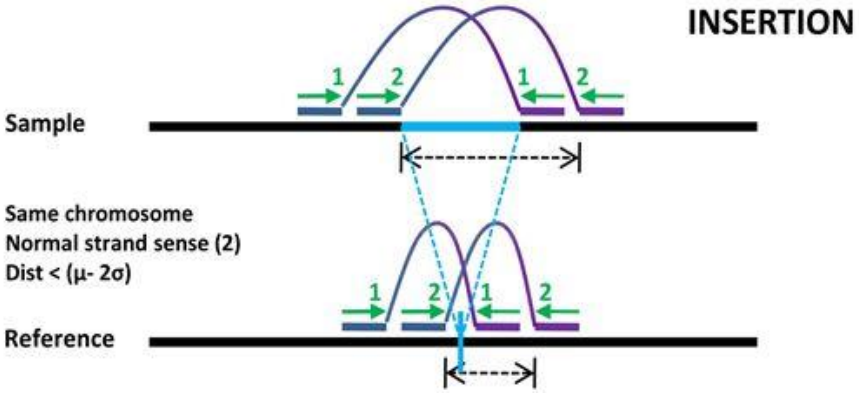



Figure 13. **SVDetect insertion event**

*The figure shows how paired-end mapping reads (PEMs) are used by SVDetect to detect a deletion event. In particular a deletion event is called when the insert size distance is lower than (μ - 2σ) and pairs are on normal strand sense.*

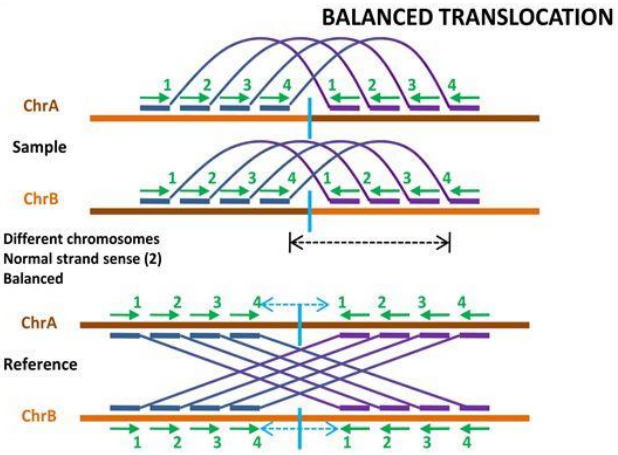



Figure 14. **SVDetect – Balanced translocation event**
*The figure shows how discordant pairs reads map to the genome when SVDetect calls a balanced translocation event. In the sample (on the top) the mates of a pairs map to different chromosomes, whereas in reference they map with the expected (normal) insert size. The translocation is balanced because in sample the genomic portions coming from different chromosomes is preserved.*

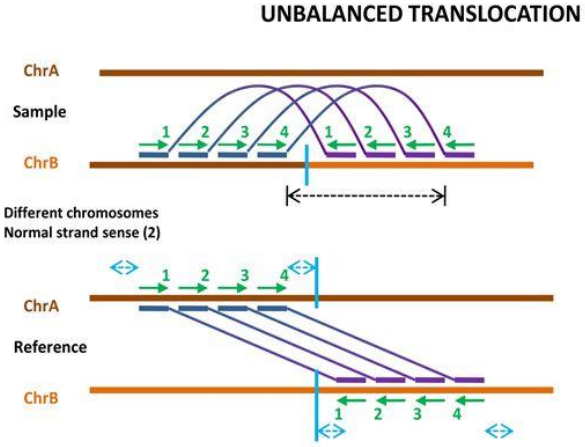



Figure 15. **SVDetect – Unbalanced translocation event**
*The figure shows how discordant pairs reads map to the genome when SVDetect calls a balanced translocation event. In the sample (on the top) the mates of a pair map to different chromosomes, whereas in reference they map with the expected (normal) insert size. The translocation is unbalanced because in sample the genomic portions coming from different chromosomes are not preserved.*

#### *3.3.2.2 ClipCrop*

ClipCrop is a bioinformatics tool that aim to identify SV by using split-read approach. ClipCrop uses clipping information to infer the breakpoints boundaries that lead to different types of SV. In particular, ClipCrop analysis consists in the following steps:

- selection of soft-clipped reads;
- discard of soft-clipped reads with both ends clipped;
- listing of breakpoints from clipping information;
- remapping of soft-clipped fragments with length greater than a specific threshold set by user around the breakpoint position
- SV type inference.

The selection of clipped reads is based on the analysis of the CIGAR field in the SAM mapping format [104]. The CIGAR string describe how the read maps to the genome. It contains both numerical and character elements, the numbers describe how many times the associated feature (expressed by the character) is contiguously detected within the read. The CIGAR string contains several character values as shown in Table 13. For instance a read with CIGAR 22S50M2I30M is a read that is clipped at the left-side for 22 nucleotides, it then match the reference sequence for 50 nucleotides, then it has an insertion of 2 bases and ends with a perfect match of 30 bp.

| Symbol | Description |
|---|---|
| **M** | alignment match (can be a sequence match or mismatch) |
| **I** | insertion to the reference |
| **D** | deletion from the reference |
| **N** | skipped region from the reference |
| **S** | soft clipping (clipped sequences present in SEQ) |
| **H** | hard clipping (clipped sequences NOT present in SEQ) |

Table 13. **Cigar field**
*The table describes values included and described in CIGAR field. On the left column the tag included in the CIGAR field in SAM files whereas on the right the description of each tag.*

ClipCrop selects all reads that contains the soft-clipped information (S) and discards the ones that contain soft-clipped portions at both sides (ends). These reads are not informative because is not possible to infer the directionality of the event they can lead to. Hence, starting from these reads as input, a list of breakpoints is created. The breakpoint is denoted as the marginal point between the clipped portion and the matched portion of the soft-clipped read. When the left-side of the read is clipped the breakpoint is designated as L-breakpoint, instead R-breakpoint when the clipped portion is on the right side. Once breakpoints are identified, they are sorted and clustered within 5-base differences. Hence, soft-clipped fragments with a length greater than 10bp are remapped by using BWA aligner (default) or Shrimp to a portion of genome flanking 1000bp (default) at both sides of each breakpoint in order to reduce the probability of wrong mapping and so minimize the a-specific mapping to other regions. This step is particularly important when the breakpoint falls near repeated sequences. Once remapped, ClipCrop infers the type of SV taking into consideration the re-mapping pattern of clipped portion of split-reads. When the clipped portion of an L-breakpoint maps to the left side of an R-breakpoint and vice versa,

it denotes a deletion event [Figure 16]. When the clipped portion of an L-breakpoint maps reversely to the left side of an R-breakpoint and vice versa, it denotes an inversion event [Figure 17]. In insertions and translocations, an L-breakpoint and an R-breakpoint are in the same position. If the clipped portion of the reads remaps to the reference genome, it denotes translocation [Figure 18], otherwise it indicates an insertion event.

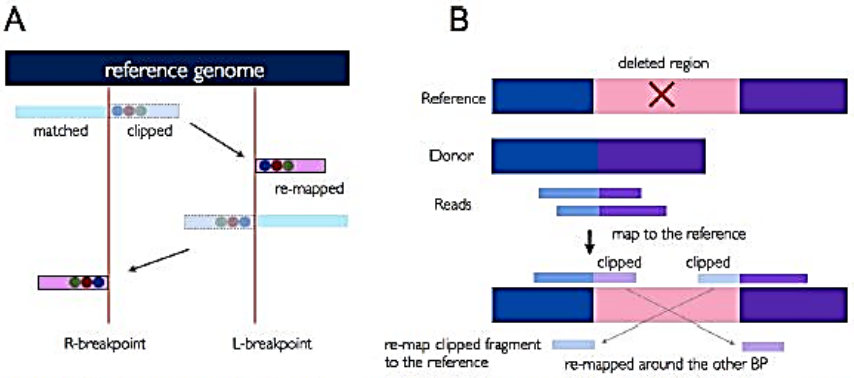



Figure 16. **ClipCrop: Deletion event**
*The figure depicts how split-reads will map when a deletion event occur in a sample. On the left, light-blue reads mapping to the reference genome. Clipped portions (L and R) are then remapped (pink remapped reads). On the right the same concept using donor and reference regions.*

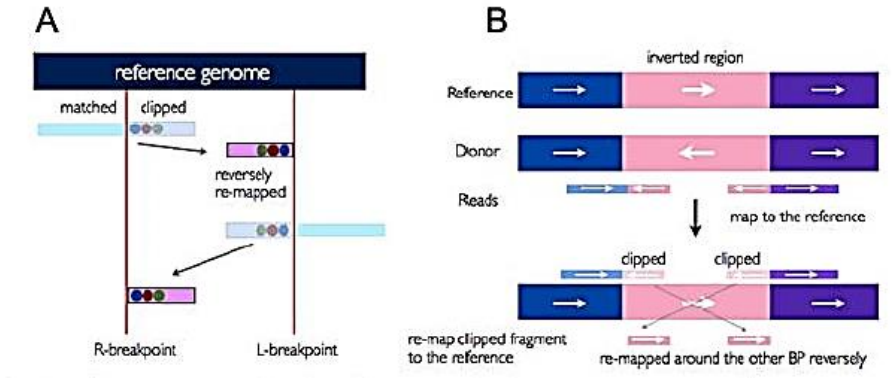



Figure 17. **ClipCrop: Inversion event**

*The figure depicts how split-reads will map when an inversion event occur. On the left, light-blue split reads mapping to the reference genome. Clipped portions (L and R) are then reversely remapped (pink remapped reads). On the right the same concept using donor and reference regions.*

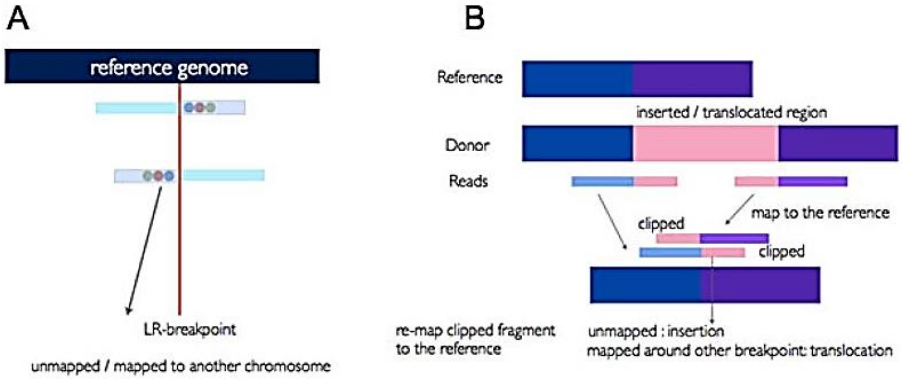



Figure 18. **ClipCrop: Insertion – Translocation event**

*The figure depicts how split-reads can spot insertion and translocation events. On the left, light-blue split reads mapping to the reference genome. Clipped portions (L and R) can unmap (insertion) or map to another chromosome (translocation). On the right the same concept using donor and reference regions.*

# 4. RESULTS

The results section will be divided in three parts: the first will be focused on the description of the target sequencing experiment; the second will deal with BCR-ABL1 breakpoints identification along with the analysis of regions surrounding the breakpoints and the final part will be dedicated to the sample-specific examination of BCR-ABL1 and ABL1-BCR breakpoints. The analyses, from raw data to breakpoints identification, has been performed for each sample in two different ways. Firstly, we identify the breakpoints by using four independent set of reads belonging to each lane, and secondly we performed the analysis merging the reads of different lanes (belonging to the same sample) and removing duplicate reads.

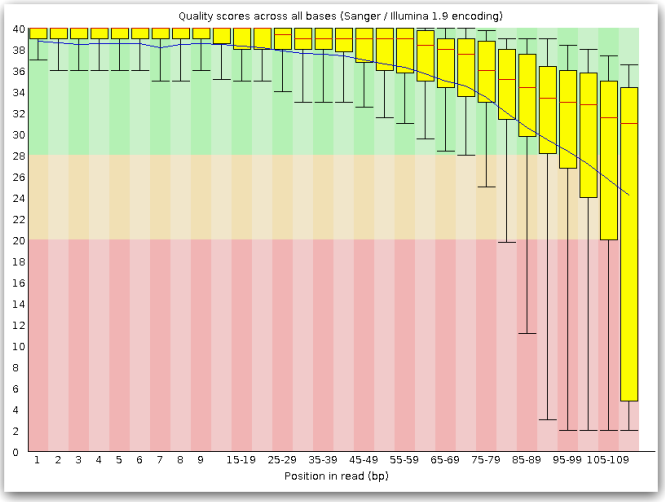
## 4.1 BCR-ABL1 target sequencing

The BCR-ABL1 oncogene target sequencing was performed in two stages. First, we assessed its feasibility with the inclusion in the experiment of known BCR-ABL1 breakpoints in 2 patients and 1 cell line; the second effort was made to reproduce the experiment in patients with unknown breakpoints. The following paragraphs will deal with the results obtained by quality control procedures and with all metrics and statistics concerning the capturing performance of the targeted regions.

### *4.1.1 Raw data and QC*

Total raw reads yield per lane ranged from 11.182.874 (sample 03a) to 15.043.842 (sample 09b). Raw reads were subjected to quality control using fastQC. Reports showed a base quality drop in the 3'-tail of both forward and reverse pair sequences [Figure 19]. In order to achieve reads of better quality, dynamic trimming with Prinseq was performed on both forward and reverse fastq files for each sample. Trimming parameters were setup in order to filter out reads with a mean quality score below 30 and length below 90 by trimming bases with quality lower than 15 at 3' tail. As

shown in Figure 19 the trimming allows reaching a very good quality over all cycles. About 15% of raw reads were filtered by trimming [Table 14]. The procedure discarded more reverse reads than forward ones. This due to the worst quality of reverse run compared to the forward one. This behavior is common and it is mainly related to loss of enzymatic efficiency with the proceeding of the sequencing run. Until the $40^{th}$ cycle, boxplots build on raw and trimmed data look very similar; the difference become impressive from $80^{th}$ nucleotide. This operation was applied at lane level for all samples. The assessment of data quality was also focused on the evaluation of additional metrics such as % of duplicates reads, overrepresented sequences, nucleotides distribution across reads and, finally, on the GC content across the whole length of each sequence. These metrics resulted appropriate for all samples [data not shown].



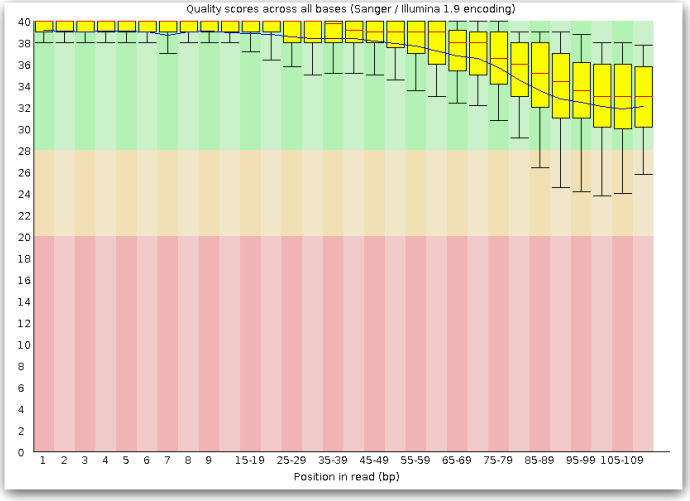




Figure 19. **Quality scores across cycles**

*The boxplot shows the base quality distribution across the read length or sequencing cycles. On the x-axis the position in the read corresponding to the n$^{th}$ cycle. The y-axis on the graph shows the quality scores in phred scale. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The top of the yellow bar identify the 75$^{th}$ quartile, whereas the bottom the 25$^{th}$. In red the median, in blue the average. A) Raw reads. B) Trimmed reads.*

| ID | Forward | Reverse | Raw Sum | Forward trimmed | Reverse trimmed | Sum trimmed | % post QC |
|---|---|---|---|---|---|---|---|
| **01a** | 6298984 | 6298984 | 12597968 | 5582294 | 5405213 | 10987507 | 87.22 |
| **01b** | 6245678 | 6245678 | 12491356 | 5424569 | 5131898 | 10556467 | 84.51 |
| **01c** | 6272058 | 6272058 | 12544116 | 5445172 | 5141276 | 10586448 | 84.39 |
| **01d** | 6215971 | 6215971 | 12431942 | 5406444 | 5100260 | 10506704 | 84.51 |
| **02a** | 6873272 | 6873272 | 13746544 | 6090082 | 5880683 | 11970765 | 87.08 |
| **02b** | 6813484 | 6813484 | 13626968 | 5928364 | 5589577 | 11517941 | 84.52 |
| **02c** | 6845934 | 6845934 | 13691868 | 5956084 | 5602601 | 11558685 | 84.42 |
| **02d** | 6787296 | 6787296 | 13574592 | 5913891 | 5552674 | 11466565 | 84.47 |
| **03a** | 5591437 | 5591437 | 11182874 | 4846211 | 4616106 | 9462317 | 84.61 |
| **03b** | 5915061 | 5915061 | 11830122 | 5097600 | 4906853 | 10004453 | 84.57 |
| **03c** | 5746450 | 5746450 | 11492900 | 5022928 | 4786576 | 9809504 | 85.35 |
| **03d** | 5682093 | 5682093 | 11364186 | 4918338 | 4715153 | 9633491 | 84.77 |
| **04a** | 6510983 | 6510983 | 13021966 | 5676865 | 5460383 | 11137248 | 85.53 |
| **04b** | 6893835 | 6893835 | 13787670 | 5982786 | 5809473 | 11792259 | 85.53 |
| **04c** | 6690856 | 6690856 | 13381712 | 5878924 | 5657736 | 11536660 | 86.21 |
| **04d** | 6605158 | 6605158 | 13210316 | 5750975 | 5567247 | 11318222 | 85.68 |
| **05a** | 6714394 | 6714394 | 13428788 | 5815957 | 5628058 | 11444015 | 85.22 |
| **05b** | 7125113 | 7125113 | 14250226 | 6139873 | 5999514 | 12139387 | 85.19 |
| **05c** | 6914936 | 6914936 | 13829872 | 6038700 | 5842232 | 11880932 | 85.91 |
| **05d** | 6827293 | 6827293 | 13654586 | 5906003 | 5752984 | 11658987 | 85.39 |
| **06a** | 6451019 | 6451019 | 12902038 | 5604187 | 5494971 | 11099158 | 86.03 |
| **06b** | 6403522 | 6403522 | 12807044 | 5451617 | 5247123 | 10698740 | 83.54 |
| **06c** | 6431008 | 6431008 | 12862016 | 5480468 | 5258877 | 10739345 | 83.50 |
| **06d** | 6375667 | 6375667 | 12751334 | 5434635 | 5223633 | 10658268 | 83.59 |
| **07a** | 6680855 | 6680855 | 13361710 | 5971620 | 5821508 | 11793128 | 88.26 |
| **07b** | 6639650 | 6639650 | 13279300 | 5835840 | 5565893 | 11401733 | 85.86 |
| **07c** | 6665204 | 6665204 | 13330408 | 5856595 | 5574653 | 11431248 | 85.75 |
| **07d** | 6625632 | 6625632 | 13251264 | 5829918 | 5540070 | 11369988 | 85.80 |
| **08a** | 6457976 | 6457976 | 12915952 | 5704888 | 5563177 | 11268065 | 87.24 |
| **08b** | 6418113 | 6418113 | 12836226 | 5554926 | 5294073 | 10848999 | 84.52 |
| **08c** | 6441913 | 6441913 | 12883826 | 5577178 | 5299877 | 10877055 | 84.42 |
| **08d** | 6388912 | 6388912 | 12777824 | 5538572 | 5257775 | 10796347 | 84.49 |
| **09a** | 7091128 | 7091128 | 14182256 | 6160652 | 5948520 | 12109172 | 85.38 |
| **09b** | 7521921 | 7521921 | 15043842 | 6506909 | 6344707 | 12851616 | 85.43 |
| **09c** | 7288332 | 7288332 | 14576664 | 6383166 | 6166581 | 12549747 | 86.09 |
| **09d** | 7194879 | 7194879 | 14389758 | 6241464 | 6071898 | 12313362 | 85.57 |
| **10a** | 6817322 | 6817322 | 13634644 | 5896515 | 5698274 | 11594789 | 85.04 |
| **10b** | 7222619 | 7222619 | 14445238 | 6222264 | 6072113 | 12294377 | 85.11 |
| **10c** | 7005020 | 7005020 | 14010040 | 6112097 | 5909014 | 12021111 | 85.80 |
| **10d** | 6925311 | 6925311 | 13850622 | 5987301 | 5825735 | 11813036 | 85.29 |

Table 14. **Reads coverage**

*In the table, the number of raw and trimmed reads for all samples in single lanes, including forward, reverse and their sum. In last column the % of reads that are available for mapping after trimming operation.*

### *4.1.2 Reads alignment*

The alignment has been performed on all samples in each lane separately. Mapped data were than submitted to data processing in GATK for ensure the best quality possible. All lanes that belonged to the same sample were merged together in a single file. The percentage of mapped reads ranged from 93.93% to 97.95%. The mapping percentage is very similar among samples, and no marked differences arise comparing the performance of cell line and patients samples. The percentage of reads duplicates is comprised from 10.32 to 17.66 with a mean of 14.74 [Table 15]. Taking into consideration the size of targeted sequence (< 500Kb), the observed value is coherent with the expected value.

| Sample | Total | Mapped | Duplicates | % mapping | % duplicates |
|---|---|---|---|---|---|
| **1 – K562** | 42637126 | 41763473 | 5961218 | 97.95 | 14.27 |
| **2** | 46513956 | 45401025 | 6444009 | 97.61 | 14.19 |
| **3** | 38909765 | 37660181 | 6235252 | 96.79 | 16.56 |
| **4** | 45784389 | 44297334 | 6386250 | 96.75 | 14.42 |
| **5** | 47123321 | 45904750 | 6275431 | 97.41 | 13.67 |
| **6** | 43195511 | 40575236 | 7164256 | 93.93 | 17.66 |
| **7** | 45996097 | 44857207 | 7125263 | 97.52 | 15.88 |
| **8** | 43790466 | 42565090 | 6534884 | 97.2 | 15.35 |
| **9** | 49823897 | 47932582 | 4947932 | 96.2 | 10.32 |
| **10** | 47723313 | 46007733 | 6956108 | 96.41 | 15.12 |

Table 15. **Mapping Statistics**
*Starting from left to right column: samples, total and mapped reads, number of duplicate reads, percentage of mapped and duplicated reads.*

### *4.1.3 Capturing performance and target enrichment*

The performance of the target enrichment experiment and the coverage distribution have been assessed both at lane and sample level, for each region separately. The target regions boundaries and features are summarized in Table 16.

Samples were multiplexed, pooled into two pools and loaded in 4 lanes. By using DepthOfCoverage module in GATK we assessed the mean coverage and the percentage of target capturing at a specific coverage depth. Mean coverage indicate the average number of reads covering each single base within the target. Higher is the coverage, higher will be the chance to detect SV, INDELs and SNV, along with a higher reliability of the detection itself. The average mean coverage reached at lane level across patient samples (from 2 to 10) was higher in BCR-major region (152.24 ± 21.92), similar for micro and minor region (130.94 ± 14.54; 133.96 ± 13.65) and lower in ABL1 (101.59 ± 12.50). The same trend was observed at sample level Table 16. The mean coverage at sample level does not correspond to the value at lane level multiplied for a factor of four because, once lanes are merged, the duplicates were removed.

| Region | Average Mean Coverage (Lane level) | Std.dev | Average Mean Coverage (Sample Level) | Std.dev |
|---|---|---|---|---|
| **Abl1** | 101.59 | 12.50 | 319.61 | 58.32 |
| **M-BCR** | 152.24 | 21.92 | 486.16 | 79.99 |
| **m-BCR** | 130.94 | 14.54 | 435.47 | 56.00 |
| **µ-BCR** | 133.96 | 13.65 | 415.62 | 64.14 |

Table 16. **Mean coverage in target regions**
*In rows the four targeted regions, in columns the average of mean coverage across patients samples, both at lane level (2nd column) and at sample level with merged lanes (4th column). In 3rd and 5th column the corresponding standard deviation for 2nd and 4th column respectively.*

Differently by patient samples, K562 cell line are covered by 2474.28 reads in ABL1, 3081.03, 1191.95 and 373.68 reads in m-BCR, M-BCR and µ-BCR region respectively [Data not shown]. The K562 cell line has been covered ~ 7.75, 7, 2.45 and 0.9 times more than the sample average in ABL1, m-BCR, M-BCR and µ-BCR region respectively. The percentage of capturing, at X coverage, is expressed as the percentage of the targeted region that is covered by at least X reads. Higher is the percentage, greater will be the uniformity of coverage at a given threshold of read depth.

The percentage has been calculated at four different thresholds, 1, 20, 50 and 100X, for each target region. In Table 17 and Table 27, Table 28, Table 29 and Table 30, the percentages of enrichment both at sample and lane level for each target region are shown. At sample level, data capturing at 1X show that ABL1 and m-BCR are not completely captured by the probes, whereas major and micro BCR target regions do not have uncovered portions Table 17. Contrariwise, the data related to the capturing at 100X reveal that BCR major and micro regions are almost completely covered (except for patient 10 with a 92% of capturing). These results are modeled both on the regions' length and on the percentage of repeated sequences included in the region. Indeed BCR-minor and ABL1 show a higher percentage compare to µ-BCR and M-BCR as well as a bigger size. The presence of many repeated regions affects the accuracy of mapping increasing the unspecific mapping through all the genome. Taking into consideration that ABL1 and m-BCR region account for 93% of the whole target size, the unspecific mapping is an issue that concern the whole experiment.

The mean on-target mapping percentage at lane level is around 2.5%. This means that 97.5% of reads map over the entire genome. This phenomenon is attributable to the small size of the target region along with the extreme high percentage of repeated sequences. In K562 cell line, this value is around 18.5 % (data not shown).

| Sample | Target | 1X | 50X | 100X |
|---|---|---|---|---|
| 1 | ABL1 | 99.9 | 96.3 | 94.3 |
| 2 | ABL1 | 99.6 | 92.9 | 87.4 |
| 3 | ABL1 | 99.5 | 91.3 | 83.6 |
| 4 | ABL1 | 99.3 | 93 | 87.6 |
| 5 | ABL1 | 99.3 | 91 | 81.6 |
| 6 | ABL1 | 99.6 | 90.3 | 75.7 |
| 7 | ABL1 | 99.2 | 92.1 | 86.1 |
| 8 | ABL1 | 99.5 | 91.3 | 84.8 |
| 9 | ABL1 | 99.4 | 92.9 | 87 |
| 10 | ABL1 | 99.7 | 92.2 | 84.9 |

| Sample | Target | 1X | 50X | 100X |
|---|---|---|---|---|
| 1 | M-BCR | 100 | 100 | 99.4 |
| 2 | M-BCR | 100 | 100 | 99.5 |
| 3 | M-BCR | 100 | 100 | 99.4 |
| 4 | M-BCR | 100 | 100 | 99.9 |
| 5 | M-BCR | 100 | 100 | 100 |
| 6 | M-BCR | 100 | 99.8 | 98.8 |
| 7 | M-BCR | 100 | 100 | 99.6 |
| 8 | M-BCR | 100 | 100 | 99.9 |
| 9 | M-BCR | 100 | 100 | 99.9 |
| 10 | M-BCR | 100 | 100 | 99.7 |

| Sample | Target | 1X | 50X | 100X |
|---|---|---|---|---|
| 1 | µ-BCR | 100 | 95.6 | 90.7 |
| 2 | µ-BCR | 100 | 100 | 98.5 |
| 3 | µ-BCR | 100 | 100 | 97.5 |
| 4 | µ-BCR | 100 | 100 | 97.8 |
| 5 | µ-BCR | 100 | 100 | 98.3 |
| 6 | µ-BCR | 100 | 100 | 98.1 |
| 7 | µ-BCR | 100 | 100 | 97.5 |
| 8 | µ-BCR | 100 | 100 | 98.3 |
| 9 | µ-BCR | 100 | 99.5 | 96.3 |
| 10 | µ-BCR | 100 | 95.7 | 92 |

| Sample | Target | 1X | 50X | 100X |
|---|---|---|---|---|
| 1 | m-BCR | 100 | 97.1 | 96.3 |
| 2 | m-BCR | 99.5 | 93.4 | 90.9 |
| 3 | m-BCR | 97.8 | 91.5 | 88.9 |
| 4 | m-BCR | 98.5 | 91.6 | 89.3 |
| 5 | m-BCR | 97.8 | 91.7 | 89.4 |
| 6 | m-BCR | 99.5 | 92.3 | 87.2 |
| 7 | m-BCR | 98.9 | 92.8 | 90 |
| 8 | m-BCR | 99.8 | 92.6 | 90.1 |
| 9 | m-BCR | 99.5 | 92.7 | 89.9 |
| 10 | m-BCR | 99.4 | 92.7 | 90.2 |

Table 17. **Sample-level % of capturing at X coverage**
*In the sub-tables, the percentage of capturing of the different regions at 1X, 50X and 100X coverage.*

The analysis at lane level shows an overall decrease of the percentage of capturing at 50 and 100X for all patients and regions, except for M-BCR where the capturing at 50X is near 100% in all samples [Figure 21]. The lane level capturing at 100X is much lower compared the one obtained at sample level, especially in ABL1 region and in the lanes belonging to sample 5 and 6 [Figure 20]. This data demonstrated that is not possible reaching a good percentage of capturing at high coverage in all regions using a single lane. However, considering the theoretical percentage of capturing calculated during probe design, the targeting efficiency is very

good. Indeed, taking as reference the percentage of capturing at 50X, the experimental capturing is more than the theoretical one, passing from 63% to 91% for ABL1, from 93% to 100% for M-BCR, from 74% to 92% and from 96% to 100% for µ-BCR. This phenomenon can be explained by the fact that probes designed in portions of repeated regions (20-40 nucleotides overlapping) can capture nonspecific regions and this can lead to decrease of capturing percentage.

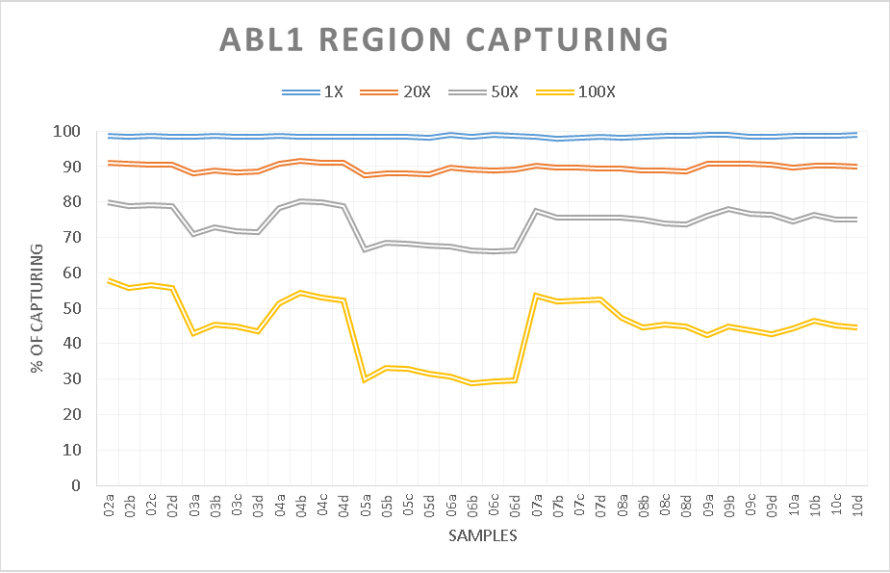



Figure 20. **ABL1 region capturing**
*On the x-axis, the samples divided in lanes. On y-axis the percentage of capturing of the ABL1 region at different depth of coverage: blue=1x, orange=20x, grey=50x and yellow=100x.This chart has been produce from data in Table 27.*

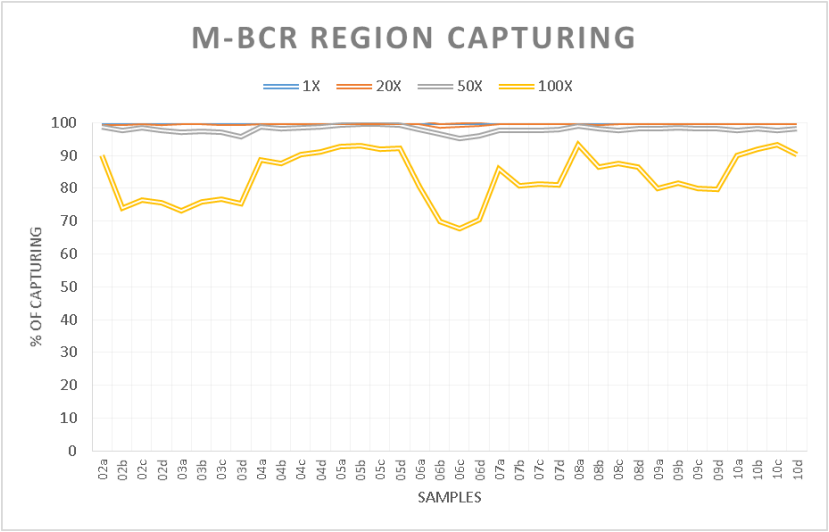



Figure 21. **M-BCR region capturing**

*On the x-axis, the samples divided in lanes. On y-axis the percentage of capturing of the M-BCR region at different depth of coverage: blue=1x, orange=20x, grey = 50x and yellow=100x.This chart has been produce from data in* **Table 28**.

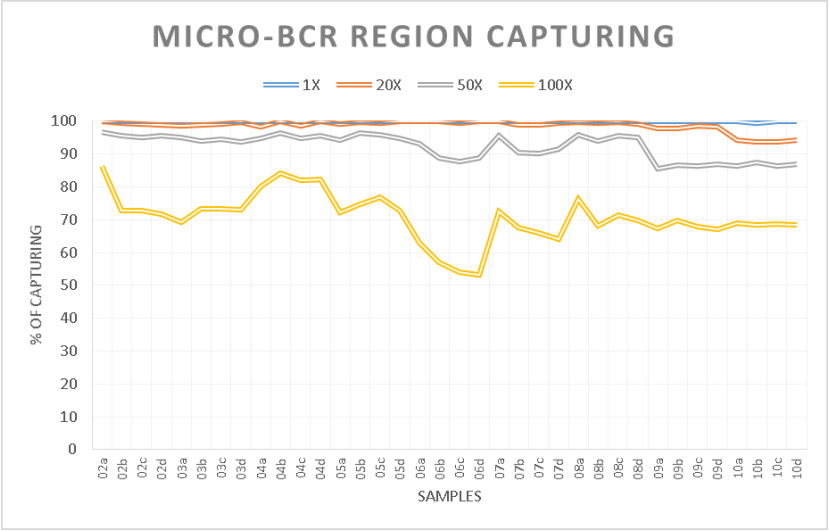



Figure 22. **Micro-BCR region capturing**

*On the x-axis, the samples divided in lanes. On y-axis the percentage of capturing of the ABL1 region at different depth of coverage: blue = 1x, orange = 20x, grey = 50x and yellow = 100x.This chart has been produce from data in* **Table 30**.

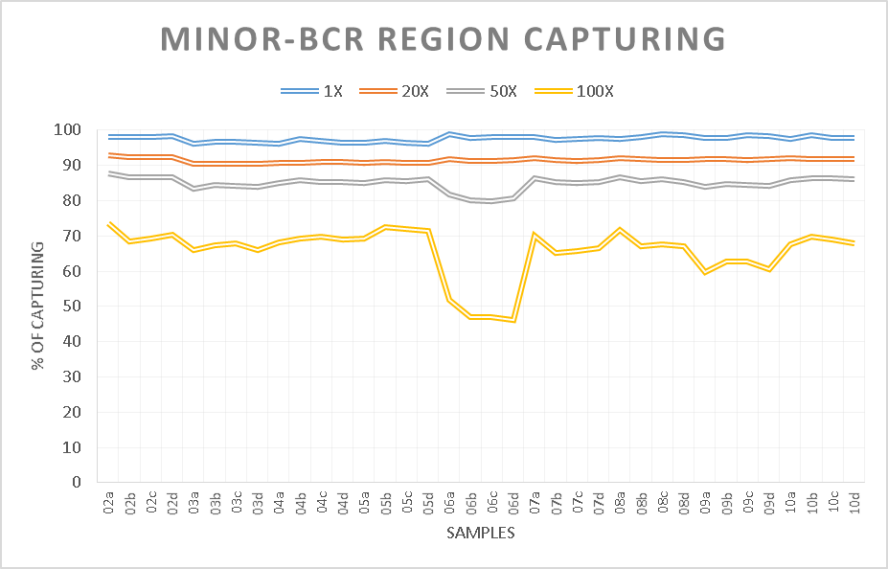



Figure 23. **m-BCR region capturing**

*On the x-axis, the samples divided in lanes. On y-axis the percentage of capturing of the ABL1 region at different depth of coverage: blue = 1x, orange = 20x, grey = 50x and yellow = 100x.This chart has been produce from data in* **Table 29**.

## 4.2 BCR-ABL1 breakpoints identification

In this chapter will be discussed the results of the identification of breakpoints that lead to the formation of BCR-ABL1 fusion oncogene. Results obtained by discordant pair approach and split-reads approach will be discussed separately. Afterwards results obtained by Sanger validation will be presented. Moreover, a comparison of breakpoint identification performance among the different methods will be discussed. Finally, a patient-specific analysis will be conducted in order to highlight genomic features and complexity of the breakpoint regions. The description will include all samples analyzed.

### *4.2.1 Breakpoints detection with SVDetect*

The detection of breakpoints was performed with SVDetect by using a discordant-pair approach. In Table 19 are described all breakpoints that identify the BCR-ABL1 oncogene and the ABL1-BCR reciprocal fusion product, if present.

In all samples two breakpoint coordinates, both in BCR and ABL1 regions, were detected, except for the K562 cell line and samples 2,7,8,9 Table 19. We have tagged the breakpoint that support the BCR-ABL1 fusion as BP1 (mapped to BCR) and BP4 (mapped to ABL1) whereas BP2 (located at BCR) and BP3 (mapped to ABL1) support the ABL1-BCR fusion.

Coordinates that define the fusion of BCR with ABL1 to form the BCR-ABL1 oncogene were identified by discordant pairs in which the forward read mapped in BCR identifying BP1 whereas its reverse mate mapped in ABL1 identifying the BP4. On the contrary, discordant pairs in which the forward read mapped to ABL1, defining the BP3, and its reverse mate mapped in BCR, spotting the BP2, detected the ABL1-BCR fusion segment. BP1 and BP4 were identified in all patients except for patient 8, whereas BP2 and BP3 were detected in six patient samples (3, 4, 5, 6, 8 and 10). By the analysis of SVDetect top results, BCR-ABL1 fusion gene seems to be

formed by the t(9;22) translocation in all samples, except for samples 4 and 5 whose BP1 breakpoints fall in m-BCR and µ-BCR region respectively. In these two samples seemed that BCR-ABL1 fusion arise from the insertion of a fragment of BCR into gene ABL1 without Ph formation. However, the refinement analysis with ClipCrop have shown that in both samples the break of chromosome 9 and 22 was blunt-like with a partial insertion of few bases at the junction point both in Ph and der (9). A deeper analysis of such breakpoints will be performed in the next sample specific analysis.

In patient 7 and 9 inverted translocations concerning ABL1-BCR fusion were detected. In particular, in patient 7 the region downstream to the ABL1 breakpoint of the BCR-ABL1 fusion segment was found to be involved in an inverted translocation with a sequence located at the following genomic interval chr9: 36900847-36901054 mapping to the FOXRED2 gene. Patient 9 showed an inverted translocation detected by discordant pair reads mapping in a region near telomere (chr9:136985521-136985696) and upstream (14 bp) to BP1. These findings will be described in detail in the sample specific analysis.

The number of pair-end reads that support the BCR-ABL1 breakpoints was variable, ranging from 38 to 348 with a mean of 174.25 ± 95.83. At first glance, the distribution of supporting reads across samples seems to correlate with the overall number of mapped reads, but this phenomenon could be associated to the concentration of $Ph^+$ cells and hence by the clinical phase of CML. Although all specimen analyzed were taken at the disease onset, no info were available to determine the disease progression of each sample. Otherwise, this phenomenon could be explained simply by the chance to capture a less number of $Ph^+$ cells from the DNA aliquot used in analyses.

### *4.2.2 IGV analysis*

In order to confirm the breakpoints coordinates identified in SVDetect, a simple quality control was performed using the Integrative Genome Browser (IGV). The confirmation of BCR-ABL1 breakpoints was carried on looking for paired-end reads in which the forward read map in 5' to 3' direction BCR and the reverse read mapped in the opposite direction in the ABL1 gene. On the contrary, ABL1-BCR fusion points were visually inspected by looking for paired-end reads in which the forward read mapped in ABL1 with 5' to 3' direction and its reverse pair mapped in the opposite direction to the BCR region. In IGV, reads whose mate pair maps to another chromosome are painted with different colors according to the specific chromosome. In particular, reads whose mate maps to the chromosome 9 are "fluorescent green", whereas reads whose mate pair maps to the chromosome 22 are "dark grey" [Figure 24]. The visual check for the breakpoints coordinates that lead to the formation of BCR-ABL1 are summarized in Table 18.

| Sample | chr9 (ABL) | chr22 (BCR) |
|---|---|---|
| **K562** | 133607140 | 23632762 |
| **2** | 133648882 | 23631930 |
| **3** | 133593828 | 23632124 |
| **4** | 133590571 | 23575268 |
| **5** | 133708492 | 23654742 |
| **6** | 133658098 | 23634051 |
| **7** | 133684300 | 23634602 |
| **8** | ---------- | ---------- |
| **9** | 133663988 | 23631897 |
| **10** | 133696999 | 23634590 |

Table 18. **IGV visual check**
*BP1 and BP4 coordinates identified by visual checking in IGV.*

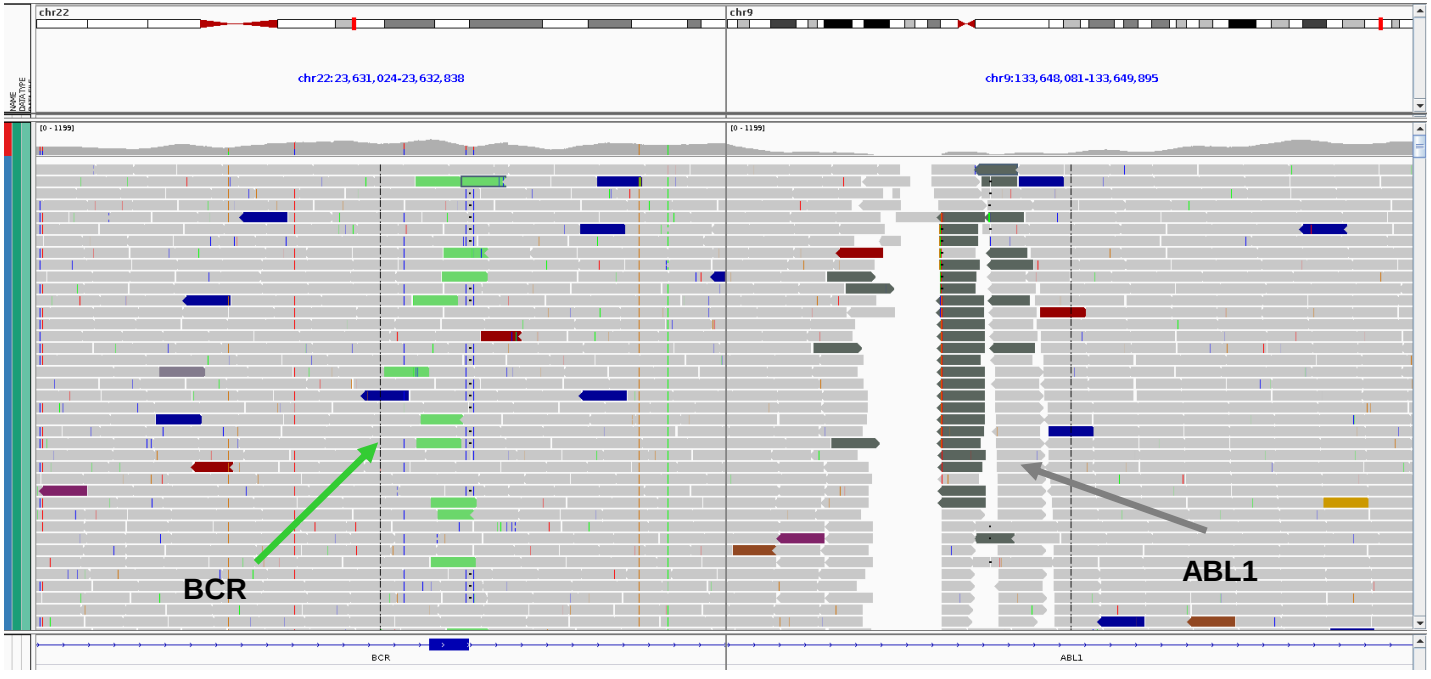



Figure 24. **BCR-ABL1 breakpoints identified by paired-end reads**

*Reads pointed with green arrow represent reads (light green) in forward orientation (5' to 3'), located in BCR gene, which have a mate pair (dark grey) in reverse orientation (3' to 5') in ABL1 gene. Light grey reads identify reads, which do not have a mate pair that map in other chromosome or to a distance that is different from the expected insert size. Reads of other colors identify the ones which have a mate pair that map in other chromosome different from 22 and 9 (each different chromosome has its own color).*

| Sample | Chr type | SV type | BAL type | N pairs | Breakpoint chr9 start-end | Breakpoint chr22 start-end |
|---|---|---|---|---|---|---|
| **K562** | INTER | Transloc | UNBAL | 1059 | 133606963-**133607140** | **23632762**-23632844 |
| **2** | INTER | Transloc | UNBAL | 146 | 133648580-**133648884** | **23631930**-23632065 |
| **3** | INTER | Transloc | UNBAL | 38 | 133593517-**133593826** | **23632124**-23632341 |
| | INTER | Transloc | UNBAL | 32 | **133593826**-133594081 | 23633315-**23633550** |
| **4** | INTER | Ins_fragmt | BAL | 348 | 133590346-**133590561**<br>133590582-133590771 | **23575268**-23575206 |
| | INTER | Ins_fragmt | BAL | 338 | 133590346-133590561<br>**133590576**-133590771 | 23575268-**23575206** |
| **5** | INTER | Ins_fragmt | BAL | 91 | 133708269-133708492<br>**133708509**-133708819 | 23654742-**23654774** |
| | INTER | Ins_fragmt | BAL | 86 | 133708269-**133708492**<br>133708502-133708819 | **23654742**-23654774 |
| **6** | INTER | Transloc | UNBAL | 227 | 133657760-**133658098** | **23634051**-23634167 |
| | INTER | Transloc | UNBAL | 42 | **133658314**-133658578 | 23634263-**23634637** |
| **7** | INTER | Transloc | UNBAL | 213 | 133684073-**133684300** | **23634602**-23634800 |
| | INTER | Inv_transloc | UNBAL | 185 | 133684252-133684469 | 36900847-36901054 |
| **8** | INTER | Transloc | UNBAL | 41 | **133694955**-133695297 | 23632677-**23632905** |
| **9** | INTER | Transloc | UNBAL | 132 | 133663848-**133663988** | **23631897**-23631966 |
| | INTER | Inv_transloc | UNBAL | 114 | 136985521-136985696 | 23631690-23631884 |
| **10** | INTER | Transloc | BAL | 204 | 133696783-**133696999** | **23634590**-23634635 |
| | INTER | Transloc | BAL | 98 | **133696783**-133696999 | 23634590-**23634658** |

Table 19. **SVDetect breakpoints detection**

*The table shows the top SV supported by PE reads in which one read maps to BCR and its mate maps to ABL1. The column describe if the SV is Inter / Intra chromosomal, the type of SV (translocation, inverted translocation, insertion fragment), the balancing, the number of pairs supporting the SV call, and the breakpoints coordinates at chr9 and chr22. The start coordinates and the end coordinates describe the most 5' and 3' coordinates of the reads included in the group of reads that support the breakpoint. In red BP1 and BP4 whereas in green BP2 and BP3.*

### *4.2.3 Breakpoints refinement with ClipCrop*

As previously explained in section 3.2.4.4.2, ClipCrop was used to refine the breakpoints coordinates identified by SVDetect. From now to onwards we will refer to ClipCrop refinement including the re-mapping procedure using the aligner BLAT. The results are listed in [Table 20] and [Table 21]. All breakpoints identified by SVDetect were refined by ClipCrop according to the clipping orientation of splitted reads. The refinement was achieved by setting the parameter “min sv cluster size” equal to 8, that allowed to spot all breakpoints identified by SVDetect. By setting the “min sv cluster size” to 20 it lead to a failure in the detection of BP4 in samples 3, BP3 in sample 5 and 8 and BP2 in sample 10. The consistency of cut-orientation with breakpoints coordinates discovered by SVDetect was perfect. As expected, the breakpoints that lead to the formation of BCR-ABL1 fusion gene are identified by R-breakpoints in BCR and L-breakpoints in ABL1 Table 20, whereas the breakpoints that give rise to ABL1-BCR counterpart are spotted by R-breakpoints in ABL1 and L-breakpoints in BCR Table 21. According to SVDetect, BP1 and BP4 were found in all patients except for patient 8.

| Patient | Chromosome | Start | End | Cut-orientation | Reads |
|---|---|---|---|---|---|
| 2 | 22 | 23631911 | 23631912 | R | 91 |
| 2 | 9 | 133648890 | 133648891 | L | 41 |
| 3 | 22 | 23632128 | 23632129 | R | 51 |
| 3 | 9 | 133593827 | 133593828 | L | 13 |
| 4 | 22 | 23575251 | 23575252 | R | 168 |
| 4 | 9 | 133590570 | 133590571 | L | 155 |
| 5 | 22 | 23654765 | 23654766 | R | 202 |
| 5 | 9 | 133708495 | 133708496 | L | 150 |
| 6 | 22 | 23634033 | 23634034 | R | 180 |
| 6 | 9 | 133658107 | 133658108 | L | 23 |
| 7 | 22 | 23634578 | 23634579 | R | 177 |
| 7 | 9 | 133684306 | 133684307 | L | 166 |
| 9 | 22 | 23631875 | 23631876 | R | 143 |
| 9 | 9 | 133663994 | 133663995 | L | 195 |
| 10 | 22 | 23634573 | 23634574 | R | 111 |
| 10 | 9 | 133697000 | 133697001 | L | 8 |

Table 20. **ClipCrop: BCR-ABL1 fusion points**
*The table lists BP1 and BP4 identified by ClipCrop according to clipped direction in the 400bp surrounding the BPs coordinates identified by SVDetect. The last column shows the number of splitted reads supporting the breakpoint.*

| Patient | Chromosome | Start | End | Cut-orientation | Reads |
|---|---|---|---|---|---|
| 3 | 22 | 23633556 | 23633557 | L | 163 |
| 3 | 9 | 133593804 | 133593805 | R | 35 |
| 4 | 22 | 23575206 | 23575207 | L | 82 |
| 4 | 9 | 133590569 | 133590570 | R | 76 |
| 5 | 22 | 23654766 | 23654767 | L | 58 |
| 5 | 9 | 133708499 | 133708500 | R | 10 |
| 6 | 22 | 23634639 | 23634640 | L | 53 |
| 6 | 9 | 133658310 | 133658311 | R | 53 |
| 8 | 22 | 23632908 | 23632909 | L | 37 |
| 8 | 9 | 133694989 | 133694990 | R | 13 |
| 10 | 22 | 23634640 | 23634641 | L | 100 |
| 10 | 9 | 133696780 | 133696781 | R | 32 |

Table 21. **ClipCrop: ABL1-BCR fusion points**
*The table lists BP2 and BP3 identified by ClipCrop according to clipped direction in the 400bp surrounding the BPs coordinates identified by SVDetect. The last column shows the number of splitted reads supporting the breakpoint.*

### 4.2.4 *Breakpoints* validation with Sanger Sequencing

This chapter will deal with the results obtained from the analysis of chromatograms obtained by the validation of the breakpoints using Sanger method. In particular, an illustrative analysis of chromatograms and their relative nucleotide sequence will be conducted on sample 6.

#### *4.2.4.1 Chromatogram Analysis*

The validation of BCR-ABL1 breakpoints coordinates was performed using Sanger sequencing as alternative sequencing method. All the chromatograms were visually inspected with Chromas [Figure 25] and associated FASTA sequences were extracted. For each sample two chromatograms were provided, one for the forward primer (FW) and one for the reverse primer (RV). The visual inspection was performed manually in order to ensure the reliability of the findings. Once FASTA sequences were obtained, Blat and Blast alignment were performed for all files. FASTA sequences obtained with FW and RV [Figure 26] were aligned to human genome with BLAT and BLAST using as reference the GRCh37p13 Primary Assembly. The sequenced fragments obtained by Sanger method had different length depending on the primers position. For instance, in sample 6, the sequence obtained by FW was 795bp long whereas the fragment got by RV was 802bp in length.

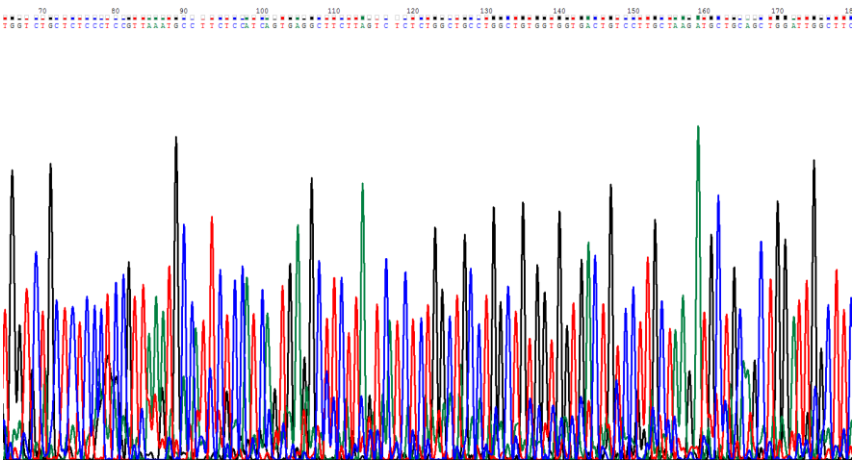


Figure 25. **Illustrative chromatogram from Sanger sequencing**
*Below, peaks color represent different nucleotides: C(Blue), G(Black), A(Green), T(Red). On the top, the position along the sequenced fragment and the signal intensity related to each base position. Higher are the peaks, higher is the signal intensity and higher the reliability of the basecall at that position.*

```
AGAATCGTAGGAGCTCACGTCAATTTAGGGGCACCACCATCCCCCGCATGGCCAGCCGAA
ACCGTGGTCTGCTCTCCCTCCGTTAAATGCCTTCTCCATCAGTGAGGCTTCTTAGTCTCT
CTGGCTGCCTGGCTGTGGTGGTGACTGTCCTTGCTAAGATGCTGCAGCTGGATTGGCTTC
GTGGGAAGTTGTGGAATTCTGTTCCCTGAGGATGTTGGGAAGGACAGATTGTACTTACCA
AGAGCATTTATTGTTGTTTTTATTTTCTTAAACATTTCCAATGAAAATCATCAAACCTAT
AAAGTTGAAAGAACAATGCACCTAGATTCATCAACTCTTAACACTTGCTATCTCTCTTAC
ATAGAAACATTTCCCTTTTACCAGTCCACTTGGAAATTGGCTATATACTCACGATAGCCA
AGAACAAGAACACTCTCTTACACCATCACATCATTGCCACACTTAAGGAAATCCACACTG
ATCCAGGAACATTATTTGATATACAGTCCATATTCAGAATCCTCCAGGTATCAGTAATAT
TTTAATTATGACCCATTTGGTGAATTTAAGATGGATTAGATGGCCTGTGGAGTTATTCAT
GGAGCTGAAGACTTCTGTGAAGTAAAAAATAACCACCCAGTGTTAAGTGTAACTCTCGTG
TGGAATATTAAAATCAATTGTGTGTTTTTTTTGGCATCCCTTGATGTGAACTCTGGAGAC
CTGATGAGATGGAGGCTGTGATTGTTTACTCTGTGTTAGCTTGGACACCTTAGCTTGGCA
GTAATCACAAGAACTTTAAGGAGTGCATCCCTTTGCTAGTTCATAAAGAATGTATTCTGA
TAACGGGGGACCAAGAGGATATATCCCGAATTGCAAATCGAGCAGTAATTACATT
```

```
AGACCTGCCCTTGTCATATCATATTTTGGTCCCCGTTATCGAATACATCTTTATGAACTA
ACAAAGGATGCACTCCCTAAAGCTCTTGTGATTACTGCCAAGCTAAATGTCCAAGCTAAC
ACAGAGTAAACAATGACAGCCTCCATCTCATCAGGTCTCCAGAGTTCACATCAAGGATGC
CAAAAAAAACACACAATTGATTTAATATTCCACACGAGAGTTACACTTAACACTGGGTGG
TTATTTTTTACTTCACAGAAGTCTTCAGCTCCATGAATAACTCCACAGGCCATCTAATCC
ATCTTAAATTCACCAAATGGGTCATAATTAAAATATTACTGATACCTGGAGGATTCTGAA
TATGGACTGTATATCAAATAATGTTCCTGGATCAGTGTGGATTTCCTTAAGTGTGGCAAT
GATGTGATGGTGTAAGAGAGTGTTCTTGTTCTTGGCTATCGTGAGTATATAGCCAATTTC
CAAGTGGACTGGTAAAAGGGAAATGTTTCTATGTAAGAGAGATAGCAAGTGTTAAGAGTT
GATGAATCTAGGTGCATTGTTCTTTCAACTTTATAGGTTTGATGATTTTCATTGGAAATG
TTTAAGAAAAATAAAAACAACAATAAATGCTCTTGGTAAGTACAATCTGTCCTTCCCAAC
ATCCTCAGGGAACAGAATTCCACAACTTCCCACGAAGCCAATCCAGCTGCAGCATCTTAG
CAAGGACAGTCACCACCACAGCCAGGCAGCCAGAGATGACTAAGAAAGCCTCACTGATGG
AGAATGGCATTTAACGGAGGAGAGCAGACCACGGTTTCTGGCTTGCCATGCGGTTGGATG
GTGGTGCCCCTCAAGGGACGGACTAAGAGGAACAAGTTTGGGGTCCATGATACATGAAAA
CG
```

Figure 26. **Sanger FASTA sequence (Patient 6)**
*FASTA sequences obtained from chromatograms of fusion fragments obtained using forward (top) and reverse (bottom) primers.*

#### *4.2.4.2 BLAT and BLAST alignments*

The BLAT alignment showed that both segments (obtained by FW and RV) spanned the BCR-ABL1 breakpoint. Indeed, by using the FW primer designed at the 5' of the BCR breakpoint, the fragment aligned to BCR from the 27$^{th}$ nucleotides to the 133$^{th}$ nucleotides (23633922-23634033) with a 100% of identity whereas it aligned to ABL1 from the 124$^{th}$ nucleotides to the 895$^{th}$ (133658102-133658868) with an identity of 98.1%. By using the RV primer designed at the 3' of the ABL1 breakpoint, the fragment aligned from the 24$^{th}$ nucleotides to the 750$^{th}$ nucleotides to ABL1 (133658102-133658827) with a 99.1% of identity whereas it aligned to BCR from the 741$^{th}$ nucleotides to the 890$^{th}$ with an identity of 98.7% [Figure 27]. The mapping coordinates are consistent among each other's, indeed, the breakpoint coordinate of BCR-ABL1 fusion gene are the same using both FW and RV primers to BCR-ABL1 amplicon. Hence, the breakpoint coordinates for BCR-ABL1 are 23634033 on chromosome 22 and 133658102 on chromosome 9. Moreover, it emerged that at the junction point BCR and ABL1 share 9 nucleotides (CTGCCTGGC). The same findings, including breakpoints coordinates and shared nucleotides were verified by using BLAST alignment [Figure 28]. In Figure 29 and Figure 30 the BLAST alignments of BCR and ABL1 portions of the fusion segment obtained by FW are showed. The matching is not perfect, due to the presence of insertions/deletions as well as SNVs. For instance, the BCR portion has a 100% of matching in BLAT and 95% in BLAST report. This behavior is due to the presence of 5 gaps that are deal by two aligners in different way. Regarding the ABL1 portion the identity in BLAST report is 98% due to the presence of 9 gaps including 7 deletions and 2 insertions both of 1bp and 5 SNVs.

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 749 | 124 | 895 | 895 | 98.1% | 9 | + | 133658102 | 133658868 | 767 |
| browser details | YourSeq | 102 | 27 | 133 | 895 | 100.0% | 22 | + | 23633922 | 23634033 | 112 |
| browser details | YourSeq | 39 | 121 | 214 | 895 | 97.7% | 3 | + | 40570639 | 40751623 | 180985 |
| browser details | YourSeq | 35 | 479 | 518 | 895 | 94.9% | 3 | + | 30403945 | 30704766 | 300822 |
| browser details | YourSeq | 21 | 494 | 516 | 895 | 95.7% | 3 | - | 64021524 | 64021546 | 23 |
| browser details | YourSeq | 21 | 257 | 277 | 895 | 100.0% | 1 | - | 97340106 | 97340126 | 21 |
| browser details | YourSeq | 21 | 265 | 285 | 895 | 100.0% | 4 | + | 79249171 | 79249191 | 21 |
| browser details | YourSeq | 21 | 496 | 516 | 895 | 100.0% | 2 | + | 201765878 | 201765898 | 21 |

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 714 | 24 | 750 | 902 | 99.1% | 9 | - | 133658102 | 133658827 | 726 |
| browser details | YourSeq | 141 | 741 | 890 | 902 | 98.7% | 22 | - | 23633880 | 23634033 | 154 |
| browser details | YourSeq | 39 | 660 | 753 | 902 | 97.7% | 3 | - | 40570639 | 40751623 | 180985 |
| browser details | YourSeq | 35 | 355 | 394 | 902 | 94.9% | 3 | - | 30403945 | 30704766 | 300822 |
| browser details | YourSeq | 23 | 72 | 97 | 902 | 83.4% | 8 | - | 131338114 | 131338137 | 24 |
| browser details | YourSeq | 21 | 588 | 608 | 902 | 100.0% | 4 | - | 79249171 | 79249191 | 21 |
| browser details | YourSeq | 21 | 357 | 377 | 902 | 100.0% | 2 | - | 201765878 | 201765898 | 21 |
| browser details | YourSeq | 21 | 357 | 379 | 902 | 95.7% | 3 | + | 64021524 | 64021546 | 23 |
| browser details | YourSeq | 20 | 433 | 454 | 902 | 95.5% | 1 | + | 228820601 | 228820622 | 22 |

Figure 27. **BLAT on fusion segment obtained by FW and RV primers**
*The figure shows the BLAT alignment report. In the 4th and 5th column the start and end positions of segments that maps at start (10th column) and end (11th column) coordinates on Hg19. In the 6th column the query size that is the length of the Sanger sequenced segment. The 7th column describe the percentage of identity without considering gaps. In the 8th and 9th column the chromosome and the strand (+/-). The last column describe how many bases of the query sequence span and match the Hg19 reference genome.*

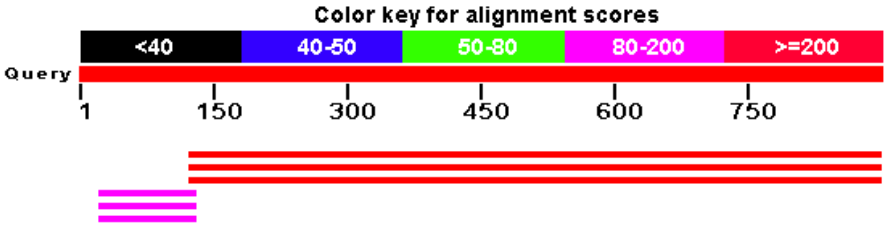
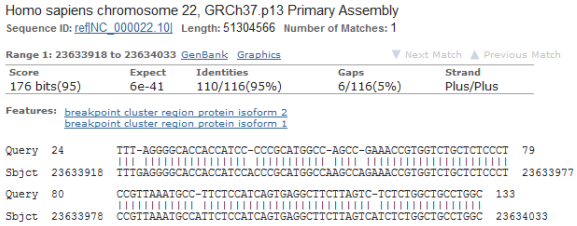



Figure 28. **Blast overview: BCR-ABL1 fusion**
*In red the ABL1 portion, whereas in violet the BCR portion. Note the overlapping near the junction point.*



Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly
Sequence ID: ref|NC_000022.10| Length: 51304566 Number of Matches: 1

Range 1: 23633918 to 23634033 GenBank Graphics Next Match Previous Match

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 176 bits(95) | 6e-41 | 110/116(95%) | 6/116(5%) | Plus/Plus |

Features: breakpoint cluster region protein isoform 2
breakpoint cluster region protein isoform 1

```
Query  24        TTT-AGGGGCACCACCATCC-CCCGCATGGCC-AGCC-GAAACCGTGGTCTGCTCTCCCT  79
                 ||| ||||||||||||||| |||||||||| |||| ||||||||||||||||||||||
Sbjct  23633918  TTTGAGGGGCACCACCATCCACCCGCATGGCCAAGCCAGAAACCGTGGTCTGCTCTCCCT  23633977

Query  80        CCGTTAAATGCC-TTCTCCATCAGTGAGGCTTCTTAGTC-TCTCTGGCTGCCTGGC  133
                 |||||||||||| |||||||||||||||||||||||| |||||||||||||||||
Sbjct  23633978  CCGTTAAATGCCATTCTCCATCAGTGAGGCTTCTTAGTCATCTCTGGCTGCCTGGC  23634033
```

Figure 29. **BLAST alignment: BCR portion**

*The figure shows the alignment report regarding the BCR portion of the BCR-ABL1 fusion segment. The range of the alignment on the NCBI reference genome (GRCh37.p13) is 23633918-23634033 (chr22). The identity is 110 out of 116 bp (95%), all unmatched bases are due to deletions in the query sequence.*

Homo sapiens chromosome 9, GRCh37.p13 Primary Assembly
Sequence ID: ref|NC_000009.11| Length: 141213431 Number of Matches: 1

Range 1: 133658102 to 133658868 GenBank Graphics Next Match Previous Match

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 1349 bits(730) | 0.0 | 761/774(98%) | 9/774(1%) | Plus/Plus |

Features: tyrosine-protein kinase ABL1 isoform b

```
Query  124        GCTGCCTGGCTGTGGTGGTGACTGTCCTTGCTAAGATGCTGCAGCTGGATTGGCTTCGTG  183
                   ||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658102  GCTGCC--GCTGTGGTGGTGACTGTCCTTGCTAAGATGCTGCAGCTGGATTGGCTTCGTG  133658159

Query  184        GGAAGTTGTGGAATTCTGTTCCCTGAGGATGTTGGGAAGGACAGATTGTACTTACCAAGA  243
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658160  GGAAGTTGTGGAATTCTGTTCCCTGAGGATGTTGGGAAGGACAGATTGTACTTACCAAGA  133658219

Query  244        GCATTTATTGTTGTTTTTATTTTCTTAAACATTTCCAATGAAAATCATCAAACCTATAAA  303
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658220  GCATTTATTGTTGTTTTTATTTTCTTAAACATTTCCAATGAAAATCATCAAACCTATAAA  133658279

Query  304        GTTGAAAGAACAATGCACCTAGATTCATCAACTCTTAACACTTGCTATCTCTCTTACATA  363
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658280  GTTGAAAGAACAATGCACCTAGATTCATCAACTCTTAACACTTGCTATCTCTCTTACATA  133658339

Query  364        GAAACATTTCCCTTTTACCAGTCCACTTGGAAATTGGCTATATACTCACGATAGCCAAGA  423
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658340  GAAACATTTCCCTTTTACCAGTCCACTTGGAAATTGGCTATATACTCACGATAGCCAAGA  133658399

Query  424        ACAAGAACACTCTCTTACACCATCACATCATTGCCACACTTAAGGAAATCCACACTGATC  483
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658400  ACAAGAACACTCTCTTACACCATCACATCATTGCCACACTTAAGGAAATCCACACTGATC  133658459

Query  484        CAGGAACATTATTTGATATACAGTCCATATTCAGAATCCTCCAGGTATCAGTAATATTTT  543
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658460  CAGGAACATTATTTGATATACAGTCCATATTCAGAATCCTCCAGGTATCAGTAATATTTT  133658519

Query  544        AATTATGACCCATTTGGTGAATTTAAGATGGATTAGATGGCCTGTGGAGTTATTCATGGA  603
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658520  AATTATGACCCATTTGGTGAATTTAAGATGGATTAGATGGCCTGTGGAGTTATTCATGGA  133658579

Query  604        GCTGAAGACTTCTGTGAAGTAAAAAATAACCACCCAGTGTTAAGTGTAACTCTCGTGTGG  663
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133658580  GCTGAAGACTTCTGTGAAGTAAAAAATAACCACCCAGTGTTAAGTGTAACTCTCGTGTGG  133658639

Query  664        AATATTAAAATCAATTGTGTGttttttttGGCATCCCTTGATGTGAACTCTGGAGACCTG  723
                   |||||| |||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct  133658640  AATATT-AAATCAATTGTGTGTTTTTTTTGGCAT-CCTTGATGTGAACTCTGGAGACCTG  133658697

Query  724        ATGAGATGGAGGCTGTGATTGTTTACTCTGTGTTAGCTTGGACACCTTAGCTTGGCAGTA  783
                   ||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  133658698  ATGAGATGGAGGCTGTGATTGTTTACTCTGTGTTAGCTTGGACA-CTTAGCTTGGCAGTA  133658756

Query  784        ATCACAAGAACTTTAAGGAGTGCATCCCTTTGCTAGTTCATAAAGAATGTATTCTGATAA  843
                   ||||||||| ||||| |||||||| |||||||||||||||||| ||||||||||||||
Sbjct  133658757  ATCACAAGAGCTTTAGGGAGTGCAT-CCTTTGCTAGTTCATAAAG-ATGTATTCTGATAA  133658814

Query  844        CGGGGGACCAAGAGGATATATCCCGAATTGCAAA-TCGAGCAGTA-ATTACATT  895
                   |||||||||| |||||||||| ||||||||| |||||||| |||||||
Sbjct  133658815  CGGGGGACCAAAAGGATATATCCTGAATTGCAAAGTCGAGCAGTAGATTACATT  133658868
```

Figure 30. **BLAST alignment: ABL1 portion**

*The figure depicts the alignment report regarding the ABL1 portion of the BCR-ABL1 fusion segment. The range of the alignment on the NCBI reference genome (GRCh37.p13) is 133658102-133658868 (chr9). The identity is 761 out of 774 bp (98%). The thirteen unmatched bases include 7 deletions, 2 insertions and 5 SNVs in the query sequence.*

This illustrative procedure was performed for all samples. After the chromatograms checking and the alignment of the BCR-ABL1 fusion segments the following breakpoints coordinates were defined [Table 22]. The Sanger validation was not performed for ABL1-BCR breakpoints.

| Sample | Breakpoint (Chr22) | Breakpoint (Chr9) |
|---|---|---|
| **2** | 23631911 | 133648891 |
| **3** | 23632128 | 133593828 |
| **4** | 23575248 | 133590571 |
| **5** | 23654765 | 133708496 |
| **6** | 23634033 | 133658102 |
| **7** | 23634578 | 133684310 |
| **8*** | 23633411 | 133694917 |
| **9** | 23631875 | 133663995 |
| **10** | 23634573 | 133697001 |

Table 22. **Sanger BCR-ABL1 breakpoints**
*In this table the breakpoints coordinates (BP1 and BP4) identified by the analysis of BCR-ABL1 fusion segments obtained by Sanger sequencing.* Sample 8 coordinates were not detected starting from NGS data results, but from PCR based techniques.*

#### *4.2.4.3 BCR-ABL1 fusion point analysis*

The analysis of BCR-ABL1 fusion segments obtained by Sanger sequencing have shown that at the junction point the BCR and ABL1 segments share several nucleotides (micro homologies), as recently demonstrated in [110], [81] and [21].The length of the homology ranged from 1bp to 9bp. Moreover, insertion events were detected in 2 samples (3 and 7). In particular, in sample 7 we validated the insertion of 33bp detected in the NGS data analysis. More details regarding this insertion will be covered in the chapter dedicated to the $7^{th}$ sample. In patient 5 and 10 neither overlapping events nor insertions were detected as expected by the results obtained in the NGS data analysis.

| Sample | Event | Size |
|---|---|---|
| **2** | Micro-homology | 1bp (G) |
| **3** | Insertion | 2bp (CC) |
| **4** | Micro-homology | 3bp (TGC) |
| **5** | None | none |
| **6** | Micro-homology | 9bp (CTGCCTGGC) |
| **7** | Insertion | 34bp |
| **9** | Micro-homology | 1bp (A) |
| **10** | None | none |

Table 23. **Arrangements at BCR-ABL1 fusion point**
*The table shows the rearrangements that occur at BCR-ABL1 junction. The analysis has been performed starting from fusion sequences obtained by Sanger sequencing.*

### *4.2.5 Breakpoints coordinates comparison*

The breakpoints coordinates were then compared to those discovered with SVdetect and ClipCrop software. No comparison was made for breakpoints in ABL1-BCR fusion segment because Sanger validation was driven to spot BCR-ABL1 breakpoints only. The difference between coordinates predicted by software and those validated with Sanger sequencing resulted in few bases [Table 24] and [Table 25].

Discordant-pair approach using SVDetect permitted to spot BCR-ABL1 breakpoints with an accuracy of 15.89 ± 7.50 bp for BCR breakpoint and 6.37 ± 2.77 bp for ABL1. Refinement with soft-clipped reads using ClipCrop allowed increasing the detection performance with an accuracy of 0.44 ± 1.01 bp for BCR breakpoints and 1.22 ± 2.22 bp for ABL1 breakpoints. The gain in accuracy achieved by ClipCrop refinement is greater in the BCR region compare to the ABL1 region, indeed the discordant-pair approach resulted to performed better in ABL1 compare to BCR, whereas ClipCrop refined the breakpoints coordinates better in BCR compared to ABL1. This behavior can be explained by the percentage of repeated regions that is much higher in ABL1. Repeated regions can be handled in an easier way

by using the discordant-pairs approach than using the split-read approach. The difference between validated breakpoints coordinates and breakpoints coordinates detected by SVDetect is strictly dependent on the coverage achieved in the targeting experiment. Higher is the coverage, higher will be the chance to have a paired-end in which the pairs map with the 3' or 5' end to the breakpoints in BCR or ABL1 respectively. In patient 8 the breakpoints coordinates for BCR-ABL1 fusion gene were not detected.

| Sample | BCR Breakpoint (SVDetect) | | BCR Breakpoint (SVDetect + ClipCrop) | | BCR Breakpoint (Sanger) |
|---|---|---|---|---|---|
| **K562** | 23632762 | 20 | 23632742 | 0 | 23632742 |
| **2** | 23631930 | 20 | 23631911 | 1 | 23631910 |
| **3** | 23632124 | 4 | 23632128 | 0 | 23632128 |
| **4** | 23575267 | 19 | 23575251 | 3 | 23575248 |
| **5** | 23654742 | 23 | 23654765 | 0 | 23654765 |
| **6** | 23634051 | 18 | 23634033 | 0 | 23634033 |
| **7** | 23634602 | 24 | 23634578 | 0 | 23634578 |
| **8** | N.D | - | N.D | - | 23633411 |
| **9** | 23631897 | 22 | 23631875 | 0 | 23631875 |
| **10** | 23634577 | 4 | 23634573 | 0 | 23634573 |

Table 24. **BP1 coordinates comparison in different detection methods**
*The table shows the BP1 coordinates identified by discordant pair approach (SVDetect), refined by split-read approach (ClipCrop with BLAT) and validated by sanger sequencing. Sub-columns describe the difference in bp from the detected and the validated coordinates.*

| Sample | ABL1 Breakpoint (SVDetect) | | ABL1 Breakpoint (SVDetect + ClipCrop) | | ABL1 Breakpoint (Sanger) |
|---|---|---|---|---|---|
| **K562** | 133607140 | 5 | 133607145 | 0 | 133607145 |
| **2** | 133648884 | 8 | 133648891 | 1 | 133648892 |
| **3** | 133593826 | 3 | 133593828 | 1 | 133593829 |
| **4** | 133590561 | 10 | 133590571 | 0 | 133590571 |
| **5** | 133708492 | 4 | 133708496 | 0 | 133708496 |
| **6** | 133658098 | 4 | 133658108 | 6 | 133658102 |
| **7** | 133684300 | 10 | 133684307 | 3 | 133684310 |
| **8** | N.D | - | N.D | - | 133694917 |
| **9** | 133663988 | 7 | 133663995 | 0 | 133663995 |
| **10** | 133696999 | 2 | 133697001 | 0 | 133697001 |

Table 25. **BP4 coordinates comparison in different detection methods**
*The table shows the BP1 coordinates identified by discordant pair approach (SVDetect), refined by split-read approach (ClipCrop with BLAT) and validated by sanger sequencing. Sub-columns describe the difference in bp from the detected and the validated coordinates.*

### *4.2.6 Scaling the breakpoints identification*

In order to establish the minimum number of reads that were needed for detecting the BCR-ABL1 breakpoints, we performed the breakpoints identification at lane level using the same procedure adopted at sample level. With the analysis at lane level, as shown in Table 31 and Table 32, the breakpoints that give rise to the formation of the BCR-ABL1 oncogene were detected only in some samples. Moreover, considering the results on the different lanes in which each sample has been loaded, we observed heterogeneity across lanes for the same sample. This is primarily due to the low number of reads that can spot the breakpoint in a single lane. In samples 1,2,4,6,7,9 and 10 the BCR-ABL1 fusion was identified in each individual lane, whereas in samples 3 and 8 breakpoint identification completely failed, while in sample 5 only some lanes gave a useful reads yield. Is it worth highlighting that in samples 4 and 5 in some lanes the best

hits are tagged as translocation events and not as insertion fragment events. This finding shows how the SV call can be highly influenced by the set of mapping reads, in this case the discordant pairs, that span the fusion points. Looking at the number of paired-end reads supporting the breakpoints, is evident how breakpoints in K562 cell-line are easily detectable (267 ± 9.8 pairs) compared to those in patients (62.2 ± 20 pairs) with a fold-value of ~4x of reads supporting the translocation event. The failure of the detection in sample 3 could be related to the lower number of reads in this sample (3.76 M reads) compared to the others samples (4.46 M ± 2.26 M reads). This hypothesis could be valid if we suppose that in all patient specimens analyzed a similar number of $Ph^+$ cells was present. If not, the failure of the detection could be explained by a different amount of $Ph^+$ cells or by chance. In sample 5, in which only some lanes are able to spot the breakpoints, the number of pairs supporting the breakpoints is very close to the cut-off threshold used in SVDetect (20 pairs). This interesting finding allow us to define 4M as the lowest number of reads necessary to spot the BCR-ABL1 junction.

## 4.3 Sample Specific Breakpoints analysis

In order to facilitate the reading of the next chapters we will used the following shortcuts:

- BP1: Breakpoint in BCR that spot the BCR-ABL1 fusion
- BP2: Breakpoint in BCR that spot the ABL1-BCR fusion
- BP3: Breakpoint in ABL1 that spot the ABL1-BCR fusion
- BP4: Breakpoint in ABL1 that spot the BCR-ABL1 fusion
- DPA: Discordant pairs approach
- SRA: Split-Reads approach

### *4.3.1 K562 cell line*

K562 cell line sample, along with samples 2 and 3, was analyzed in the pilot phase of the project with a prior knowledge of the breakpoints coordinates that give rise to Ph and hence to BCR-ABL1 oncogene. The known breakpoints coordinates of K562 cell line were chr22:23632742 and chr9:133607147. By using SVDetect we detected two breakpoints coordinates in each chromosome. In particular BP1 mapped at chr9: 133607140, whereas BP4 mapped at chr22:23632762 [Table 19]. We then refined these coordinates to chr9:133607145 and chr22:23632742 respectively [Table 20]. There was no difference in bp with known coordinates both for BP4 [Table 25] and BP1 [Table 24]. The fact that we identified only two breakpoint is probably due to the fact that ABL1-BCR could be loss, as already reported in literature [46], [111], [112].

By enabling the visualization of the soft-clipped reads we can see also how single reads are splitted at the breakpoint. In particular, in BCR region, the breakpoint (BP2) is identified by reverse R-clipped split reads, whereas in ABL1 region the breakpoint (BP4) is identified by forward L-clipped split reads. This pattern allows to spot BCR-ABL1 breakpoints in all samples, as previously described in methods. Reads spanning the BCR-ABL1 junction can map in two ways. The first way is when one mate maps in BCR and the other maps into ABL1 regions (discordant pairs approach). The second way (split-reads approach) occurs when only one of the two mates span the junction point with a portion of its sequence mapping to BCR and the other portion mapping to ABL1. For this reason, in BCR the split reads are in reverse orientation and R-clipped. If the right clipped read were the forward oriented and not the reverse oriented, it would mean that an inversion occur at that point. This procedure of analysis will be carry out for all samples. A deeper analysis of the Figure 31 shows that both at the BCR and ABL1

breakpoint a dramatically coverage drop (from ~2500X to ~200X) occurs [Figure 32] and Figure 31.

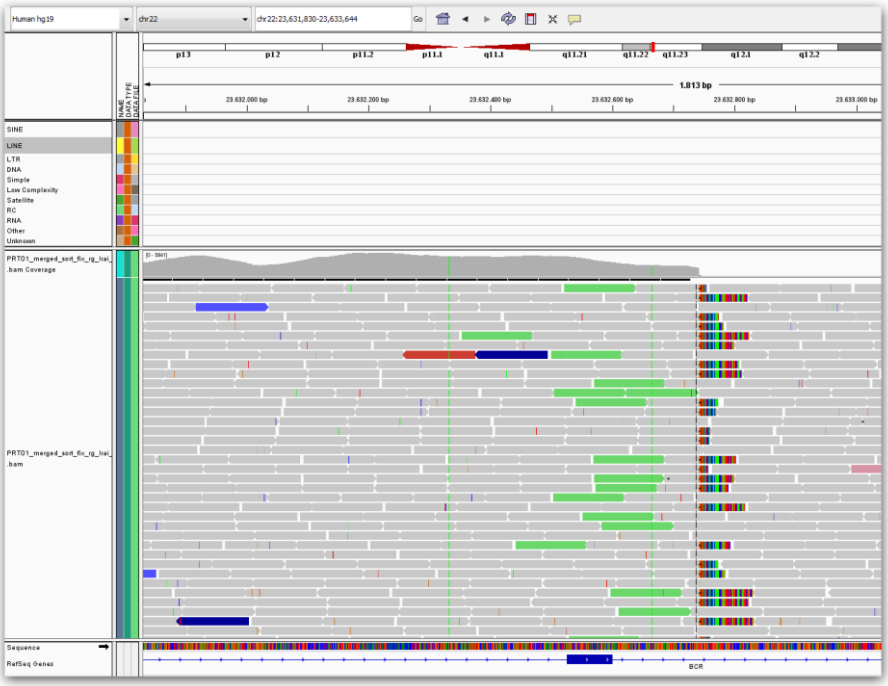



Figure 31. **K562 cell line: BCR-ABL1 fusion point at BCR gene**
*This picture shows how BP1 is spotted by discordant pairs (green forward reads) and right-clipped split reads (grey reads with a coloured unmapped portion). On the top the chromosomal localization and the genomic coordinates on the reference human genome (Hg19). Light-grey reads are normally mapped reads.*

This occurrence is a tangle demonstration of the der(9) loss. In Figure 32 reads were sorted to highlight the paired-end reads whose mate maps to different chromosomes. In Figure 32 is clear how the junction point on ABL1 could be detected by using discordant pair approach only.

BP4 is localized within a L1ME3A [Table 33], a member of the L1 repeat family, which belongs to the Long Interspersed Elements (LINEs) class. This repeat sequence maps to the Hg19 with the following coordinates:

chr9:133606693-133607233. Near BP4 there are also other repeated elements such as AluSx1 and L1MD [Table 33].

BP1 is far from repeated sequences but is proximal (142bp) to the end of the 14$^{th}$ exon (22:23632526-23632600) [Table 38].

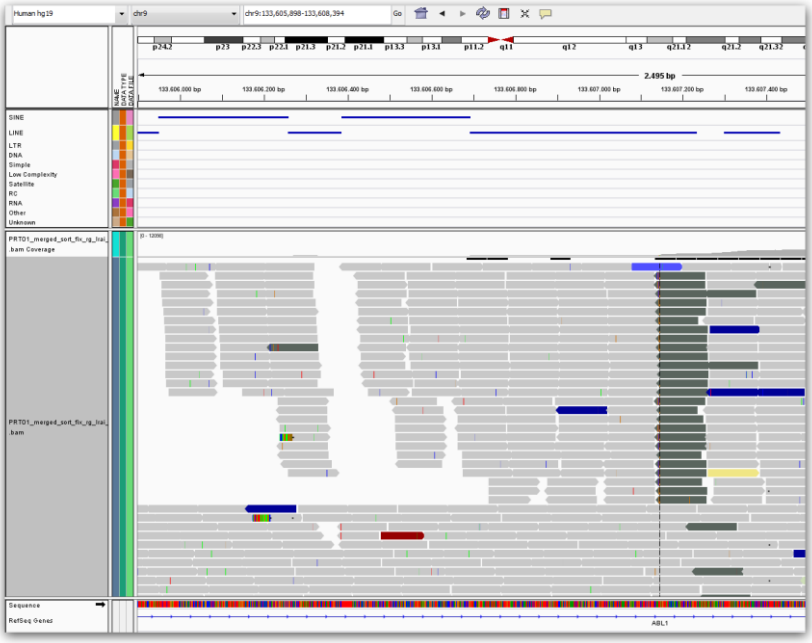



Figure 32. **K562 cell line: BCR-ABL1 fusion point at ABL1 gene**
*This picture shows how discordant pairs (dark grey reverse reads) spot the BP4 at ABL1. On the top, the chromosomal localization and the genomic coordinates on the reference human genome (Hg19) and the annotation of repeated regions located at these coordinates [blue lines].*

### *4.3.2 Patient 2*

Patient 2 was analyzed during the pilot phase together with the cell line and patient 3. For patient 2 BP1 and BP4 were already known (chr22:23631910 and chr9:133648892). By using SVDetect we identified BP1 and BP4 only. In particular, BP1 was localized at chr22:23631930, whereas BP4 at chr9:133648884. The difference in nucleotides between the SVdetect findings and the known coordinates was 20bp and 8bp for BP1 and BP4 respectively [Table 24] and [Table 25]. The refinement using ClipCrop allowed reaching an accuracy of 1bp for both BP1 and BP4 (chr22:23631911 and chr9:133648891) [Table 24], [Table 25]. BP1 is located 102 nucleotides from the 3' end of the 13th exon (23631704-23631808) [Table 38] in a region devoid of repeated sequences [Figure 33]. On the contrary BP4 is located within the AluJb (chr9:133648840-133649166) a repeat element belonging to the Short Interspersed Elements (SINEs) and near (64bp) to the 3' end of L1MEd, a LINE located at chr9:133648742-133648826 [Table 33]. BP2 and BP3 were not found and this is likely due to the low number of discordant pairs and split reads supporting them. Indeed there are only 9 paired-end reads that span the ABL1-BCR fusion point that are not enough to call the translocation event. Split reads analysis with ClipCrop was also unable to detect the BP3 due to the extremely low number of split reads (only 4) supporting it [Figure 35], even if in the BCR region the number of split reads that support BP2 is much higher [Figure 33]. This phenomenon could arise from the high density of repeated elements in the ABL1 region that make the mapping difficult. As depicted in Figure 35, both BP3 and BP4 are located within a region dense of repeated elements (AluJb, L1MEd, AluSp). The visual inspection enabling split reads view allows localizing the ABL1-BCR junction point at the following coordinates (chr9:133648781 - chr22:23632974). BP2 and BP3 were then confirmed manually by aligning

clipped reads using BLAT. Both BP2 and BP3 fall within a repeated element: L1ME1 [Table 36] and L1MEd [Table 37].

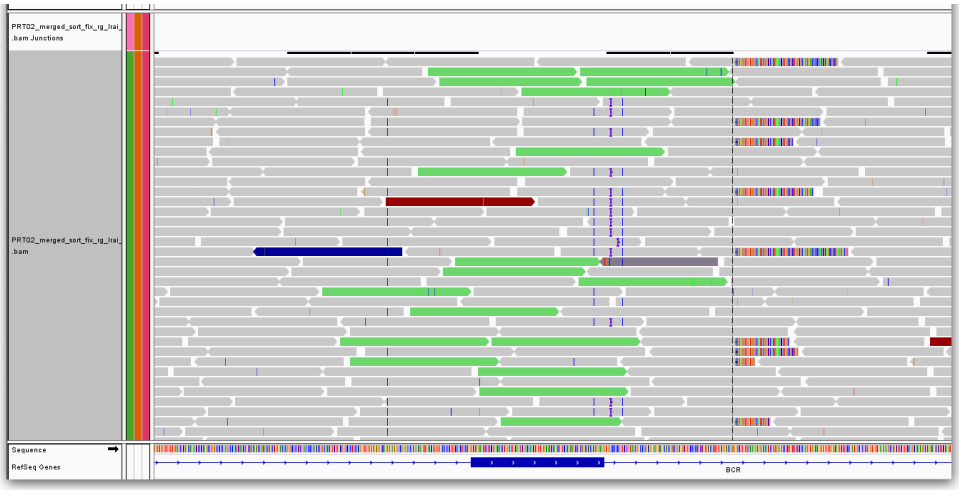



Figure 33. **Sample 2: BP1 at BCR gene**
*Discordant pair reads (DPR) whose mate maps to chromosome 9 are colored in light green. DPR with 5' → 3' direction along with R-clipped reads spot BP1. On the bottom (blue box) the exon 13 in BCR. Clipped portion is characterized by colored nucleotides.*

This manual and deeper analysis had pinpoint complications that can arise when the breakpoint region falls into a repeat-rich region. The low coverage of the ABL1-BCR junction could also be explained biologically by hypothesizing an unbalanced keeping of the der(9) in the cells analyzed. Moreover, considering the coordinates of the fusion points of both BCR-ABL1 and ABL1-BCR we could infer the following chromosomal break [Figure 34].


Figure 34. **Sample 2: DNA break**
*On the top to the left the BCR gene (green) with BP1 and BP2, on the top to the right the ABL1 with BP3 and BP4. On the bottom to the left the BCR-ABL1 fusion, whereas on the bottom to the right the ABL1-BCR fusion product. All the four images depict genomic sequences with 5' to 3' direction. In both BCR and ABL1 genes a deletion occurs after the break.*

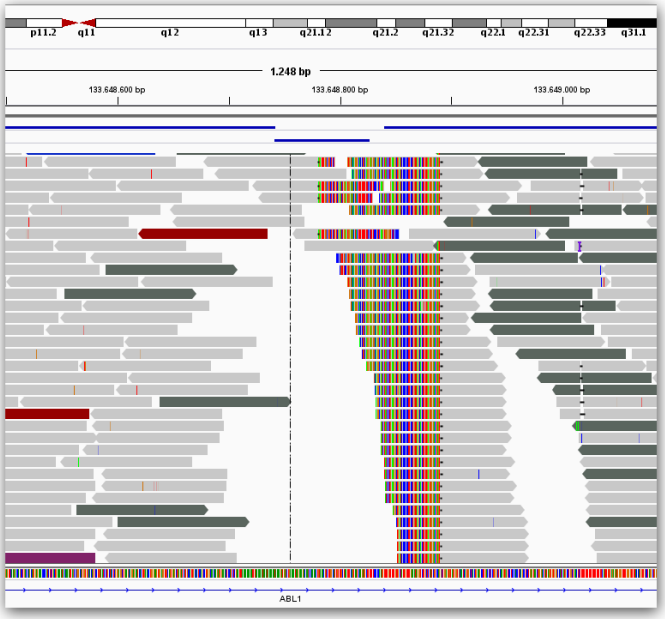



Figure 35. **Sample 2: ABL1 breakpoints**
*The figure depicts the region in which BP3 and BP4 occur. Dark grey reads (5' → 3') and R-clipped reads spot BP3 whereas BP4 is detected by dark grey reads (3' → 5') and L-clipped reads. L and R-clipped reads denote left and right clipping direction.*

### *4.3.3 Patient 3*

In patient 3, BP1 and BP4 were already known (chr22:23632128 and chr9:133593829). The analysis with SVDetect returned four breakpoints. BP1 and BP4 were located to chr22:23632124 and chr9:133593826, whereas the coordinates for BP2 and BP3 were chr22:23633550 and chr9:133593826 respectively [Table 19]. The number of discordant pairs supporting the junction points for BCR-ABL1 oncogene and ABL1-BCR fusion segment were 32 and 36 respectively. Taking into account the coordinates identified by SVDetect, BP1 and BP2 are 1426 bp distant, whereas BP3 and BP4 share the same coordinates supposing that in ABL1 there is neither loss nor gain of genomic material on Ph or der(9). The refinement with ClipCrop along set BP1 at chr22:23632128 and BP4 at chr9:133593828 [Table 20], whereas BP2 at chr22:23633556 and BP3 at chr9:133593805 [Table 21]. Hence the distance between BP1 and BP2 is 1428bp whereas between BP3 and BP4 only 23 bp [Figure 36]. The BP1 coordinate confirm that the breakpoint in BCR is located between exon 13 and 14 leading to BCR-ABL1 b2a2 transcript. The translocation is molecularly unbalanced and we can suppose that lead to the formation of both Ph and der(9), losing 1428bp in chromosome 22 and 23 bp in chromosome 9 . Regarding the proximity to repeated sequences, both BP3 and BP4 do not map within repeat elements but are flanked by two Charlie1a (DNA repeat AT-rich elements) which map to the following intervals: chr9:133593729-133593794 and chr9:133593835-133593985 [Figure 37] [Table 33]. These two repeated DNA elements are 9bp and 7bp far from BP3 and BP4 respectively [Table 37],[Table 33].

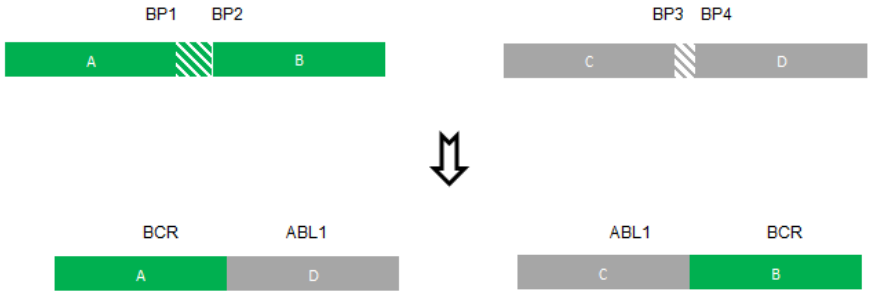



Figure 36. **Sample 3: DNA break**
*BCR (green) and ABL1(grey) genes both have two breakpoints. A+D segment depicts the BCR-ABL1 fusion whereas C+B segment depicts the ABL1-BCR reciprocal fusion product. Both at BCR and ABL1 genes a deletion after the translocation event occur (striped rectangles).*

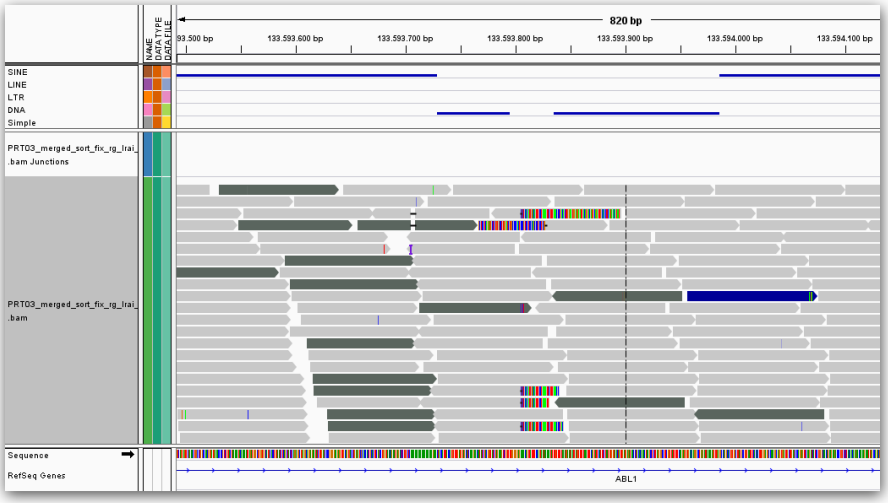



Figure 37. **Sample 3: ABL1 breakpoints**
*On the top the genomic coordinates at ABL1 gene. Blue lines depict repeated elements such as SINE (1st row) and low-complexity DNA elements (4th row). DPR with 5' to 3' direction and R-clipped reads spot BP3 whereas DPR with 3' to 5' direction and L-clipped reads identify BP4.*

In the BCR region, the breakpoint BP1 that spot the BCR-ABL1 junction maps in a genomic region free from repeated sequences, whereas BP2 maps within a L1ME1 (chr22:23633414-23633642) [Table 36] a L1 repeat that belongs to the LINEs [Figure 38].


Figure 38. **Sample 3: BP2 spotted by L-clipped split reads**
*Left-clipped (colored nucleotides) split reads spotting BP2. The blue box near the top is the L1ME1 repeated reads located at the BP2 coordinates. On the extreme top the cytogenetic band.*

Hence, in sample 3 both fusion products (BCR-ABL1 and ABL1-BCR) were detected. Clipped reads orientation was as expected by theory and no strange phenomena were detected except for the loss of 1426 bp in chromosome 22 and a dramatic decrease in coverage in the region upstream (100-200bp) to BP3 [Figure 37]. This latter event can be linked to the extremely high density of repeated sequences such as Alu elements and DNA AT-rich repeat elements.

### *4.3.4 Patient 4*

In this sample and in the following as well, BCR-ABL1 breakpoints were unknown. In these patient only the detection of mRNA had been performed by RT-PCR [Table 2]. In sample 4, by using SVDetect, four possible likely breakpoints were identified. However, the best results identified by SVDetect are tagged as an “Insertion fragment” event and not as a translocation event. This phenomena, has been observed in patients 4 and 5 whose breakpoints map in m-BCR and µ-BCR [Table 38]. In Figure 39 discordant pairs that map at the breakpoints in BCR and ABL1 are shown respectively. As we can see, both in BCR and ABL1 the paired-end reads that span the junction point of the BCR-ABL1 oncogene and ABL1-BCR fusion segment map very close (sometimes overlapping as in BCR) to the breakpoint with opposite direction [Figure 39].

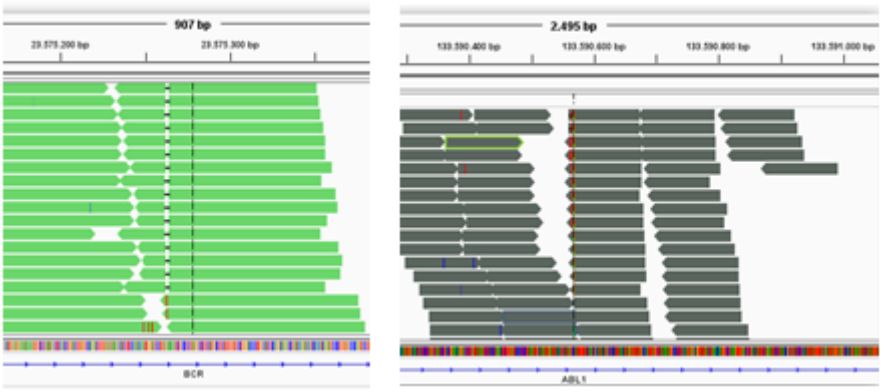



Figure 39. **Sample 4: Overlapping discordant pairs**
*On the left discordant pair reads (mapped at BCR), whose mate maps into chromosome 9 (ABL1), overlap with each other’s’ (5’→ 3’ direction with 3’ → 5’ direction . The same event occurs at ABL1 gene.*

This phenomenon could occur when the break is “blunt-like” and does not bring to loss of genetic material. In this case, the overlapping of forward with reverse paired-end reads that span the breakpoint in BCR [Figure 39] indicates that BP2 is upstream to BP1. This could be explained by a dsDNA blunt break in ABL1 and by a dsDNA protruding break (or ssDNA in two different point) in BCR gene [Figure 40].

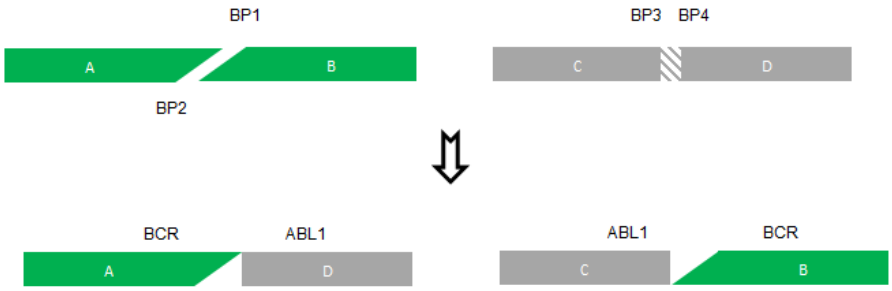



Figure 40. **Sample 4: DNA break**
*BCR (green) and ABL1(grey) genes both have two breakpoints. A+D segment depicts the BCR-ABL1 fusion whereas C+B segment depicts the ABL1-BCR reciprocal fusion product. In ABL1 a deletion after the translocation event occur (striped rectangles). In BCR, a strand specific protruding dsDNA break occurs.*

The BP1 and BP4 coordinates identified by SVDetect were chr22:23575267 and chr9:133590561 whereas those that identify the ABL1-BCR fusion (BP2 and BP3) were chr22:23575206 and chr9:133590576 [Table 19]. The refinement with ClipCrop along with the visual check using IGV have spot the following breakpoints coordinates: chr22:23575252 - chr9:133590570 [Table 20] and chr22:23575206 – chr9:133590570 [Table 21] for BCR-ABL1 and ABL1-BCR fusion respectively. Split-read approach had highlight very clearly how BCR and ABL1 fragments were joined. In Figure 41 we can see that in ABL1 at the coordinate chr9:133590570 there are two types of clipped reads, L-clipped that spot the BCR-ABL1 junction and R-clipped that identify the ABL1-BCR fusion. In patient 4 the break of chromosome 9 occurs in a single point, or

better, is not followed by a gain or a loss of genetic material that would have spaced BP3 and BP4.

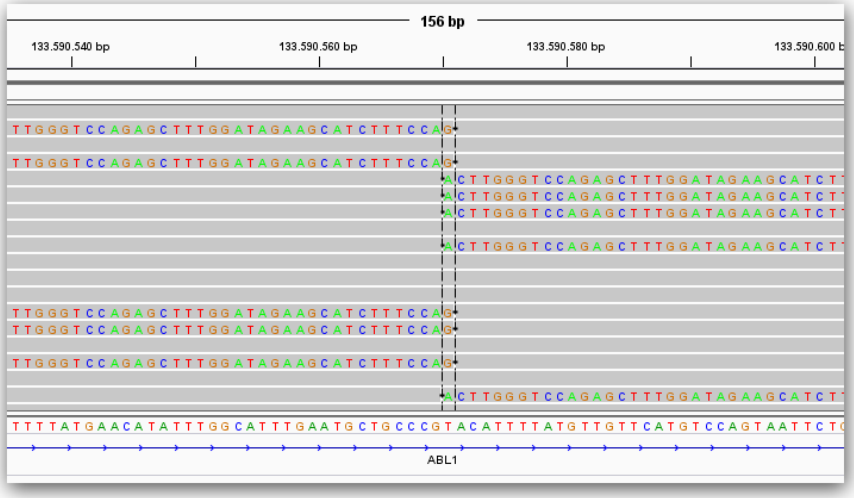



Figure 41. **Sample 4: Split reads spot BP3 and BP4**
*L and R clipped split reads at ABL1 detect the same breakpoint. This result lead to consider the DNA break in this patient "blunt like" on the ABL1 gene. On the top the genomic coordinates on the Hg19 reference genome.*

Regions sorrounding both breakpoint in BCR and ABL1 lack of repeated elements. The only repeated sequence in proximity of the breakpoints is the AT-rich repeat DNA element named Charlie4z (chr9:133590242-133590294) [Table 33], [Table 37].

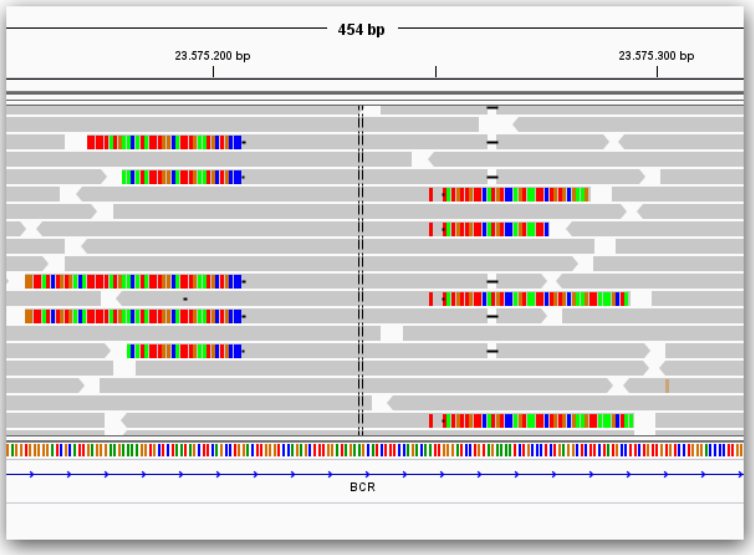



Figure 42. **Sample 4: Split-Read identify breakpoints in BCR**
*Split-Reads (L-clipped and R-clipped) identify BP2 and BP1 respectively. BP2 is upstream to BP1.*

### 4.3.5 Patient 5

Sample 5 was from a CML patient hat express the p230 transcript and related oncoprotein. This fusion transcript is formed when BP1 is located in μ-BCR between exon 19 and 20 [Table 2]. The analysis with SVDetect had define the rearrangement as an insertion fragment, similarly to that identified in patient 4. The breakpoints coordinates identified with SVDetect were four: chr22:23654742 (BP1), chr9:133708492 (BP4) and chr22:23654774 (BP2), chr9:133708509 (BP3) that identify the BCR-ABL1 and ABL1-BCR junction points respectively [Table 19]. ClipCrop analysis had refined BP1 to chr22:23654766 and BP4 to chr9:133708495 [Table 20], whereas BP3 chr9:133708500 and BP2 to chr22:23654766 [Table 21]. This breakpoint distribution is opposite to that occurs in patient 4. In sample 5 there is a unique breakpoint (BP1 equal to BP2) in BCR [Figure 44], whereas in ABL1 two breakpoints occurs [Figure 45]. In this case BP4 (133708495) is upstream to BP3 (133708500) of only 5 bp. This event is a balanced translocation with neither gain nor loss of material [Figure 43].

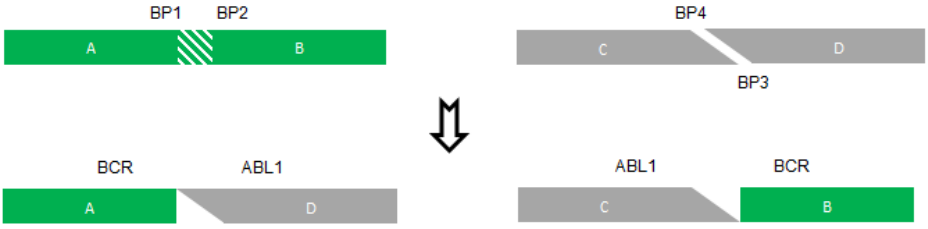



Figure 43. **Sample 5: DNA break**
*In BCR two breakpoints occur in a blunt-like way and are followed by nucleotide loss. In ABL1 the DNA break is protruding with BP3 downstream to BP4.On the bottom the two fusion products: BCR-ABL1 and ABL1-BCR.*

The explanation of this rearrangement is similar to the one hypothesized before but in this case the protruding break occurs at chromosome 9 instead of chromosome 22. The tag “insertion fragment” provided by SVDetect is probably a mistake that arise from the overlap of discordant pairs that support BCR-ABL1 and ABL1-BCR fusion points. Among the non-best results at sample level, there are several records tagged as translocation events with the same coordinates assigned to the “insertion fragment” findings [data not shown]. Moreover, at lane-level analysis the presence of the translocation event has been found as best hits [Table 31] and [Table 32]. The coordinates that give rise to BCR-ABL1 oncogene were validated with Sanger sequencing [Table 22]. As in patient 4, breakpoints in patients 5 are far from repeated sequences. Breakpoints that occur at BCR are locate between exon 19 and exon 20, confirming the previous info collected for patient [Table 38].

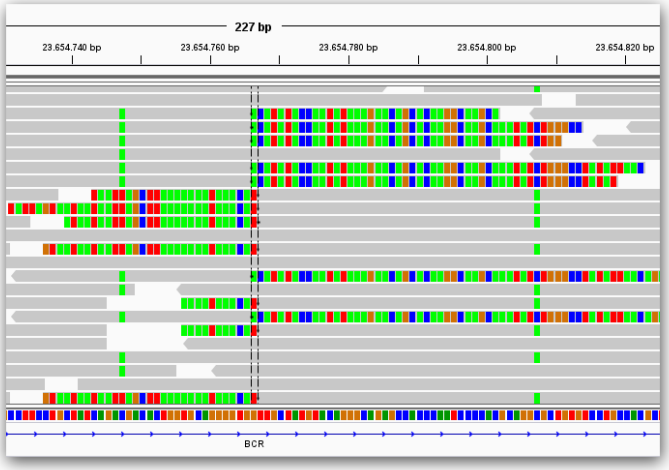



Figure 44. **Sample 5: split-reads identify BP1 and BP2**
*Split-Reads (L-clipped and R-clipped) identify BP2 and BP1 respectively. In this case BP1 and BP2 are coincident, denoting a blunt chromosomal break.*

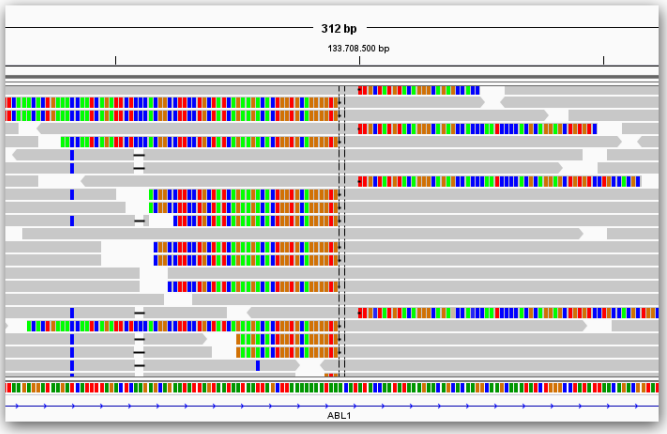



Figure 45. **Sample 5: Split reads spot breakpoints in ABL1**
*Split-Reads (L-clipped and R-clipped) identify BP2 and BP1 respectively. In this case BP1 and BP2 are coincident, denoting a blunt chromosomal break.*

### *4.3.6 Patient* 6

The only prior knowledge concerning this patient was that it expressed the p210 (b3a2) isoform of BCR-ABL1 transcript [Table 2]. This means that the breakpoints in BCR should map between exons 14 and 15. The analysis with SVDetect has indicated four breakpoints. BP1 at chr22:23634051, BP4 at chr9:133658098, BP2 at chr22:23634637 and BP3 at chr9:133658314 [Table 19]. Breakpoints that occur at BCR are located between exon 14 and exon 15, confirming the previous info collected for patient [Table 38].

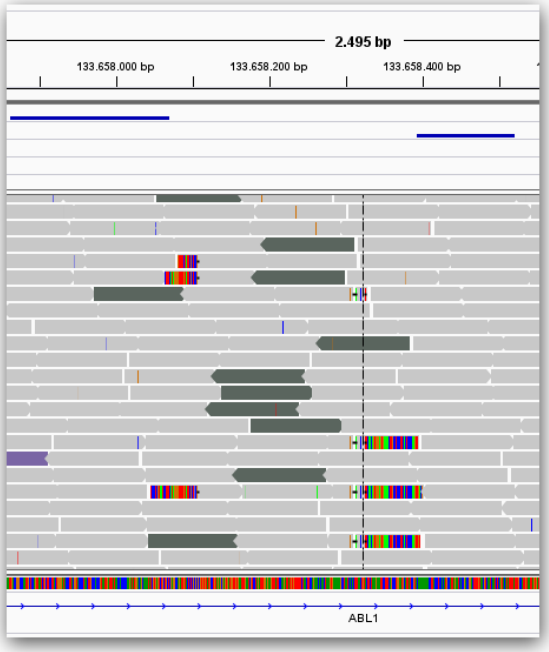



Figure 46. **Sample 6: Spotting BP3 and BP4 in ABL1 gene**
*Discordant pair reads (forward and reverse oriented) and split-reads (right and left clipped) identify BP3 and BP4 respectively. On the top, flanking repeated elements as blue lines (SINE miRb on the left and LINE L2b on the right).*

In ABL1 the discordant-pair reads that spot BP3 and BP4 overlap with each other's. This means that BP4 is upstream to BP3. Instead BP2 is downstream to BP1, supposing the loss of genetic material ranging between these coordinates [Figure 48]. The number of reads pairs supporting the translocations identified in SVDetect results unbalanced towards the BCR-ABL1 fusion (227 reads) compared to the ABL1-BCR fusion (42 reads) [Table 19]. The refined analysis by using ClipCrop and the visual check with IGV have spotted the following coordinates: chr22:23634034 – chr9:133658107 [Table 20] and chr22:23634639 – chr9:133658311 [Table 21] for BCR-ABL1 and ABL1-BCR fusions respectively.

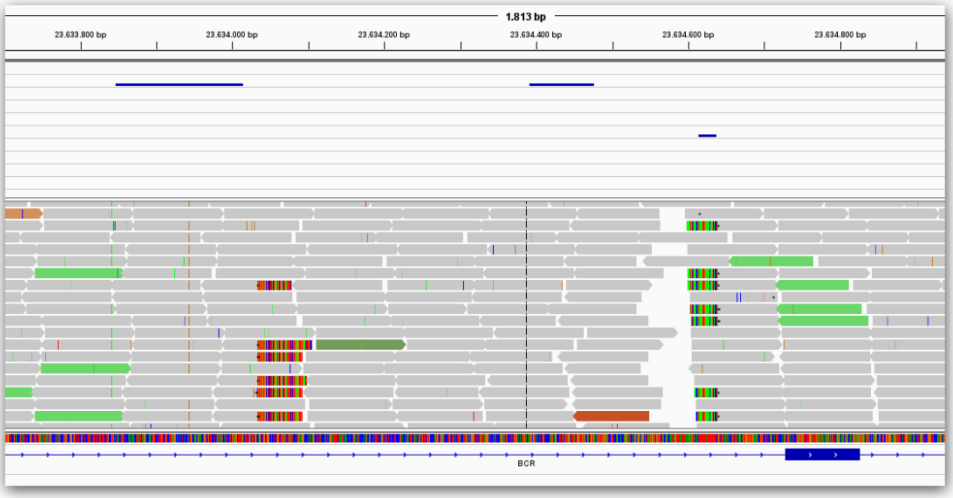



Figure 47. **Sample 6: DPRs and split-reads identify BP1 and BP2.** *Discordant pair reads (forward and reverse oriented) and split-reads (right and left clipped) identify BP1 and BP2 respectively. On the top, flanking repeated elements as blue lines (SINE on the left and AT-RICH on the right).*

The number of clipped reads that support the BCR-ABL1 fusion (180 R-clipped and 23 L-clipped) [Table 20] is higher than that spotting the ABL1-BCR (53 right-clipped and 53 left-clipped) [Table 21] . The low number of left-clipped reads in ABL1 is probably due to their proximity (37bp) to the SINE MIRb (chr9:133657864 - 133658070) [Table 37]. The low number of supporting clipped reads for ABL1-BCR could be ascribed to deletion of der(9) genomic regions [46]. This trend is common to all samples having both BCR-ABL1 and ABL1-BCR junction except for patient 3 in which the number of clipped reads that support BCR-ABL1 fusion is lesser than the one that support the ABL1-BCR fusion [Table 20], [Table 21]. Hence, BCR region probably underwent to dsDNA break with a loss of genomic material (605bp), whereas in ABL1, an unbalanced protruding dsDNA break occurred differently in chromosome 9 strands [Figure 48]. BP1 and BP4 that support the BCR-ABL1 fusion were found near (20bp) to the LINE L2b (chr22:23633847-23634014) [Table 35] and proximal (37bp) to the SINE

MIRb (chr9:133657864-133658070), [Table 34]. BP2 and BP3 that support the ABL1-BCR fusion were proximal (83bp) to the LINE L1MC4 (chr9:133658394 - 133658520) [Table 37] and very close (2bp) to the AT-rich DNA element (chr22:23634615-23634637) [Table 36]. Validation with Sanger sequencing provided the following breakpoints coordinates: chr22:23634033 (BP1) and chr9:133658102 (BP4) [Table 22]. The difference between such coordinates with the ones identified by the software analysis is 1bp and 5bp for BP1 and BP4 respectively.


Figure 48. **Sample 6: DNA break**
*In BCR two breakpoints occur in a blunt-like way and are followed by nucleotide loss. In ABL1 the DNA break is protruding with BP3 downstream to BP4.On the bottom the two fusion products: BCR-ABL1 and ABL1-BCR.*

### *4.3.7 Patient 7*

SVDetect results revealed that patient 7 presents a translocation event spotting BCR-ABL1 fusion gene and an inverted translocation event that involves the ABL1 and FOXRED2 genes Table 19]. More details about this latter rearrangement will be discussed further in the chapter after the description of the BCR-ABL1 fusion. About patient 7, it was only known that the p210 (b3a2) transcript form was expressed [Table 2]. Breakpoints coordinates (chr22:23634602 and chr9:133684300) that identify the BCR-ABL1 fusion gene confirmed that the junction is located in BCR between the 14$^{th}$ and 15$^{th}$ exon [Table 38]. Refinement with ClipCrop and visual check using IGV have rounded the coordinates to the following boundaries:

chr22:23634578 and chr9:133684307 [Table 20]. The number of paired-end reads supporting the junction as well as the number of right-clipped (BCR) and left-clipped (ABL1) reads was very high: 227, 177 and 166 reads respectively [Table 19, Table 20]. BP1 is located 37bp upstream to the AT-rich DNA element (chr22:23634615-23634637) [Table 35], [Figure 49].

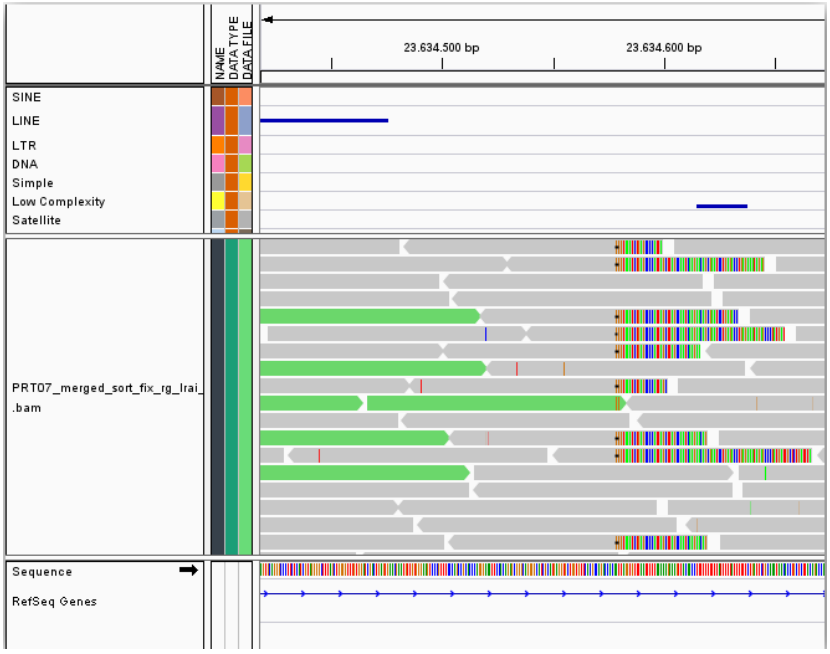
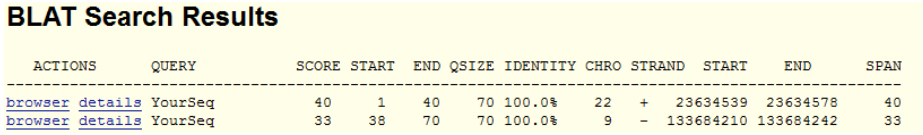



Figure 49. **Sample 7: BCR-ABL1 fusion point at BCR**
*This figure shows how DPR whose mate map to chromosome 9 (light green forward reads) as well as R-clipped reverse reads can spot BP1. AT-Rich low complex element (small blue line at right-top) is located 37 bp downstream BP1.*

This DNA element is the same found very close to the BP2 in sample 6 [Figure 47]. This finding could support the hypothesis that this low-complexity genomic region can actively promote the DNA break.

Regarding the ABL1 region, BP4 mapped within the L1MEc (chr9:133684219-133684749) LINE [Figure 52], [Table 34]. By performing the re-alignment of the longest clipped portion of the L-split reads around

BP4 by using BLAT, it was noticed that it did not map completely into BCR but also in a region upstream to the breakpoint itself in ABL1 at chr9:133684210-133684242 with an inverted orientation [Figure 50]. This is a clear evidence that these clipped reads are able to span both the junctions between BCR (5' to 3') with an inverted fragment of 33bp (chr9:133684210-133684242) and the fusion of the latter with the remnant ABL1 portion (5' to 3'). This phenomenon lead to suppose that at the BCR-ABL1 junction point an insertion of the fragment chr9:133684210-133684242 is present.



**BLAT Search Results**

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 40 | 1 | 40 | 70 | 100.0% | 22 | + | 23634539 | 23634578 | 40 |
| browser details | YourSeq | 33 | 38 | 70 | 70 | 100.0% | 9 | - | 133684210 | 133684242 | 33 |

Figure 50. **BLAT alignment: L-clipped split reads at ABL1**
*The figure shows the BLAT alignment of the L-clipped read at ABL1 spotting BP4. The first 40$^{th}$ nucleotides map to BCR, whereas from 38$^{th}$ to 70$^{th}$ nucleotide the fragment map to ABL1 reversely.*

Sanger sequencing validated the breakpoints coordinates (chr22:23634578 and chr9:133684310) [Table 22]. The difference among the calculated and the validated coordinates was 0 and 3bp for breakpoints coordinates in BCR and ABL1 respectively. The hypothesis of the inserted fragment was validated. In fact, the alignment of the FASTA sequence obtained by Sanger sequencing with BLAT and BLAST demonstrated that the fusion segment contains the insertion of a small fragment (37bp), which maps from chr9:133684206 to chr9:133684242 [Figure 51].

**BLAT Search Results**

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 101 | 276 | 438 | 513 | 96.4% | 22 | - | 23634032 | 23634578 | 547 |
| browser details | YourSeq | 59 | 184 | 242 | 513 | 100.0% | 9 | - | 133684310 | 133684368 | 59 |
| browser details | YourSeq | 42 | 126 | 173 | 513 | 97.8% | 6 | - | 67713001 | 67713049 | 49 |
| browser details | YourSeq | 37 | 242 | 278 | 513 | 100.0% | 9 | + | 133684206 | 133684242 | 37 |
| browser details | YourSeq | 31 | 188 | 226 | 513 | 75.8% | 1 | - | 240196778 | 240196810 | 33 |
| browser details | YourSeq | 28 | 146 | 183 | 513 | 93.8% | 6 | + | 19045378 | 19045416 | 39 |

Figure 51. **BCR-ABL1 fusion: Sanger Sequencing Validation**
*The figure shows the BLAT alignment of the fusion fragment sequenced by sanger method.*

The alignment with BLAT shows also a micro homology (3bp) between the BCR segment and the 37bp-inserted fragment, verifying the length (33bp) observed in reads analysis. SVDetect found an inverted translocation with the following coordinates: chr9:133684252 and chr22:36901054 [Table 19] that were refined to chr9:133684236 and chr22:36900838 [Table 21]. The analysis of this inverted translocation has pinpointed a very complex rearrangement that occur during the DNA break in the BCR region. Considering the breakpoints coordinates obtained by SVDetect analysis, it seems that instead of ABL1-BCR fusion segment, an ABL1-FOXRED2 fusion segment is formed without BCR implication [Figure 54]. The manual and curated analysis of the breakpoints regions allowed to speculate that during the formation of der(9), a fragment of FOXRED2 has been inserted at the fusion point between ABL1 and BCR.

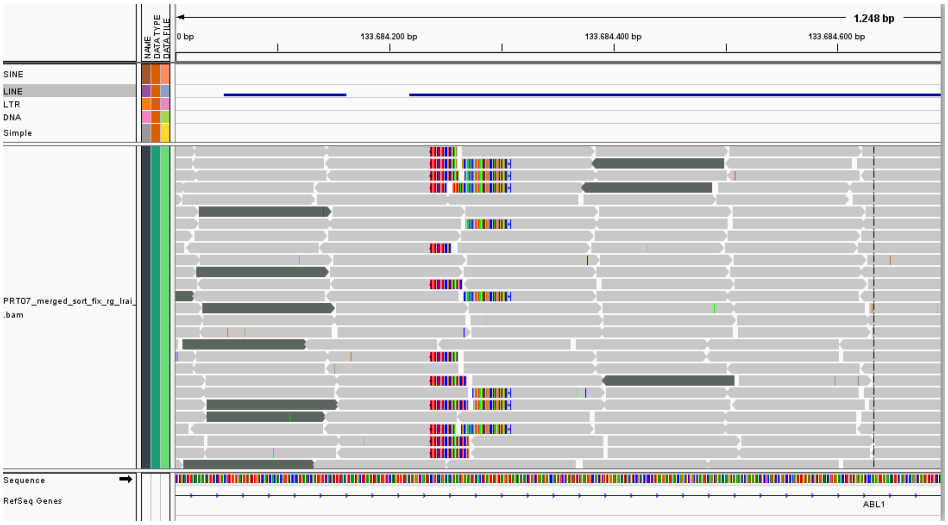



Figure 52. **Sample 7: DPRs and split-reads spot BP3 and BP4**
*The figure shows BP3 and BP4 detected by R-clipped and L-clipped split-reads respectively as well as forward and reverse discordant pair reads whose mate maps to chromosome 22 (dark grey). On the top blue lines describe repeated elements (LINEs).*

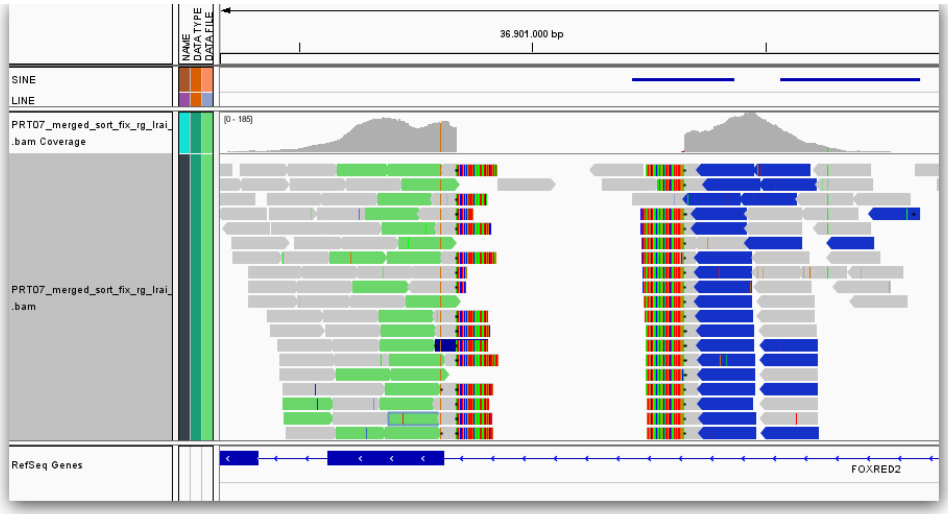



Figure 53. **Sample 7: FOXRED2 gene region (zoom in)**
*The figure depicts the FOXRED2 gene region involved in the formation of the ABL1-BCR fusion segment. Blue reads describe an insertion event, whereas green forward reads, which spot BP2, are the mate pair of discordant pair reads spotting BP3 at ABL1 gene. The inverted orientation of green reads as well as that of split reads is due to the inversion event occurring at BCR in the region upstream to BP1* [Figure 55].

To demonstrate this hypothesis, all the information available from the reads (split-reads, discordant pair, strand orientation, insert size and so on) were used. We noticed that mate pairs of the forward-oriented discordant reads, mapping in ABL1, instead of mapping in BCR with reverse direction as expected, they mapped to the FOXRED2 gene with forward direction [Figure 53]. This means that ABL1 is fused with FOXRED2 gene that appear inverted. This is also demonstrated by: 1) the clipped portion of L-split reads in ABL1 that map inverted to FOXRED2 and 2) by the clipped portion of R-split reads in FOXRED2 that map to ABL1 inversely and contiguously to that at point 1) [Figure 52] and [Figure 53].


Figure 54. **ABL1-FOXRED2 fusion detected by SVDetect**
*The figure shows the fusion ABL1-FOXRED2 detected by both discordant pairs and split reads (half orange and blue). The fragment is positioned with a 5' to 3' direction. The inverted FOXRED2 gene fragment is fused with ABL1.*

Once established that ABL1 was fused with FOXRED2, in the way depicted in Figure 54, we analyzed the clipped portion of L-split reads mapping to FOXRED2 [Figure 53]. What came out was that this clipped portion map to BCR inversely and contiguously to the clipped portion of R-split reads that instead map to FOXRED2 inversely and contiguously to the former clipped portion. This means that FOXRED2 (inverted) is fused with BCR [Figure 56]. This arrangement is also confirmed by blue colored paired-end reads that spot insertion events. In this case, the insertion size computed over blue-reads in reverse direction mapping into BCR is positive meaning that the pairs of such reads map downstream to them [Figure 55].

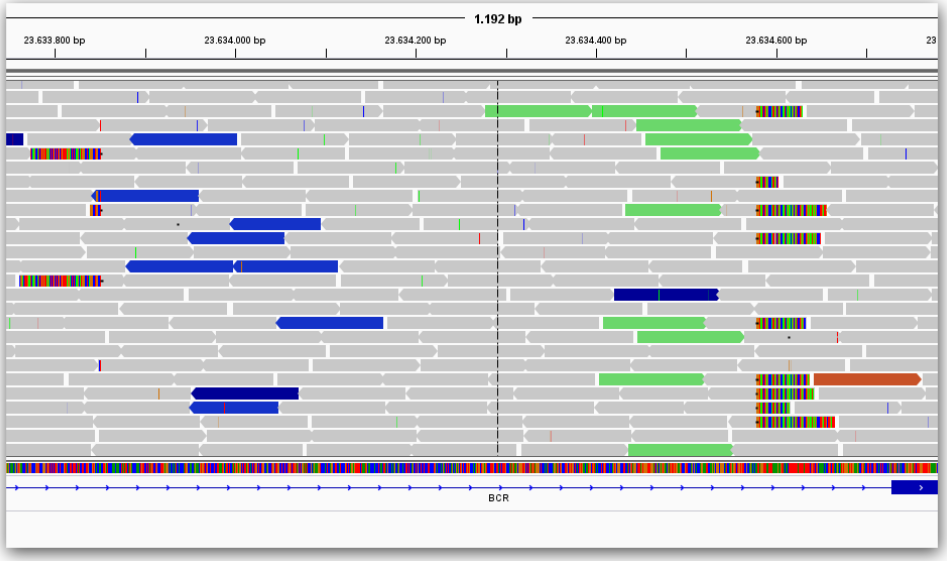



Figure 55. **FOXRED2 insertion at BCR**
*The figure shows the insertion (FOXRED2 fragment) identified by reverse blue reads upstream to BP1.*

Indeed FOXRED2 is downstream to BCR. The fact that these pairs mapped not only downstream but also inverted, in reverse direction as their mate in BCR gene, demonstrated that FOXRED2 is inversely fused with BCR [Figure 53]. These findings lead to define the following chromosomal rearrangement [Figure 56].

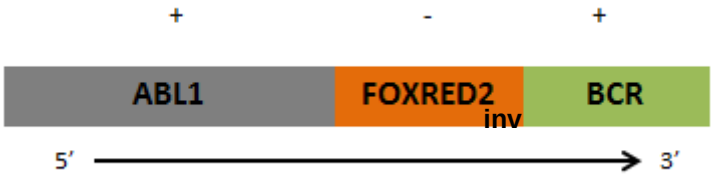



Figure 56. **Sample 7: ABL1/FOXRED2/BCR fusion segment**
*The figure shows the ABL1/FOXRED2/BCR fusion taking as reference the forward direction 5' → 3'. The fragment of the FOXRED2 gene is inserted inversely at the ABL1 and BCR junction. +/- described the strand.*

The analysis of this patient had pinpointed how complex could be the rearrangements at the breakpoint. Both BCR-ABL1 oncogene and ABL1-BCR fusion segment underwent to the insertion of an inverted fragment of DNA. In the case of BCR-ABL1 oncogene, the insertion of an inverted ABL1 sequence upstream to the BCR-ABL1 breakpoint at ABL1 occurred, whereas regarding the ABL1-BCR fusion segment, the insertion of an inverted fragment of the FOXRED2 gene occurred. Sanger sequencing validated the breakpoints coordinates that give rise to the BCR-ABL1 oncogene and the ABL1 inserted fragment [Table 22]. No validation were carried out for the ABL1-BCR fusion segment.

### *4.3.8 Patient 8*

Patient 8 was the only one in which the identification of BP1 and BP4 breakpoints failed. The only coordinates identified with SVDetect were those related to BP2 and BP3 with the following coordinates (chr22:23632905 and chr9:133694955) [Table 19]. Such coordinates were then refined with ClipCrop to chr22:23632908 and chr9:133694990 [Table 21]. These coordinates were used to set up PCR for amplifying the BCR-ABL1 fusion segment. After several try, we succeeded to amplify the BCR-ABL1 fusion segment. Sanger sequencing of this amplicon provided the following coordinates: chr22:23633411 (BP1s) and chr9:133694917 (BP4s). The first coordinate map to the 3' end of the SINE, AluSx1 (chr22:23633113-23633411) and 3bp upstream to the 5' end of L1ME1 (chr22:23633414-23633642), member of LINEs [Table 35]. The breakpoint in the ABL1 region is located 3bp upstream to the 5' end of the AluSx3 (chr9:133694920-133695226) [Table 34].

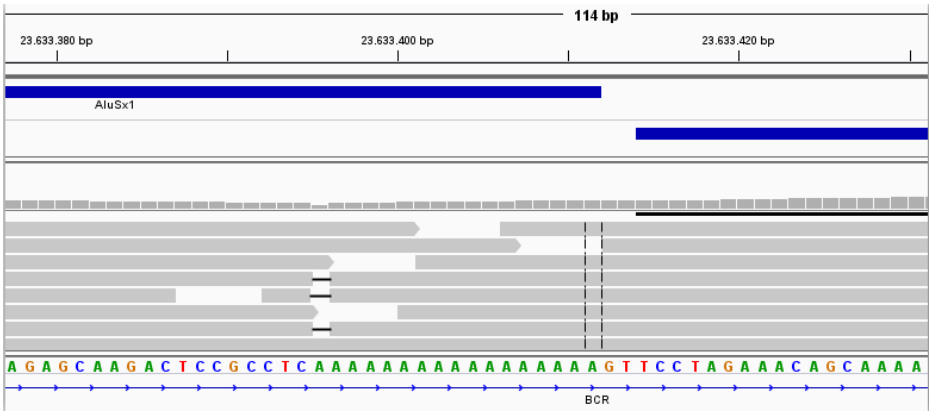



Figure 57. **Sample 8: BP1 at BCR**
*The figure shows the genomic region in which the breakpoint BP1 is identified by PCR map. In particular the two dashed vertical lines sport the nucleotide at which the break occurs. BP1 is flanked by two repeated elements such as AluSx1 (left blue line) and L1ME1 (right blue line). At the bottom the nucleotide sequence with a long stretch of A (green).*

The proximity to repeated sequences (2 SINEs and 1 LINE) along with the low-complexity of both regions has made it impossible to detect BP1 and BP4 both with discordant pair approach and split read approach. The region upstream to the BP1s is featured by a long stretch of nucleotide A [Figure 57], whereas sequence downstream to BP4s is characterized by a long stretch of Thymine (23) [Figure 58]. Achieve a reasonable coverage of this genomic region is very difficult both for the low-complexity of the genomic sequence and for the high density of SINEs and LINEs that exacerbate this issue. BP1 coordinate confirm the b2a3 transcript isoform previously detected in RT-PCR [Table 2], indeed it maps between exon 14 and 15 [Table 38].

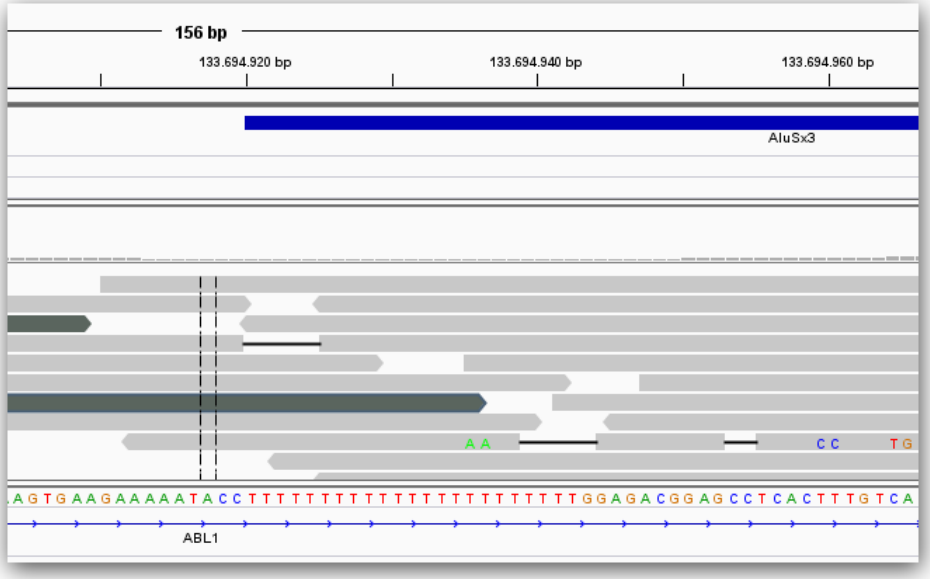



Figure 58. **Sample 8: BCR-ABL1 breakpoint at ABL1.**
*The figure shows the genomic region in which the breakpoint BP4 is identified by PCR map. In particular the two dashed vertical lines sport the nucleotide at which the break occurs. BP4 is very close at 3' to the AluSx3 (blue box). At the bottom the nucleotide sequence with a long stretch of T (red).*

Moreover, considering the coordinates of the fusion points of both BCR-ABL1 and ABL1-BCR we could infer the following chromosomal break [Figure 59].

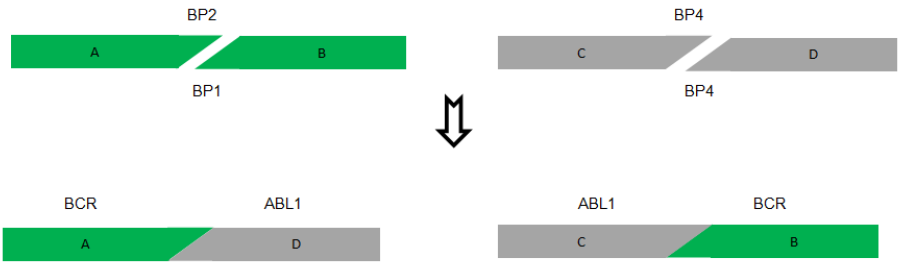



Figure 59. **Sample 8: DNA break**
*In BCR (green rectangle) the break occur in a protruding way producing two breakpoints in which BP2 is upstream to BP1. Also in ABL1 the DNA break is protruding with BP4 upstream to BP3.On the bottom the two fusion products: BCR-ABL1 and ABL1-BCR.*

### *4.3.9 Patient 9*

Sample 9 belongs to a patient in which has been monitored the levels of the p210 b3a2 transcript. BP1 and BP4 have the following refined coordinates: chr22:23631876 and chr9:133663994 [Table 20]. No breakpoints leading to ABL1-BCR fusion product were identified. BP1 is located in a region free from repeated sequences, whereas BP4 maps within the LINE, L1M4c (chr9:133663585-133664351) [Table 34]. In addition to the BCR-ABL1 fusion points, SVDetect has identified a reverse translocation with the following coordinates: chr9:136985521 and chr22:23631884 [Table 19]. Mapping with BLAT the clipped portion of the L-splitted reads at chr9: 136985521 to the reference genome it resulted that it mapped partially to very small fragments that belong to chromosome 2 and chromosome 3. This findings could be explained by the presence of the AluSz (chr9:136985256-136985553). Probably a piece of chromosome

9 near to the telomere rearranged with BCR. The same finding has been reported previously in [91]. Considering the breakpoints coordinates identified we could infer the DNA break depicted in [Figure 62].

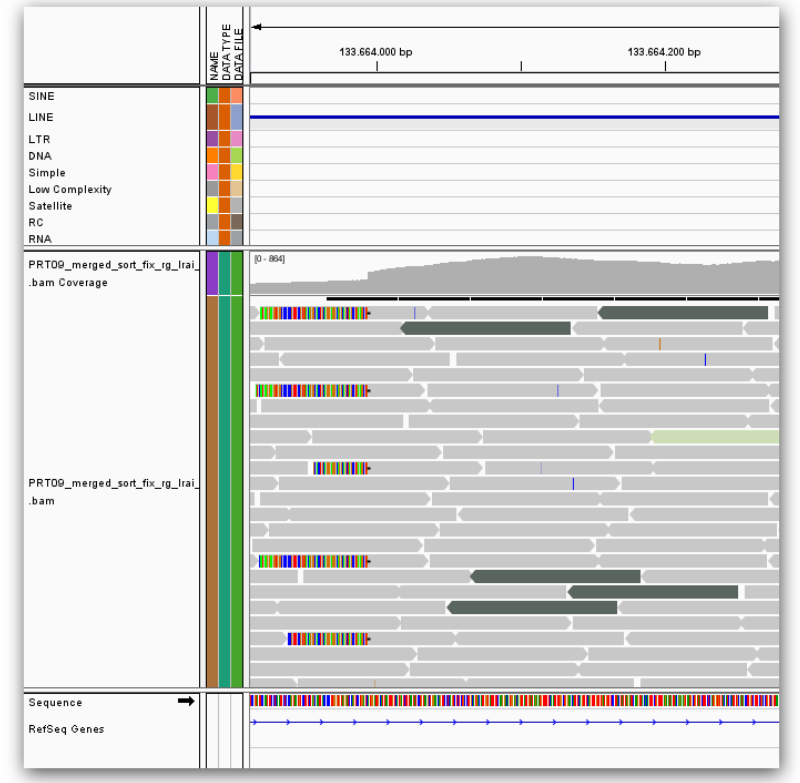



Figure 60. **Sample 9: BP4 at ABL1 region**
*The figure shows how split-reads (half light grey and half mixed colored) and reverse discordant pair reads (dark grey) spot BP4 at ABL1 gene. On the top the blue line that describe the LINE L1M4c.*

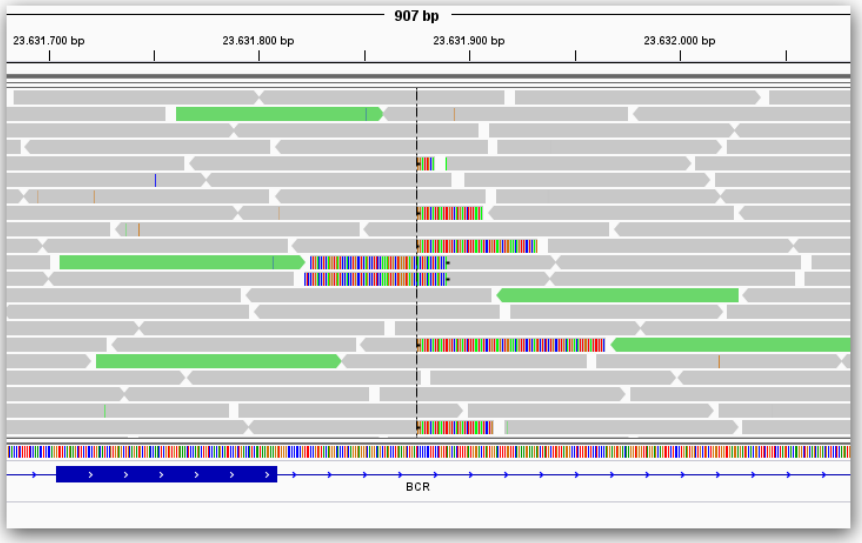



Figure 61. **Sample 9: BP1 and BP2 at BCR region**
*The figure shows BP1 and BP2 spotted by R-clipped and L-clipped split reads respectively. These breakpoints are also identified by forward and reverse discordant pair reads (light green colored). At the bottom the blue line that describes the exon 13. At the top the genomic coordinates according to Hg19.*

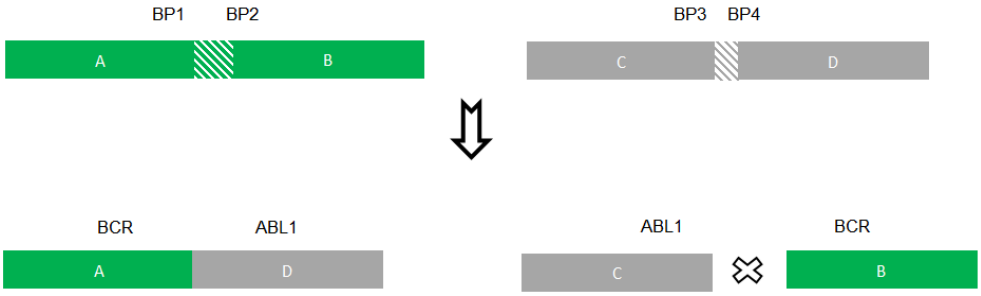



Figure 62. **Sample 9: DNA break**
*The figure shows how DNA could be broken in this sample. BP1 is upstream to BP2 as well as BP3 is upstream to BP4. Such feature lead to hypothesize that DNA break is blunt like and is followed by loss of nucleotides. Only the fusion product BCR-ABL1 is formed.*

### *4.3.10 Patient 10*

Sample 10 belongs to a patient in which has been found a translocation t(9;22) by using CBA technique. Moreover, the p210 (b3a2) transcript has been identified by RT-PCR. This prior knowledge was verified by the analysis. SVDetect reported the following breakpoints that give rise to the BCR-ABL1 oncogene: chr22:23634590 (BP1) and chr9:133696999 (BP4) [Table 19]. These coordinates were then refined by ClipCrop and visually checked in IGV. The results are the following: chr22:23634574 (BP1) and chr9:133697000 (BP4) [Table 20]. Sanger sequencing validated these breakpoints coordinates [Table 22].

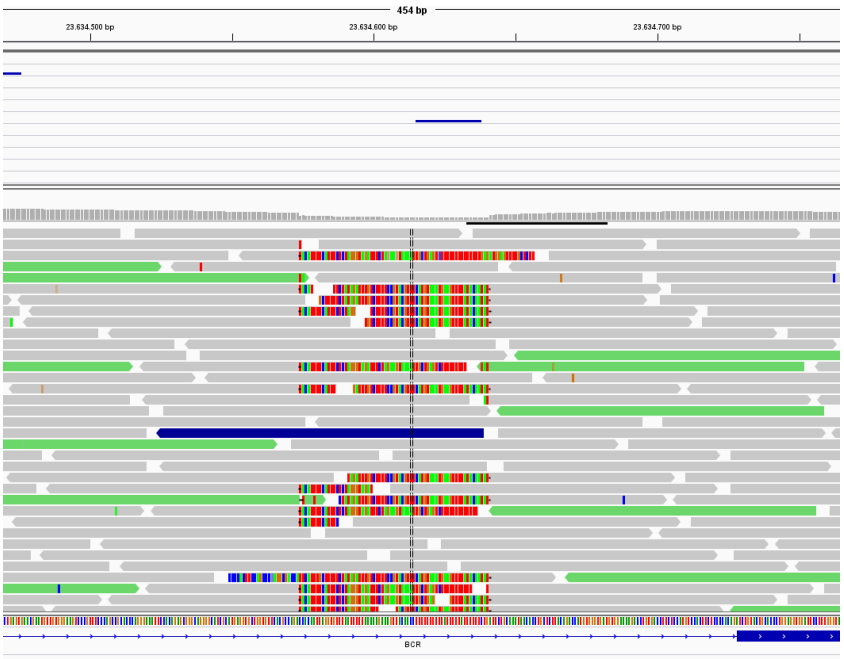



Figure 63. **Sample 10: BP1 and BP2 at BCR**
*The figure shows BP1 and BP2 spotted by R-clipped and L-clipped split reads respectively. These breakpoints are also identified by forward and reverse discordant pair reads (light green colored). At the top the blue box that describe a DNA element of low complexity very close to BP2. At the top the genomic coordinates according to Hg19.*

In addition others two breakpoints spotting the ABL1-BCR fusion segment were found by SVDetect and refined with ClipCrop [Table 21]. BP2 and BP3 identified by ClipCrop were chr22:23634640 and chr9:133696781 respectively [Table 21]. The coordinates of BP2 and BP3 are easily viewable in Figure 63 and Figure 64 respectively. The number of discordant pairs that span both fusions segments was very high, 204 reads supporting the BCR-ABL1 fusion and 98 spotting the ABL1-BCR fusion [Table 19]. Differently the number of split-read, which support both fusions in ABL1, was very low (8 and 32 reads for BCR-ABL1 and ABL1-BCR fusion segment) [Table 20] and [Table 21]. This phenomenon arises from the high density of repeated elements that surround the breakpoints. In particular, both breakpoints that occur in ABL1 region mapped within a L1ME1 (chr9:133696617-133697051) LINE and near to several SINE such as AluSx (chr9:133697052-133697227), AluY(chr9:133697228-133697523) and the AluSx8 (chr9:133696313-133696616) [Table 34][Table 37]. The Figure 64 shows very well how these repeated elements are thickened and how their tightly consecution can lead to a very low capturing. The high number of discordant pair reads that spot the BCR-ABL1 fusion in ABL1 gene can be explained with the efficient capturing of their mates that map in BCR. Indeed, very few clipped-reads (only 8) were able to identify the BCR-ABL1 breakpoint [Table 20]. Regarding the breakpoints found at BCR, BP1 is located in a region free from repeated elements, whereas BP2 falls very near to a DNA region with low complexity [Figure 63]. Sanger sequencing validated the breakpoints coordinates that spot the junction BCR-ABL1 [Table 22]. Concerning the mechanism of DNA break that occurred in this sample we can infer that both chromosomes (22 and 9) underwent to dsDNA break with a partial loss of genetic material: 219 bp at chromosome 9 and 66bp at chromosome 22 [Figure 65]. BP1 coordinate confirm the b2a3 transcript isoform previously detected in RT-PCR [Table 2], indeed it maps between exon 14 and 15 [Table 38].


Figure 64. **Sample 10: Breakpoints in ABL1 region**
*The figure describes the genomic region at which BP3 and BP4 are located. R-clipped and L-clipped split reads spot BP3 and BP4 respectively. These breakpoints are also spotted by reverse and forward discordant pair reads (dark grey). Both BP3 and BP4 reside within a L1ME1 repeated element. This genomic region is highly dense of repeated sequences (AluSx, AluY, AluSc8).*

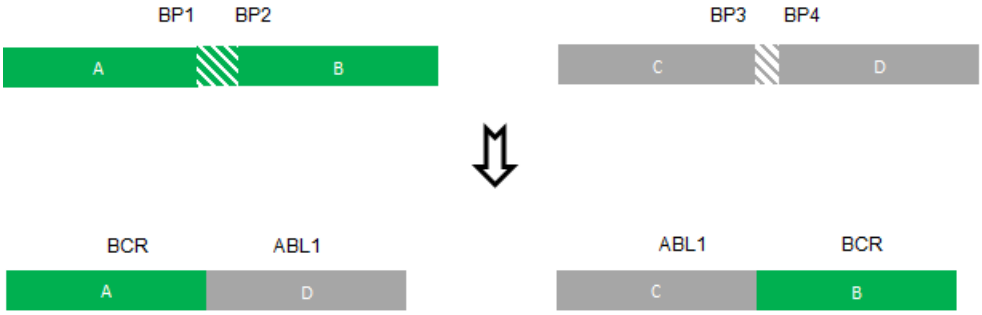



Figure 65. **Sample 10: DNA break**
*The figure shows how BCR and ABL1 are broken and fused in two hybrid products: BCR-ABL1 and ABL1-BCR. In this samples the breaks at chromosome 9 and 22 are blunt-like with loss of nucleotides. Indeed BP1 is upstream BP2 as well as BP3 is upstream BP4.*

## 4.4 Discussion

We have described: 1) the specific capture and enrichment of the genomic region spanning the breakpoints in ABL1 and BCR that produce the chimeric gene BCR-ABL1, 2) the NGS protocol, 3) the bioinformatics analysis used to identify the BCR-ABL1 fusion points. This approach has been applied to a cell line (K562) and to 9 DNAs from CML patients.

Breakpoint identification was successful for K562 cell line and for 8 over 9 CML patients analyzed.

As a general comment, we observed that the identification of BCR-ABL1 fusion points in K562 cell lines is much easier than in CML patients. The average reads coverage, as well as the capturing efficiency is in fact much higher in the cell line compared to that observed in patients. This is primarily due to the fact that in cell lines all cells carry the BCR-ABL1 fusion gene. On the contrary, patient DNAs were obtained from peripheral blood (PB) at disease onset. At this stage the % of cells carrying the translocation is variable and normal cells are also present. The lower is the % of BCR-ABL1$^{+}$ cells, the higher is the difficulty to identify the translocation event.

In one patient (Patient 8), the precise identification of BCR-ABL1 fusion points failed and we could only detect the coordinates of the reciprocal ABL1-BCR fusion. Based on these coordinates, we could restrict the genomic region to set up a PCR assay to amplify BCR-ABL1 and sequence this amplicon in Sanger. After having resolved the breakpoints in Sanger sequencing, we could establish that the initial failure in breakpoints detection was mainly attributable to the very low number of reads covering these region, due to their low complexity (long stretches of A and T at both portions of the fusion) that can affect sequencing performance.

Sanger sequencing validated the identified BCR-ABL1 fusion coordinates in all samples. The bioinformatics pipeline implemented in this PhD project allowed us to identify the BCR-ABL1 fusion points with an accuracy of 1.01 ± 0.44 bp in BCR and 1.22 ± 2.22 bp in ABL1. In fact, in most samples the

breakpoints coordinates identified by NGS were identical to those obtained using Sanger method.

As supplementary analysis besides the identification of BCR-ABL1 breakpoints in each sample, we used NGS data to infer additional information on DNA breaks and rearrangement events that are described below.

Using the pipeline, the reciprocal ABL1-BCR fusion was found in 6 out of 9 samples (excluding the cell line) and was detected by curated manual analysis in patients 2 and 7 in which the number of supporting reads did not pass the software thresholds (sample 2) and a DNA insertion occurred between the ABL1 and BCR segments (sample 7).

The analysis of sequencing data in each patient suggests possible explanation on how the breaks occur at chromosomes level and why a variable number of breakpoints are detected on chromosome 9 and 22. In fact, the analysis has shown that most patients present four breakpoints, 2 on chromosome 22 and 2 on chromosome 9. What we expected was to find a single breakpoint in each chromosome. The presence of two breakpoints in both chromosomes could be explained either by the loss of genomic material or by a protruding break in which the two strands are differently broken. Taking into account this concept we could infer that in patients 2, 3 and 10 the dsDNA break was blunt and followed by a DNA loss. The loss of genomic DNA was observed thanks to the coordinates of both the fusion products (BCR-ABL1 and ABL1-BCR) that are produced by the reciprocal translocation t(9;22). The difference in coordinates between the breakpoints that spot the two fusion products in each chromosome has shown that the loss of DNA is variable ranging from few bases to thousands. The analysis of these breakpoints suggests that initially a single dsDNA break occurred both in chromosome 9 and chromosome 22 and this was followed by a DNA loss at the ends of the broken chromosomes.

The analysis of patients 4 and 5 has shown that the DNA break can be blunt without loss of DNA at one chromosome end and protruding in the other chromosome. In particular, in patient 4 the DNA break on chromosome 22 (BCR) is protruding producing two breakpoints coordinates in which the BP2 (which spots the ABL1-BCR fusion) is upstream BP1 (which spots the BCR-ABL1 fusion). On chromosome 9 the break is blunt without loss of genomic DNA. On the contrary, patient 5 has a protruding DNA break on chromosome 9 (ABL1) producing two breakpoints coordinates in which BP3 (which spot the ABL1-BCR fusion) is downstream BP4.

Patient 6 is characterized by protruding break at ABL1 and blunt break at BCR with loss of DNA at the ends.

The analysis of patient 7 was peculiar and has demonstrated the complexity of DNA rearrangement that follows chromosome break. We established that a protruding DNA break occurred at BCR, whereas a DNA break with loss of DNA occurred at ABL1. In particular, in this patient both BCR-ABL1 and ABL1-BCR fusion products have a DNA insertion at the juntion. BCR-ABL1 fusion segment includes a reverse inverted fragment which is located upstream BP3, whereas the ABL1-BCR fusion segment includes a reverse inverted portion of the FOXRED2 gene located downstream BP1 and BP2 on chromosome 22. Finally the analysis of patient 9 identified the BCR-ABL1 fusion product but could not identify the reciprocal ABL1-BCR. The results of the global analysis have shown that many breakpoints occur within or very close to repeated elements such as SINEs (Alu family) or LINEs (L1 family). In particular, in samples 4 and 5, the breakpoints in BCR occur in m-BCR and µ-BCR respectively and are located in regions that lack of repeated elements. These breakpoints derive from blunt DNA break without loss of nucleotides. On the contrary, breakpoints in samples 6 e 7 that could derive from protruding DNA breaks are located close to repeat elements. A similar feature was detected in

samples 2,3 and 10 in which blunt DNA breaks were followed by nucleotides loss. However it has to be noted that the sample size of the present study does not allow to establish if sequence itself can really affect the type of break (blunt or protruding) and nucleotide loss.

# 5. CONCLUSIONS

Next-generation sequencing (NGS) is gradually making its way into clinical laboratories due to the decreasing of costs.

NGS can be used to identify both sequence and structural variants.

The work described in this thesis has demonstrated that NGS technology can be applied to monitor minimal residual disease (MRD) in CML. The workflow implemented to analyze NGS data allowed us to identify breakpoints that give rise to BCR-ABL1 oncogene, with single nucleotide accuracy and in a genomic landscape featured by a high density of repeated sequences.

This is important because the initial idea was to use breakpoints identification with NGS as the starting point for the design of a patient specific qPCR assay to be used in patient follow up during IM treatment.

Targeted next generation sequencing has been previously used to describe translocations in leukemia. In particular, Duncavage et al [113], focused on the identification of gene mutations such as translocations, SNVs and INDELs in five leukemia cell lines including K562,NB4,OCI-AML3, kasumi-1 and MV4-11. The authors also applied this method to identify the t(9;11) in a bone marrow sample from a patient affected by AML. Shibata et al, [114] developed Anchor chrom-PET technique to capture the BCR-ABL1 translocation in K562 cell line and in Ph+ and Ph- patients. In this study they targeted only M-BCR region by using a custom capture and enrichment protocol not applicable for clinical purpose. Finally Chen et al. [115], performed shotgun sequencing on flow sorted derivative chromosomes using NGS in lymphoblastoid cell lines.

The present PhD work has innovative character because is the first NGS approach specifically designed to target all the potential breakpoints that give rise to the BCR-ABL1 fusion gene.

The method developed in our lab can detect patient-specific breakpoints simultaneously in multiple samples.

Moreover, we demonstrated that we could spot small insertions / deletions that can occur during the translocation event. These secondary rearrangements are undetectable by cytogenetic techniques such as CBA or FISH. The additional original character of this work is the fine analysis of the BCR-ABL1 fusion points along with their reciprocal (ABL1-BCR), to infer DNA break mechanism in CML using NGS data. Different features of DNA break in CML have been observed in this analysis. This could offer new clues to understand the mechanism of the translocation. A possible future analysis could focus on the regions surrounding the breakpoints in order to spot the presence of recurrent DNA patterns that can promote DNA break or the translocation event.

In order to increase the statistical power of the study we are collecting DNAs of further 40 CML samples in major molecular response (no BCR-ABL1 transcript detected for two years). We plan to apply our capturing protocol in these patients, identify breakpoints, and design the patient specific assay for DNA quantification in real time during the follow-up.

DNA quantification in these samples will help to take clinical decisions on disease treatment.

## BIBLIOGRAPHY


[1] Sanger, F., Nicklen, S. Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.U.S.A*, 74, 12 (1977), 5463-5467.

[2] Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. Smith, M. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265, 5596 (1977), 687-695.

[3] Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. Hood, L.E. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321, 6071 (1986), 674-679.

[4] Marziali, A. Akeson, M. New DNA sequencing methods. *Annual Review of Biomedical Engineering*, 3 (2001), 195-223.

[5] Shendure, J. Ji, H. Next-generation DNA sequencing. *Nat Biotech*, 26, 10 (2008), 1135-1145.

[6] Venter, J.C., et al. The sequence of the human genome. *Science*, 291, 5507 (2001), 1304-1351.

[7] Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature*, 409, 6822 (2001), 860-921.

[8] Collins, F.S., Lander, E.S., Rogers, J. Waterson, R.H. Finishing the euchromatic sequence of the human genome", *Nature*, 431, 7011 (2004), 931-945.

[9] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437 (2005), 376–380.

[10] Bentley, D.R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 7218 (2008), 53-59.

[11] Sambrook J, Russell DW. Fragmentation of DNA by nebulization. *CSH Protoc*, 4 (2006).

[12] Fedurco, M., Romieu, A., Williams, S., Lawrence, I. Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, 34, 3 (2006).

[13] Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P. Kawashima E. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research*, 28, 20 (2000).

[14] Mamanova, L. , A. J. Coffey , C. E. Scott , I. Kozareva, E. H.

Turner, A. Kumar , E. Howard , et al. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7 (2010), 111-118.

[15] Chandra Shekhar Pareek, Rafal Smoczynski, Andrzej Tretyn. Sequencing technologies and genome sequencing. *J Appl Genetics*, 52 (2011), 413-435.

[16] Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J. M. L. M., van der Zaag, P. J., Franke, A., Brookes, A. J. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in functional genomics*, 10, 6 (2011), 374–86.

[17] Wang, W., Wei, Z., Lam, T.-W., Wang, J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Scientific reports*, 1, 55 (2011).

[18] Smit, AFA, Hubley, R Green, P. RepeatMasker Open-3.0 (2010).

[19] EB, Hook. Unbalanced Robertsonian translocations associated with Down's syndrome or Patau's syndrome: chromosome subtype, proportion inherited, mutation rates, and sex ratio. *Hum Genet*, 59, 3 (1981), 235-9.

[20] Povirk, L. F. Biochemical mechanisms of chromosomal translocations resulting from DNA double-strand breaks. *DNA repair*, 9, 5 (2006), 1199–1212.

[21] Score, J., Calasanz, M. J., Ottman, O., Pane, F., Yeh, R. F., Sobrinho-Simões, M. a, Grand, F. H. Analysis of genomic breakpoints in p190 and p210 BCR-ABL indicate distinct mechanisms of formation. *Leukemia*, 24, 10 (2010), 1742–50.

[22] Mitelman, F., Johansson, B., Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature reviews Cancer*, 7, 4 (2007), 233-45.

[23] Roth, D. B., Porter, T. N. Wilson, J. H. Mechanisms of nonhomologous recombination in mammalian cells. *Mol. Cell. Biol*, 5 (1985), 2599–2607.

[24] Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem*, 79 (2010), 181-211.

[25] Simsek, D. Jasin, M. Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4-ligase IV during chromosomal translocation formation. *Nature Struct. Mol. Biol*, 17 (2010), 410-416.

[26] Medvedev P., Stanciu M., Brudno M. Computational methods for discovering structural variation with next-generation sequencing.

*Nature Methods Supplement*, 6, 11s (2009), 13-20.

[27] Le Scouarnec, S., Gribble, S. M Heredity. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. *Heredity*, 108, 1 (2012), 75-85.

[28] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF et al. Paired- end mapping reveals extensive structural variation in the human genome. *Science*, 318 (2007), 420-426.

[29] Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., Gerstein, M. Identification of genomic indels and structural variations using split reads. *BMC genomics*, 12, 375 (2011).

[30] Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, 21 (2011), 974-984.

[31] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6, 9 (2009), 677-81.

[32] Zeitouni B., Boeva V., Janoueix-Lerosey I., Loeillet S., Legoix-né P., Nicolas A., Delattre O., Barillot E. SVDetect: a tool to identify genomic structural variations from pair-end and mate-pair sequencing data. *Bioinformatics*, 26, 15 (2010), 1895-18.

[33] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25, 21 (2009), 2865–71.

[34] Abel, H. J., Duncavage, E. J., Becker, N., Armstrong, J. R., Magrini, V. J., Pfeifer, J. D. SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics*, 26, 21 (2010), 2684-8.

[35] Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics*, 12, Suppl 14:S7 (December 14, 2011).

[36] R., Li H. and Durbin. Fast and accurate short read alignment with Burrows-Wheeler. *Bioinformatics*, 25 (2009), 1754-60.

[37] Deininger, M. W. N., Goldman, J. M., Melo, J. V. The molecular biology of chronic myeloid leukemia. *Blood*, 96, 10 (2000), 3343–3356.

[38] Faderl S., Talpaz M., Estrov Z., O'Brien S., Kurzrock R., Kantarjian

HM. The biology of chronic myeloid leukemia. *New England Journal of Medicine*, 341, 3 (1999), 164-72.

[39] Howlader N, Noone AM, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA. SEER Cancer Statistics Review 1975-2010 (2013).

[40] Cortes, J., Kantarjian, H. How I treat newly diagnosed chronic phase CML. *Blood*, 120, 7 (2012), 1390–7.

[41] Radich, J. P. How I monitor residual disease in chronic myeloid leukemia. *Blood*, 114, 16 (2009), 3376–81.

[42] Gibson, J., Iland, H. J., Larsen, S. R., Brown, C. M. S., Joshua, D. E. Leukaemias into the 21st century. Part 2: the chronic leukaemias. *Internal medicine journal*, 45, 3 (2013), 484–94.

[43] Cortes E.J., Silver R,T., Kantarjian H. Chronic Myeloid Leukemia, Cancer Management: 14th ed. *Cancer Network* (2011).

[44] Akimichi Ohsaka, Shigeo Shiina, Masaru Kobayashi, Hideyuki Kudo and Ryuji Kawaguchi. Philadelphia Chromosome-Positive Chronic Myeloid Leukemia Expressing p190 BCR-ABL. *Internal Medicine*, 41, 12 (2002), 1183–1187.

[45] Baccarani M., Pileri S., Steegmann J.L., Muller M., Soverini S., Dreyling M., “Chronic myeloid leukemia:ESMO clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 23, Supplement 7 (2012), 72-77.

[46] Albano F., Anelli L., Zagaria A., Coccaro N., Casieri P., Russo Rossi A., Vicari L., Liso V., Rocchi M., Specchia G. Non random distribution of genomic features in breakpoint regions involved in chronic myeloid leukemia cases with variant t(9;22) or additional chromosomal rearrangements. *Molecular Cancer*, 9, 120 (2010).

[47] Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of myeloid neoplasm. *Blood*, 100 (2002), 2292-2302.

[48] Cortes J.E., Talpaz M., O’Brien S., Faderl S., Garcia-Manero G., Ferrajoli A., Verstovsek S., Rios M.B., Shan J., Kantarjian H.M. Staging of Chronic Myeloid Leukemia in the Imatinib era. *Cancer*, 106, 6 (2006), 1306-1315.

[49] Sokal JE, Baccarani M, Russo D, et al. Staging and prognosis in chronic myelogenous leukemia. *Semin Hemato*, 25, 1 (1988), 49-61.

[50] Hasford J, Pfirrmann M, Hehlmann R et al. A new prognostic score for survival of patient with chronic myeloid leukemia treated with interferon alfa. *J Natl Cancer Inst* (1998), 850-858.

[51] Hasford, J., Baccarani, M., Hoffmann, V., Guilhot, J., Saussele, S., Rosti, G., Hehlmann, R. Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with CML on Imatinib treatment: the EUTOS score. *Blood*, 18 (2011), 686-692.

[52] B.J. Druker, C.L. Sawyers, H. Kantarjian, D.J. Resta, S.F. Reese, J.M. Ford, R. Capdeville, M. Talpaz. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronicmyeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med* (2001), 1038–1042.

[53] Renaud Capdeville, Elisabeth Buchdunger, Juerg Zimmermann Alex Matter. Glivec (STI571, Imatinib), a rationally developed, targeted anticancer drug. *Nature Reviews Drug Discovery* (2002), 493-502.

[54] Saussele, S., Pfirrmann, M. Clinical trials in chronic myeloid leukemia. *Current hematologic malignancy reports*, 7, 2 (2012), 109–15.

[55] Chandra, H. S., Heisterkamp, N. C., Hungerford, A., Morrissette, J. J. D., Nowell, P. C., Rowley, J. D., Testa, J. R. Philadelphia Chromosome Symposium: commemoration of the 50th anniversary of the discovery of the Ph chromosome. *Cancer genetics* (2011), 171-9.

[56] O'Hare, T., Eide, C. a, Deininger, M. W. N. Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood*, 110, 7 (2007), 2242–9.

[57] Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nature reviews Cancer*, 5, 3 (2005), 172–83.

[58] Rea, D., Rousselot, P., Guilhot, J., Guilhot, F., Mahon, F.-X. Curing chronic myeloid leukemia. *Current hematologic malignancy reports*, 7, 2 (2012), 103-8.

[59] Baccarani M., Deininger M.W., Rosti G., Hochhaus A., Soverini S., Apperley J.F., Cervantes F., Clark R.E., Cortes J.E., Guilhot F., Hjorth-Hansen H., Hughes T.P., Kantarjian H.M., Kim D.W., Larson R.A., Lipton J.H., Mahon F.X., Martinelli G., Mayer J. European LeukemiaNet recommendations for the management of chronic myeloid leukemia: 2013. *Blood*, 122, 6 (2013), 872-84.

[60] Mattarucchi, E., Spinelli, O., Rambaldi, A., Pasquali, F., Lo Curto, F., Campiotti, L., Porta, G. Molecular monitoring of residual disease in chronic myeloid leukemia by genomic DNA compared with conventional mRNA analysis ((2009)), 482-7.

[61] Kim, Y.-J., Kim, D.-W., Lee, S., Kim, H.-J., Kim, Y.-L., Hwang, J.-Y.,

Kim, C.-C. Comprehensive comparison of FISH, RT-PCR, and RQ-PCR for monitoring the BCR-ABL gene after hematopoietic stem cell transplantation in CML. *European journal of haematology*, 68, 5 (2002), 272-80.

[62] Zhang, J. G., Lin, F., Chase, a, Goldman, J. M., Cross, N. C. Comparison of genomic DNA and cDNA for detection of residual disease after treatment of chronic myeloid leukemia with allogeneic bone marrow transplantation. *Blood*, 87, 6 (1996), 2588–93.

[63] Bartley, P. a, Ross, D. M., Latham, S., Martin-Harris, M. H., Budgen, B., Wilczek, V., Morley, A. a. Sensitive detection and quantification of minimal residual disease in chronic myeloid leukaemia using nested quantitative PCR for BCR-ABL DNA. *International journal of laboratory hematology*, 32, 6 (2010), e222-8.

[64] Kurzrock R., Kantarjian H.,M., Drunker B.J., Talpaz M. Philadelphia Chromosome-Positive leukemias: from basic mechanism to molecular therapeutics. *Ann Intern Med*, 138 (2003), 819-830.

[65] Paul Flicek et al. Ensembl 2013. *Nucleic Acids Research*, 41, D48-D55 (2013).

[66] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res*, 12, 6 (2002), 996-1006.

[67] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue) (2004), D493-6.

[68] Laurent, E., Talpaz, M., Wetzler, M., and Kurzrock, R. Cytoplasmic and nuclear localization of the 130 and 160 kDa Bcr proteins. *Leukemia*, 14 (2000), 1892-1897.

[69] Laurent, E., Talpaz, M., Kantarjian, H. The BCR Gene and Philadelphia Chromosome-positive Leukemogenesis The BCR Gene and Philadelphia Chromosome-positive Leukemogenesis. *Cancer Res*, 61 (2001), 2343–2355.

[70] Quintás-Cardama, A., Cortes, J. Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood*, 113, 8 (2009), 1619-30.

[71] Consortium, The UniProt. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*, 41 (2013), D43-D47.

[72] De Braekeleer, E., Douet-Guilbert, N., Rowe, D., Bown, N., Morel, F., Berthou, C., De Braekeleer, M. ABL1 fusion genes in hematological malignancies: a review. *European journal of haematology*, 86, 5 (2011), 361–71.

[73] Greuber, E. K., Smith-Pearson, P., Wang, J., Pendergast, A. M.

Role of ABL family kinases in cancer: from leukaemia to solid tumours. *Nature reviews. Cancer*, 13, 8 (2013), 559–71.

[74] Raghavan, S. C., Lieber, M. R. DNA structures at chromosomal translocation sites. *Bioessay: news and reviews in molecular, cellular and developmental biology*, 28, 5 (2006).

[75] Greuber, E. K., Smith-Pearson, P., Wang, J., Pendergast, A. M. Role of ABL family kinases in cancer: from leukaemia to solid tumours. *Nature Reviews. Cancer*, 13, 8 (2013), 559–71.

[76] Panjarian, S., Iacob, R. E., Chen, S., Engen, J. R., Smithgall, T. E. Structure and dynamic regulation of Abl kinases. *The Journal of biological chemistry*, 288, 8 (2013), 5443–50.

[77] Hantschel, O., Nagar, B., Guettler, S., Kretzschmar, J., Dorey, K., Kuriyan, J., and Superti-Furga, G. A myristoyl/phosphotyrosine switch regulates c-Abl. *Cell*, 112 (2003), 845–857.

[78] Groffen J, Stephenson Jr, Heisterkamp N, de Klein A, Bartram CR, Grosveld G. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell*, 36 (1984), 93-99.

[79] Hermans A, Heisterkamp N, von Linden M, van Baal S, Meijer D, van der Plas D et al. Unique fusion of bcr and c-abl genes in Philadelphia chromosome positive acute lymphoblastic leukemia. *Cell*, 51 (1987), 33-40.

[80] Pane F, Frigeri F, Sindona M, Luciano L, Ferrara F, Cimino R et al. Neutrophilic-chronic myeloid leukemia: a distinct disease with a specific molecular marker (BCR-ABL with C3/a2 junction). *Blood*, 88 (1996), 2410-2414.

[81] Krumbholz, M., Karl, M., Tauer, J. T., Thiede, C., Rascher, W., Suttorp, M., Metzler, M. Genomic BCR - ABL1 Breakpoints in Pediatric Chronic Myeloid Leukemia (July 2012), 1045–1053.

[82] Bennour, A., Ouahchi, I., Achour, B., Zaier, M., Youssef, Y. Ben, Khelif, A., Sennana, H. Analysis of the clinico-hematological relevance of the breakpoint location within M-BCR in chronic myeloid leukemia. *Medical oncology*, 30, 1 (2013), 348.

[83] Jones D, Luthra R, Cortes J, Thomas D, O'Brien S, Bueso-Ramos C, Hai S, Ravandi F, de Lima M, Kantarjian H, Jorgensen JL. BCR-ABL fusion transcript types and levels and their interaction with secondary genetic changes in determining the phenotype of Philadelphia chromosome-positive leukemias. *Blood*, 112, 13 (2008), 5190-2.

[84] Yaghmaie, M., Ghaffari, S. H., Ghavamzadeh, A., Alimoghaddam, K., Jahani, M., Mousavi, S.-A., Bibordi, I. Frequency of BCR-ABL

fusion transcripts in Iranian patients with chronic myeloid leukemia. *Archives of Iranian medicine*, 11, 3 (2008), 247-251.

[85] Arana-Trejo, R. M., Ruíz Sánchez, E., Ignacio-Ibarra, G., Báez de la Fuente, E., Garces, O., Gómez Morales, E., Kofman, S. BCR/ABL p210, p190 and p230 fusion genes in 250 Mexican patients with chronic myeloid leukaemia. *Clinical and laboratory haematology*, 24, 3 (2002), 145–50.

[86] Melo, J.V. The diversity of BCR-ABL fusion proteins and their relationship to leukemia phenotype. *Blood*, 88 (1996), 2375–2384.

[87] Weerkamp, F., Dekking, E., Ng, Y. Y., van der Velden, V. H. J., Wai, H., Böttcher, S., van Dongen, J. J. M. Flow cytometric immunobead assay for the detection of BCR-ABL fusion proteins in leukemia patients. *Leukemia*, 23, 6 (2009), 1106–17.

[88] Chen, Y., Peng, C., Li, D., Li, S. Molecular and cellular bases of chronic myeloid leukemia. *Protein cell*, 1, 2 (2010), 124–32.

[89] Cilloni, D., Saglio, G. Molecular pathways: BCR-ABL. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18, 4 (2012), 930–7.

[90] Yoshida, C., Melo, J. V. Biology of Chronic Myeloid Leukemia and Possible Therapeutic Approaches to Imatinib-Resistant Disease. *International Journal of Hematology*, 79, 5, 420–433.

[91] Burmeister, T., Gröger, D., Kühn, A., Hoelzer, D., Thiel, E., Reinhardt, R. Fine structure of translocation breakpoints within the major breakpoint region in BCR-ABL1-positive leukemias (2011), 1131–7.

[92] Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, 22, 2 (2006), 134-41.

[93] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[94] Schmieder R. and Edwards R., Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27 (2011), 863-864.

[95] Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13, 1 (2011), 36-46.

[96] Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Sun, S. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData mining*, 5, 1 (2012), 6.

[97] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8, 3 (1998), 186–194.

[98] R., Li H. and Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25 (2009), 1754-60.

[99] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E.E. Eichler, S.C.Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods*, 7, 8, 576-7.

[100] http://hgdownload.cse.ucsc.edu/downloads.html#human

[101] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20, 1297 (2010), 1297-303.

[102] DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43 (2011), 491-498.

[103] WJ., Kent. BLAT--the BLAST-like alignment tool. *Genome Res*, 12, 4 (2002), 656-64.

[104] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 16 (2009), 2078–9.

[105] Rumble, S. M., Lacroute, P., Dalca, A. V, Fiume, M., Sidow, A., Brudno, M. SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology*, 5, 5 (2009).

[106] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* , 215 (1990), 403-410.

[107] Sindi, S., Helman, E., Bashir, A., Raphael, B. J. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25, 12 (2009), 222–30.

[108] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6, 9 (2009), 677-81.

[109] Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res*, 19 (2009), 1639–1645.

[110] Mattarucchi, E., Guerini, V., Rambaldi, A., Campiotti, L., Venco, A., Pasquali, F., Porta, G. Microhomologies and Interspersed Repeat Elements at Genomic Breakpoints in Chronic Myeloid Leukemia. *Genes, Chromosomes, Cancer*, 47 (2008), 625–632.

[111] Loncarevic IF, Römer J, Starke H, Heller A, Bleck C, Ziegler M, Fiedler W, Liehr T, Clement JH, Claussen U. Heterogenic molecular basis for loss of ABL1-BCR transcription: deletions in der(9)t(9;22) and variants of standard t(9;22) in BCR-ABL1-positive chronic myeloid leukemia. *Genes Chromosomes Cancer*, 34, 2 (2002), 193-200.

[112] De Melo, V. a S., Milojkovic, D., Marin, D., Apperley, J. F., Nacheva, E. P., Reid, A. G. Deletions adjacent to BCR and ABL1 breakpoints occur in a substantial minority of chronic myeloid leukemia patients with masked Philadelphia rearrangements. *Cancer genetics and cytogenetics*, 182, 2 (2008), 111-5.

[113] Duncavage, E. J., Abel, H. J., Szankasi, P., Kelley, T. W., Pfeifer, J. D. Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia. *Modern pathology*, 25, 6 (2012), 795-804.

[114] Shibata Y, Malhotra A, Dutta A. Detection of DNA fusion junctions for BCR-ABL1 translocations by Anchored ChromPET. *Genome Medicine*, 2, 70 (2010).

[115] Chen, W., Kalscheuer, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M. H., Ropers, H. H. Mapping translocation breakpoints by next-generation sequencing. *Genome Research*, 18 (2008), 1143–1149.

[116] http://picard.sourceforge.net

## APPENDIX

| Breakpoint | Chromosome | Coordinate | Annotation | N° of probes |
|---|---|---|---|---|
| **BP1** | 22 | 23632041 | intron 13-14 | 5 |
| **BP2** | 22 | 23631917 | intron 13-14 | 5 |
| **BP3** | 22 | 23632170 | intron 13-14 | 5 |
| **BP4** | 22 | 23631956 | intron 13-14 | 5 |
| **BP5** | 22 | 23632256 | intron 13-14 | 5 |
| **BP6** | 22 | 23632142 | intron 13-14 | 5 |
| **BP7** | 22 | 23632501 | intron 13-14 | 5 |
| **BP8** | 22 | 23632360 | intron 13-14 | 5 |
| **BP9** | 22 | 23632133 | intron 13-14 | 5 |
| **BP10** | 22 | 23631981 | intron 13-14 | 5 |
| **BP11** | 22 | 23632365 | intron 13-14 | 5 |
| **BP12** | 22 | 23633068 | intron 14-15 | 2 |
| **BP13** | 22 | 23633696 | intron 14-15 | 5 |
| **BP14** | 22 | 23632845 | intron 14-15 | 5 |
| **BP15** | 22 | 23632443 | intron 13-14 | 5 |
| **BP16** | 22 | 23633595 | intron 14-15 | 5 |
| **BP17** | 22 | 23633803 | intron 14-15 | 5 |
| **BP18** | 22 | 23632580 | exon 14 | 5 |
| **BP19** | 22 | 23633263 | intron 14-15 | not covered |
| **BP20** | 22 | 23632883 | intron 14-15 | 5 |
| **BP21** | 22 | 23632682 | intron 14-15 | 5 |
| **BP22** | 22 | 23634158 | intron 14-15 | 5 |
| **BP23** | 22 | 23632742 | intron 14-15 | 5 |
| **BP24** | 22 | 23631916 | intron 13-14 | 5 |
| **BP25** | 22 | 23633377 | intron 14-15 | not covered |
| **BP26** | 22 | 23631838 | intron 13-14 | 5 |
| **BP27** | 22 | 23632196 | intron 13-14 | 5 |

Table 26. **Theoretical capturing evaluation of known breakpoints**
*The table shows the testing of the best probe set for M-BCR region in order to evaluate if this probe set can theoretically cover 27 breakpoints already detected by Prof. G. Porta. Only two breakpoints were not covered because they fall within an AluSx1 element.*

| Sample | Target | Total cov | Average cov | Q1 | median | Q3 | 1X | 20X | 50X | 100X |
|---|---|---|---|---|---|---|---|---|---|---|
| **02a** | ABL1 | 19021875 | 124.15 | 63 | 119 | 180 | 98.7 | 91 | 80 | 58 |
| **02b** | ABL1 | 18362295 | 119.85 | 60 | 113 | 172 | 98.5 | 90.7 | 78.9 | 55.7 |
| **02c** | ABL1 | 18281256 | 119.32 | 60 | 114 | 171 | 98.6 | 90.6 | 79 | 56.6 |
| **02d** | ABL1 | 18195013 | 118.75 | 59 | 114 | 172 | 98.5 | 90.4 | 78.9 | 55.7 |
| **03a** | ABL1 | 15268209 | 99.65 | 44 | 87 | 144 | 98.5 | 88.2 | 70.9 | 43.1 |
| **03b** | ABL1 | 16024382 | 104.59 | 47 | 92 | 151 | 98.6 | 88.8 | 72.9 | 45.5 |
| **03c** | ABL1 | 15823051 | 103.27 | 46 | 89 | 148 | 98.4 | 88.4 | 71.9 | 44.9 |
| **03d** | ABL1 | 15462313 | 100.92 | 45 | 89 | 145 | 98.4 | 88.6 | 71.6 | 43.4 |
| **04a** | ABL1 | 16996561 | 110.93 | 57 | 103 | 155 | 98.6 | 90.9 | 78.3 | 51.5 |
| **04b** | ABL1 | 17956429 | 117.2 | 62 | 109 | 166 | 98.5 | 91.5 | 80.1 | 54.4 |
| **04c** | ABL1 | 17566485 | 114.65 | 61 | 106 | 162 | 98.3 | 91.2 | 79.8 | 53 |
| **04d** | ABL1 | 17221366 | 112.4 | 59 | 105 | 159 | 98.5 | 91 | 78.9 | 52.1 |
| **05a** | ABL1 | 12363399 | 80.69 | 40 | 71 | 111 | 98.5 | 87.5 | 66.7 | 29.8 |
| **05b** | ABL1 | 12994710 | 84.81 | 42 | 74 | 117 | 98.4 | 88.2 | 68.5 | 33.1 |
| **05c** | ABL1 | 12843026 | 83.82 | 42 | 73 | 116 | 98.3 | 88.1 | 68.3 | 32.9 |
| **05d** | ABL1 | 12645128 | 82.53 | 41 | 72 | 113 | 98.2 | 87.9 | 67.8 | 31.5 |
| **06a** | ABL1 | 12817986 | 83.66 | 41 | 71 | 112 | 99 | 89.6 | 67.5 | 30.8 |
| **06b** | ABL1 | 12473458 | 81.41 | 41 | 70 | 108 | 98.4 | 89.1 | 66.3 | 28.8 |
| **06c** | ABL1 | 12499065 | 81.58 | 40 | 70 | 109 | 99 | 88.9 | 66.1 | 29.4 |
| **06d** | ABL1 | 12472852 | 81.41 | 41 | 69 | 110 | 98.7 | 89.1 | 66.3 | 29.7 |
| **07a** | ABL1 | 17342929 | 113.19 | 56 | 108 | 164 | 98.5 | 90.2 | 77.4 | 53.7 |
| **07b** | ABL1 | 16638946 | 108.6 | 52 | 105 | 158 | 98 | 89.7 | 75.6 | 52 |
| **07c** | ABL1 | 16738957 | 109.25 | 52 | 106 | 159 | 98.1 | 89.6 | 75.7 | 52.3 |
| **07d** | ABL1 | 16739703 | 109.26 | 52 | 106 | 160 | 98.4 | 89.4 | 75.7 | 52.5 |
| **08a** | ABL1 | 16058491 | 104.81 | 51 | 96 | 148 | 98.2 | 89.4 | 75.5 | 47.2 |
| **08b** | ABL1 | 15413341 | 100.6 | 51 | 91 | 142 | 98.5 | 88.9 | 75 | 44.6 |
| **08c** | ABL1 | 15381300 | 100.39 | 49 | 92 | 142 | 98.6 | 88.9 | 74 | 45.3 |
| **08d** | ABL1 | 15289529 | 99.79 | 48 | 91 | 141 | 98.6 | 88.7 | 73.7 | 45 |
| **09a** | ABL1 | 14592066 | 95.24 | 52 | 88 | 130 | 98.9 | 90.9 | 76 | 42.3 |
| **09b** | ABL1 | 15353248 | 100.21 | 55 | 92 | 138 | 99 | 90.9 | 77.9 | 45 |
| **09c** | ABL1 | 15046751 | 98.21 | 53 | 90 | 134 | 98.4 | 90.7 | 76.7 | 43.7 |
| **09d** | ABL1 | 14682601 | 95.83 | 53 | 88 | 132 | 98.5 | 90.4 | 76.4 | 42.7 |
| **10a** | ABL1 | 15505023 | 101.2 | 50 | 89 | 136 | 98.6 | 89.7 | 74.4 | 44.2 |
| **10b** | ABL1 | 16400009 | 107.04 | 53 | 94 | 144 | 98.8 | 90.3 | 76.3 | 46.5 |
| **10c** | ABL1 | 16070544 | 104.89 | 51 | 92 | 143 | 98.7 | 90.2 | 75.1 | 45.2 |
| **10d** | ABL1 | 15801952 | 103.14 | 51 | 91 | 140 | 98.9 | 90.1 | 75.1 | 44.6 |

Table 27. **ABL1 lane-level capturing and coverage**

*The table shows the coverage and capturing statistics related to ABL1 targeted region across lanes. From left to right column: sample and lane, region name, number of bp that cover the region, average coverage, coverage at 1$^{st}$ quartile (Q1), median coverage, coverage at 3$^{rd}$ quartile (Q3), percentage of targeted region captured with an average coverage greater or equal to 1X,20X,50X and 100X. Data included in this table are plotted in* **Figure 20**.

| Sample | Target | Total cov | Average cov | Q1 | median | Q3 | 1X | 20X | 50X | 100X |
|---|---|---|---|---|---|---|---|---|---|---|
| **02a** | BCR-MAJOR | 1387439 | 154.19 | 123 | 155 | 186 | 100 | 99.8 | 98.7 | 90.1 |
| **02b** | BCR-MAJOR | 1199975 | 133.36 | 99 | 132 | 168 | 100 | 99.9 | 97.7 | 74.1 |
| **02c** | BCR-MAJOR | 1224217 | 136.05 | 102 | 133 | 169 | 100 | 100 | 98.5 | 76.4 |
| **02d** | BCR-MAJOR | 1201963 | 133.58 | 101 | 135 | 165 | 100 | 99.9 | 97.6 | 75.5 |
| **03a** | BCR-MAJOR | 1204656 | 133.88 | 96 | 132 | 171 | 100 | 100 | 97.1 | 73.1 |
| **03b** | BCR-MAJOR | 1256548 | 139.65 | 102 | 141 | 178 | 100 | 100 | 97.4 | 75.9 |
| **03c** | BCR-MAJOR | 1278401 | 142.08 | 105 | 145 | 176 | 100 | 99.8 | 97 | 76.7 |
| **03d** | BCR-MAJOR | 1233997 | 137.14 | 101 | 140 | 176 | 100 | 99.8 | 95.8 | 75.4 |
| **04a** | BCR-MAJOR | 1423376 | 158.19 | 129 | 158 | 188 | 100 | 100 | 98.6 | 88.6 |
| **04b** | BCR-MAJOR | 1479552 | 164.43 | 130 | 162 | 197 | 100 | 100 | 98.2 | 87.5 |
| **04c** | BCR-MAJOR | 1481600 | 164.66 | 133 | 162 | 201 | 100 | 100 | 98.4 | 90.3 |
| **04d** | BCR-MAJOR | 1439454 | 159.97 | 130 | 162 | 190 | 100 | 100 | 98.7 | 91 |
| **05a** | BCR-MAJOR | 1679913 | 186.7 | 150 | 179 | 216 | 100 | 100 | 99.3 | 92.8 |
| **05b** | BCR-MAJOR | 1771687 | 196.9 | 159 | 188 | 232 | 100 | 100 | 99.4 | 93 |
| **05c** | BCR-MAJOR | 1743931 | 193.81 | 159 | 188 | 229 | 100 | 100 | 99.5 | 92 |
| **05d** | BCR-MAJOR | 1723298 | 191.52 | 160 | 186 | 222 | 100 | 100 | 99.3 | 92.1 |
| **06a** | BCR-MAJOR | 1179753 | 131.11 | 106 | 125 | 153 | 100 | 100 | 97.8 | 80.6 |
| **06b** | BCR-MAJOR | 1081331 | 120.17 | 96 | 113 | 146 | 100 | 99.1 | 96.6 | 70 |
| **06c** | BCR-MAJOR | 1065464 | 118.41 | 93 | 114 | 140 | 100 | 99.3 | 95.2 | 67.7 |
| **06d** | BCR-MAJOR | 1069290 | 118.84 | 97 | 114 | 138 | 100 | 99.5 | 96.1 | 70.4 |
| **07a** | BCR-MAJOR | 1394863 | 155.02 | 124 | 152 | 184 | 100 | 100 | 97.7 | 86 |
| **07b** | BCR-MAJOR | 1269227 | 141.06 | 109 | 143 | 178 | 100 | 100 | 97.7 | 80.8 |
| **07c** | BCR-MAJOR | 1253465 | 139.3 | 109 | 136 | 173 | 100 | 100 | 97.6 | 81.2 |
| **07d** | BCR-MAJOR | 1262030 | 140.26 | 109 | 139 | 175 | 100 | 100 | 98 | 81.1 |
| **08a** | BCR-MAJOR | 1622035 | 180.27 | 149 | 174 | 212 | 100 | 100 | 99 | 93.4 |
| **08b** | BCR-MAJOR | 1383234 | 153.73 | 120 | 149 | 186 | 100 | 99.9 | 98.1 | 86.6 |
| **08c** | BCR-MAJOR | 1383015 | 153.7 | 120 | 150 | 183 | 100 | 100 | 97.7 | 87.5 |
| **08d** | BCR-MAJOR | 1386669 | 154.11 | 122 | 151 | 188 | 100 | 100 | 98.3 | 86.6 |
| **09a** | BCR-MAJOR | 1198975 | 133.25 | 106 | 129 | 159 | 100 | 100 | 98.1 | 79.9 |
| **09b** | BCR-MAJOR | 1267942 | 140.91 | 107 | 138 | 171 | 100 | 100 | 98.5 | 81.6 |
| **09c** | BCR-MAJOR | 1259015 | 139.92 | 108 | 138 | 170 | 100 | 100 | 98.2 | 80.1 |
| **09d** | BCR-MAJOR | 1199192 | 133.27 | 106 | 128 | 160 | 100 | 100 | 98.1 | 79.8 |
| **10a** | BCR-MAJOR | 1545958 | 171.81 | 139 | 166 | 206 | 100 | 100 | 97.7 | 89.9 |
| **10b** | BCR-MAJOR | 1603633 | 178.22 | 147 | 171 | 211 | 100 | 100 | 98.3 | 91.8 |
| **10c** | BCR-MAJOR | 1617798 | 179.8 | 146 | 173 | 214 | 100 | 100 | 97.6 | 93.4 |
| **10d** | BCR-MAJOR | 1543451 | 171.53 | 136 | 162 | 209 | 100 | 100 | 98.2 | 90.2 |

Table 28. **major-BCR lane-level capturing**

*The table shows the coverage and capturing statistics related to ABL1 targeted region across lanes. From left to right column: sample and lane, region name, number of bp that cover the region, average coverage, coverage at 1$^{st}$ quartile (Q1), median coverage, coverage at 3$^{rd}$ quartile (Q3), percentage of targeted region captured with an average coverage greater or equal to 1X,20X,50X and 100X. Data included in this table are plotted in* **Figure 21***.*

| Sample | Target | total cov | average cov | Q1 | median | Q3 | 1X | 20X | 50X | 100X |
|---|---|---|---|---|---|---|---|---|---|---|
| **02a** | BCR-MINOR | 10990319 | 150.51 | 96 | 157 | 213 | 98.2 | 92.8 | 87.8 | 73.7 |
| **02b** | BCR-MINOR | 9961355 | 136.42 | 86 | 140 | 194 | 98.1 | 92.5 | 86.6 | 68.5 |
| **02c** | BCR-MINOR | 10082112 | 138.07 | 87 | 144 | 196 | 98.1 | 92.3 | 86.6 | 69.3 |
| **02d** | BCR-MINOR | 10061237 | 137.79 | 89 | 141 | 195 | 98.4 | 92.3 | 86.6 | 70.4 |
| **03a** | BCR-MINOR | 9596364 | 131.42 | 78 | 132 | 189 | 96.3 | 90.5 | 83.5 | 66 |
| **03b** | BCR-MINOR | 10020216 | 137.23 | 83 | 137 | 197 | 96.8 | 90.5 | 84.4 | 67.4 |
| **03c** | BCR-MINOR | 9949180 | 136.25 | 83 | 137 | 193 | 96.7 | 90.4 | 84.3 | 67.8 |
| **03d** | BCR-MINOR | 9666079 | 132.38 | 80 | 133 | 191 | 96.5 | 90.4 | 83.9 | 66 |
| **04a** | BCR-MINOR | 9610234 | 131.61 | 84 | 136 | 186 | 96.3 | 90.8 | 85.1 | 68.2 |
| **04b** | BCR-MINOR | 10139793 | 138.86 | 87 | 144 | 197 | 97.6 | 90.8 | 85.8 | 69.3 |
| **04c** | BCR-MINOR | 10038410 | 137.47 | 87 | 142 | 193 | 96.9 | 90.9 | 85.4 | 69.9 |
| **04d** | BCR-MINOR | 9857992 | 135 | 87 | 140 | 190 | 96.5 | 91 | 85.2 | 69.1 |
| **05a** | BCR-MINOR | 10549734 | 144.48 | 86 | 145 | 202 | 96.4 | 90.8 | 85.1 | 69.3 |
| **05b** | BCR-MINOR | 11151755 | 152.72 | 91 | 152 | 215 | 97 | 90.9 | 85.8 | 72.4 |
| **05c** | BCR-MINOR | 11012295 | 150.81 | 90 | 150 | 213 | 96.5 | 90.7 | 85.7 | 71.9 |
| **05d** | BCR-MINOR | 10738066 | 147.06 | 88 | 147 | 206 | 96.2 | 90.7 | 86 | 71.4 |
| **06a** | BCR-MINOR | 7695616 | 105.39 | 63 | 104 | 145 | 98.8 | 91.7 | 81.8 | 51.9 |
| **06b** | BCR-MINOR | 7124091 | 97.56 | 59 | 96 | 135 | 97.7 | 91.2 | 80.1 | 47.1 |
| **06c** | BCR-MINOR | 7133764 | 97.7 | 59 | 96 | 134 | 98 | 91.3 | 80 | 47 |
| **06d** | BCR-MINOR | 7130775 | 97.66 | 60 | 95 | 135 | 98.1 | 91.5 | 80.8 | 46.1 |
| **07a** | BCR-MINOR | 10054197 | 137.69 | 88 | 144 | 194 | 98 | 92 | 86.5 | 70.2 |
| **07b** | BCR-MINOR | 9263296 | 126.86 | 80 | 129 | 180 | 97.3 | 91.5 | 85.3 | 65.2 |
| **07c** | BCR-MINOR | 9262246 | 126.85 | 81 | 130 | 180 | 97.5 | 91.4 | 85 | 65.6 |
| **07d** | BCR-MINOR | 9349568 | 128.04 | 82 | 132 | 181 | 97.9 | 91.5 | 85.4 | 66.6 |
| **08a** | BCR-MINOR | 10411220 | 142.58 | 92 | 144 | 197 | 97.5 | 92 | 86.6 | 71.8 |
| **08b** | BCR-MINOR | 9647763 | 132.12 | 83 | 133 | 186 | 98.2 | 91.7 | 85.6 | 67.1 |
| **08c** | BCR-MINOR | 9616149 | 131.69 | 83 | 134 | 184 | 98.8 | 91.5 | 86.2 | 67.7 |
| **08d** | BCR-MINOR | 9522240 | 130.41 | 82 | 131 | 184 | 98.5 | 91.6 | 85.2 | 67 |
| **09a** | BCR-MINOR | 8302666 | 113.7 | 72 | 116 | 158 | 97.8 | 91.7 | 84 | 59.7 |
| **09b** | BCR-MINOR | 8690212 | 119.01 | 77 | 120 | 165 | 97.9 | 91.8 | 84.8 | 62.8 |
| **09c** | BCR-MINOR | 8550245 | 117.09 | 77 | 118 | 162 | 98.6 | 91.6 | 84.5 | 62.7 |
| **09d** | BCR-MINOR | 8404465 | 115.1 | 74 | 116 | 159 | 98.3 | 91.7 | 84.1 | 60.6 |
| **10a** | BCR-MINOR | 9919724 | 135.85 | 84 | 134 | 189 | 97.6 | 92.1 | 85.9 | 67.7 |
| **10b** | BCR-MINOR | 10386588 | 142.24 | 88 | 139 | 197 | 98.7 | 91.8 | 86.4 | 69.8 |
| **10c** | BCR-MINOR | 10321866 | 141.36 | 86 | 141 | 195 | 97.7 | 91.8 | 86.4 | 69 |
| **10d** | BCR-MINOR | 9993624 | 136.86 | 85 | 135 | 190 | 97.8 | 91.8 | 86.2 | 67.9 |

Table 29. **minor-BCR lane-level capturing**

*The table shows the coverage and capturing statistics related to ABL1 targeted region across lanes. From left to right column: sample and lane, region name, number of bp that cover the region, average coverage, coverage at 1$^{st}$ quartile (Q1), median coverage, coverage at 3$^{rd}$ quartile (Q3), percentage of targeted region captured with an average coverage greater or equal to 1X,20X,50X and 100X. Data included in this table are plotted in* **Figure 23**.

| Sample | Target | Total cov | Average cov | Q1 | median | Q3 | 1X | 20X | 50X | 100X |
|---|---|---|---|---|---|---|---|---|---|---|
| **02a** | BCR-MICRO | 1208070 | 156.63 | 119 | 152 | 188 | 100 | 100 | 96.6 | 86.4 |
| **02b** | BCR-MICRO | 1041932 | 135.09 | 98 | 134 | 163 | 100 | 99.4 | 95.7 | 72.9 |
| **02c** | BCR-MICRO | 1026152 | 133.04 | 97 | 130 | 162 | 100 | 99.2 | 95.2 | 72.9 |
| **02d** | BCR-MICRO | 1022616 | 132.58 | 96 | 130 | 164 | 100 | 98.8 | 95.7 | 71.6 |
| **03a** | BCR-MICRO | 1061446 | 137.62 | 90 | 141 | 172 | 100 | 98.7 | 95 | 69.2 |
| **03b** | BCR-MICRO | 1105325 | 143.31 | 98 | 141 | 181 | 100 | 98.8 | 94 | 73.4 |
| **03c** | BCR-MICRO | 1120843 | 145.32 | 97 | 147 | 188 | 100 | 99.2 | 94.6 | 73.4 |
| **03d** | BCR-MICRO | 1073728 | 139.21 | 96 | 138 | 169 | 100 | 99.7 | 93.8 | 73 |
| **04a** | BCR-MICRO | 1149866 | 149.08 | 109 | 148 | 187 | 100 | 98.3 | 94.7 | 80 |
| **04b** | BCR-MICRO | 1204271 | 156.14 | 118 | 153 | 190 | 100 | 100 | 96.3 | 84.1 |
| **04c** | BCR-MICRO | 1199496 | 155.52 | 113 | 154 | 195 | 100 | 98.5 | 94.9 | 82.1 |
| **04d** | BCR-MICRO | 1185594 | 153.71 | 115 | 155 | 188 | 100 | 100 | 95.7 | 82.4 |
| **05a** | BCR-MICRO | 1089888 | 141.31 | 94 | 137 | 179 | 100 | 99.1 | 94.3 | 72.3 |
| **05b** | BCR-MICRO | 1133742 | 146.99 | 99 | 143 | 180 | 100 | 99.6 | 96.3 | 74.6 |
| **05c** | BCR-MICRO | 1131154 | 146.66 | 105 | 147 | 179 | 100 | 99.5 | 96 | 76.8 |
| **05d** | BCR-MICRO | 1070474 | 138.79 | 91 | 137 | 168 | 100 | 100 | 94.8 | 72.9 |
| **06a** | BCR-MICRO | 908549 | 117.79 | 82 | 116 | 146 | 100 | 100 | 93.1 | 63.1 |
| **06b** | BCR-MICRO | 834879 | 108.24 | 71 | 110 | 138 | 100 | 100 | 88.9 | 56.9 |
| **06c** | BCR-MICRO | 827638 | 107.3 | 75 | 106 | 137 | 100 | 99.5 | 87.7 | 54 |
| **06d** | BCR-MICRO | 803957 | 104.23 | 72 | 104 | 133 | 100 | 100 | 88.7 | 53.1 |
| **07a** | BCR-MICRO | 1075784 | 139.48 | 96 | 134 | 174 | 100 | 100 | 95.7 | 72.5 |
| **07b** | BCR-MICRO | 951647 | 123.38 | 83 | 121 | 158 | 100 | 99 | 90.4 | 67.5 |
| **07c** | BCR-MICRO | 933486 | 121.03 | 81 | 120 | 150 | 100 | 98.9 | 90.2 | 65.9 |
| **07d** | BCR-MICRO | 955228 | 123.85 | 83 | 122 | 154 | 100 | 99.3 | 91.6 | 64.1 |
| **08a** | BCR-MICRO | 1177187 | 152.62 | 102 | 156 | 193 | 100 | 99.8 | 96 | 76.2 |
| **08b** | BCR-MICRO | 1007750 | 130.66 | 87 | 135 | 169 | 100 | 99.3 | 94 | 68.2 |
| **08c** | BCR-MICRO | 1028521 | 133.35 | 94 | 135 | 167 | 100 | 99.8 | 95.6 | 71.4 |
| **08d** | BCR-MICRO | 1033451 | 133.99 | 91 | 136 | 175 | 100 | 99.2 | 95.1 | 69.7 |
| **09a** | BCR-MICRO | 945414 | 122.57 | 84 | 126 | 156 | 100 | 97.8 | 85.5 | 67.4 |
| **09b** | BCR-MICRO | 968251 | 125.53 | 88 | 129 | 156 | 100 | 97.9 | 86.7 | 69.8 |
| **09c** | BCR-MICRO | 975760 | 126.51 | 85 | 131 | 159 | 100 | 98.7 | 86.3 | 67.9 |
| **09d** | BCR-MICRO | 945139 | 122.54 | 88 | 126 | 154 | 100 | 98.4 | 87 | 67.1 |
| **10a** | BCR-MICRO | 982413 | 127.37 | 84 | 130 | 167 | 100 | 94.2 | 86.4 | 68.9 |
| **10b** | BCR-MICRO | 1016777 | 131.83 | 89 | 134 | 177 | 99.3 | 93.7 | 87.4 | 68.4 |
| **10c** | BCR-MICRO | 1011911 | 131.2 | 87 | 138 | 174 | 100 | 93.8 | 86.4 | 68.6 |
| **10d** | BCR-MICRO | 989168 | 128.25 | 87 | 132 | 172 | 100 | 94.2 | 86.8 | 68.3 |

Table 30. **micro-BCR lane-level capturing**

*The table shows the coverage and capturing statistics related to ABL1 targeted region across lanes. From left to right column: sample and lane, region name, number of bp that cover the region, average coverage, coverage at 1$^{st}$ quartile (Q1), median coverage, coverage at 3$^{rd}$ quartile (Q3), percentage of targeted region captured with an average coverage greater or equal to 1X,20X,50X and 100X. Data included in this table are plotted in* **Figure 22**.

| Sample Lane | SV_type | type | pairs | bp1_start1-end1 (chr9) | bp2_start2-end2 (chr22) |
|---|---|---|---|---|---|
| **1a** | TRANSLOC | UNBAL | 270 | 133606918-133607140 | 23632762-23632892 |
| **2a** | TRANSLOC | UNBAL | 43 | 133648580-133648887 | 23631908-23632079 |
| **4a** | INS_FRAGMT | BAL | 91 | 133590218-133590566133590575-133590896 | 23575267-23575207 |
| **4a** | TRANSLOC | BAL | 80 | 133590575-133590566 | 23575267-23575207 |
| **6a** | TRANSLOC | UNBAL | 60 | 133657725-133658098 | 23634042-23634274 |
| **7a** | TRANSLOC | UNBAL | 49 | 133683996-133684302 | 23634584-23634859 |
| **7a** | INV_TRANSLOC | UNBAL | 47 | 133684244-133684515 | 36900842-36901057 |
| **9a** | TRANSLOC | UNBAL | 35 | 133663802-133663988 | 23631896-23632112 |
| **10a** | TRANSLOC | BAL | 32 | 133696775-133697006 | 23634479-23634637 |

| Sample Lane | SV_type | type | pairs | bp1_start1-end1 | bp2_start2-end2 |
|---|---|---|---|---|---|
| **1b** | TRANSLOC | UNBAL | 253 | 133606948-133607140 | 23632757-23632890 |
| **2b** | TRANSLOC | UNBAL | 35 | 133648594-133648885 | 23631883-23632051 |
| **4b** | INS_FRAGMT | BAL | 86 | 133590333-133590561 - 133590576-133590916 | 23575268-23575209 |
| **4b** | INS_FRAGMT | BAL | 81 | 133590206-133590561 - 133590576-133590829 | 23575251-23575209 |
| **6b** | TRANSLOC | UNBAL | 64 | 133657743-133658098 | 23634051-23634315 |
| **7b** | TRANSLOC | UNBAL | 53 | 133684061-133684302 | 23634592-23634863 |
| **7b** | INV_TRANSLOC | UNBAL | 49 | 133684247-133684481 | 36900847-36901145 |
| **9b** | INV_TRANSLOC | UNBAL | 30 | 136985435-136985697 | 23631670-23631890 |
| **9b** | TRANSLOC | UNBAL | 29 | 133663780-133663990 | 23631882-23632117 |
| **10b** | TRANSLOC | UNBAL | 30 | 133696781-133697038 | 23634348-23634646 |
| **10b** | TRANSLOC | UNBAL | 21 | 133696693-133697006 | 23634577-23634793 |

Table 31. **SVDetect breakpoints detection: A and B lanes**

*Table shows the breakpoints coordinates identified by discordant pair reads using SVDetect. The columns describe: patient and lane number, SV type, balancing, number of pairs supporting the SV event, start and end of breakpoints coordinates (chromosome 9 and chromosome 22).*

| Sample Lane | SV_type | type | pairs | bp1_start1-end1 | bp2_start2-end2 |
|---|---|---|---|---|---|
| **1c** | TRANSLOC | UNBAL | 273 | 133606917-133607140 | 23632756-23632890 |
| **2c** | TRANSLOC | UNBAL | 29 | 133648582-133648884 | 23631918-23632210 |
| **4c** | INS_FRAGMT | BAL | 85 | 133590292-133590563 - 133590576-133590916 | 23575267-23575207 |
| **4c** | TRANSLOC | BAL | 76 | 133590582-133590572 | 23575267-23575207 |
| **5c** | INS_FRAGMT | BAL | 21 | 133708178-133708501 - 133708502-133708895 | 23654742-23654792 |
| **6c** | TRANSLOC | UNBAL | 57 | 133657679-133658098 | 23634042-23634181 |
| **7c** | INV_TRANSLOC | UNBAL | 52 | 133684252-133684542 | 36900846-36901211 |
| **7c** | TRANSLOC | UNBAL | 41 | 133683951-133684306 | 23634592-23634979 |
| **9c** | TRANSLOC | UNBAL | 37 | 133663859-133663988 | 23631897-23631955 |
| **9c** | INV_TRANSLOC | UNBAL | 33 | 136985532-136985698 | 23631680-23631884 |
| **10c** | TRANSLOC | UNBAL | 37 | 133696775-133696978 | 23634410-23634637 |
| **10c** | TRANSLOC | BAL | 32 | 133696775-133697005 | 23634504-23634640 |

| Sample Lane | SV_type | type | pairs | bp1_start1-end1 | bp2_start2-end2 |
|---|---|---|---|---|---|
| **1d** | TRANSLOC | UNBAL | 274 | 133606908-133607140 | 23632757-23632916 |
| **2d** | TRANSLOC | UNBAL | 39 | 133648547-133648886 | 23631930-23632155 |
| **4d** | TRANSLOC | BAL | 62 | 133590569-133590563 | 23575249-23575214 |
| **4d** | INS_FRAGMT | BAL | 59 | 133590249 -133590561 - 133590561-133590908 | 23575249-23575214 |
| **5d** | TRANSLOC | BAL | 24 | 133708451-133708500 | 23654734-23654777 |
| **5d** | TRANSLOC | BAL | 21 | 133708502-133708500 | 23654734-23654777 |
| **6d** | TRANSLOC | UNBAL | 46 | 133657686-133658101 | 23634042-23634259 |
| **7d** | TRANSLOC | UNBAL | 70 | 133684020-133684300 | 23634602-23634852 |
| **7d** | INV_TRANSLOC | UNBAL | 37 | 133684242-133684530 | 36900845-36901164 |
| **9d** | INV_TRANSLOC | UNBAL | 38 | 136985515-136985697 | 23631698-23631905 |
| **9d** | TRANSLOC | UNBAL | 31 | 133663703-133663995 | 23631873-23632051 |
| **10d** | TRANSLOC | UNBAL | 31 | 133696775-133696992 | 23634342-23634635 |
| **10d** | TRANSLOC | BAL | 21 | 133696775-133696999 | 23634569-23634645 |

Table 32. **SVDetect breakpoints detection: C and D lanes**

*Table shows the breakpoints coordinates identified by discordant pair reads using SVDetect. The columns describe: patient and lane number, SV type, balancing, number of pairs supporting the SV event, start and end of breakpoints coordinates (chromosome 9 and chromosome 22).*

| Chr | Start | End | Name | K562 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | S | E | S | E | S | E | S | E | S | E |
| **9** | 133590053 | 133590178 | MIR3 | | | | | | | 518 | 393 | | |
| **9** | 133590241 | 133590294 | Charlie4z | | | | | | | 330 | 277 | | |
| **9** | 133593422 | 133593728 | AluSz | | | | | 406 | 100 | | | | |
| **9** | 133593728 | 133593794 | Charlie1a | | | | | 100 | 34 | | | | |
| **9** | 133593834 | 133593985 | Charlie1a | | | | | -6 | -157 | | | | |
| **9** | 133593985 | 133594276 | AluJo | | | | | -157 | -448 | | | | |
| **9** | 133606692 | 133607233 | **L1ME3A** | **453** | **-88** | | | | | | | | |
| **9** | 133607299 | 133607432 | L1MD | -154 | -287 | | | | | | | | |
| **9** | 133648425 | 133648741 | AluSp | | | 466 | 150 | | | | | | |
| **9** | 133648741 | 133648826 | L1MEd | | | 150 | 65 | | | | | | |
| **9** | 133648839 | 133649166 | **AluJb** | | | **52** | **-275** | | | | | | |
| **9** | 133649174 | 133649743 | L1ME1 | | | -283 | -852 | | | | | | |
| **9** | 133707880 | 133708012 | MIR | | | | | | | | | | 484 |

Table 33. **BP4 distance to repeated elements: samples 2,3,4,5 and K562 cell lines.**
*The table shows the distance of BP4 (which spots the BCR-ABL1 fusion at ABL1) to repeated elements in a region of 800bp (400 upstream and 400 downstream) surrounding the breakpoint. From left to right: the chromosome, start and end of the repeated element, the name of the repeated element, samples column (S indicates the start of the repeated element and E the distance from the end of the repeated element). Bold and underline numbers denote that the BP4 in that sample maps within the repeated element. In this case sample 2 and K562 cell line have the BP4 that maps within AluJb and L1ME3A respectively.*

| Chr | Start | End | Name | 6 | | 7 | | 8 | | 9 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | S | E | S | E | S | E | S | E | S | E |
| **9** | 133657438 | 133657747 | AluSz | 664 | 355 | | | | | | | | |
| **9** | 133657747 | 133657844 | MER112 | 355 | 258 | | | | | | | | |
| **9** | 133657863 | 133658070 | MIRb | 239 | 32 | | | | | | | | |
| **9** | 133658393 | 133658520 | L1MC4 | -291 | -418 | | | | | | | | |
| **9** | 133663584 | 133664351 | **L1MC4** | | | | | | | **411** | **-356** | | |
| **9** | 133664415 | 133664487 | L1MC4 | | | | | | | -420 | -492 | | |
| **9** | 133683619 | 133683929 | AluSx | | | 691 | 381 | | | | | | |
| **9** | 133683947 | 133683994 | (TAGG)n | | | 363 | 316 | | | | | | |
| **9** | 133684052 | 133684161 | L2 | | | 258 | 149 | | | | | | |
| **9** | 133684218 | 133684749 | **L1MEc** | | | **92** | **-439** | | | | | | |
| **9** | 133694459 | 133694769 | AluSg | | | | | 458 | 148 | | | | |
| **9** | 133694919 | 133695226 | AluSx3 | | | | | -2 | -309 | | | | |
| **9** | 133696312 | 133696616 | AluSc8 | | | | | | | | | 689 | 385 |
| **9** | 133696616 | 133697051 | **L1ME1** | | | | | | | | | **385** | **-50** |
| **9** | 133697051 | 133697227 | AluSx3 | | | | | | | | | -50 | -226 |

Table 34. **BP4 distance to repeated elements: patients 6,7,9,10**

*The table shows the distance of BP4 (which spots the BCR-ABL1 fusion at ABL1) to repeated elements in a region of 800bp (400 upstream and 400 downstream) surrounding the breakpoint. From left to right: the chromosome, start and end of the repeated element, the name of the repeated element, samples column (S indicates the start of the repeated element and E the distance from the end of the repeated element). Bold and underline numbers denote that the BP4 in that sample map within the repeated element. In this case sample 7,9 and 10 have the BP4 that maps within L1MEc, L1MC4 and L1ME1 respectively.*

| chr | Start | End | Name | 6 S | 6 E | 7 S | 7 E | 8 S | 8 E | 10 S | 10 E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **22** | 23633112 | 23633411 | AluSx1 | | | | | 299 | 0 | | |
| **22** | 23633413 | 23633642 | L1ME1 | | | | | -2 | -231 | | |
| **22** | 23633846 | 23634014 | L2b | 187 | 19 | | | | | | |
| **22** | 23634391 | 23634475 | L2b | | | 187 | 103 | | | 182 | 98 |
| **22** | 23634614 | 23634637 | AT_rich | | | -36 | -59 | | | -41 | -64 |

Table 35. **BP1 distance to repeated elements**

*The table shows the distance of BP1 (which spots the BCR-ABL1 fusion at BCR) to repeated elements in a region of 800bp (400 upstream and 400 downstream) surrounding the breakpoint. From left to right: the chromosome, start and end of the repeated element, the name of the repeated element, samples column (S indicates the start of the repeated element and E the distance from the end of the repeated element).*

| chr | Start | End | Name | K562 S | K562 E | 2 S | 2 E | 5 S | 5 E | 7 S | 7 E | 8 S | 8 E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **22** | 23633112 | 23633411 | AluSx1 | -138 | -437 | 444 | 145 | | | -204 | -503 | | |
| **22** | 23633413 | 23633642 | L1ME1 | | | **143** | **-86** | | | | | | |
| **22** | 23633846 | 23634014 | L2b | | | -290 | -458 | | | | | | |
| **22** | 23634391 | 23634475 | L2b | | | | | 248 | 164 | | | 249 | 165 |
| **22** | 23634614 | 23634637 | AT_rich | | | | | 25 | 2 | | | 26 | 3 |

Table 36. **BP2 distance to repeated elements**

*The table shows the distance of BP2 (which spots the ABL1-BCR fusion at BCR) to repeated elements in a region of 800bp (400 upstream and 400 downstream) surrounding the breakpoint. From left to right: the chromosome, start and end of the repeated element, the name of the repeated element, samples column (S indicates the start of the repeated element and E the distance from the end of the repeated element). Bold and underline numbers denote that the BP4 in that sample map within the repeated element. For instance sample 2 has the BP2 that maps within a L1ME1.*

| chr | Start | End | Name | 2 S | 2 E | 3 S | 3 E | 4 S | 4 E | 6 S | 6 E | 8 S | 8 E | 10 S | 10 E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9** | 133590241 | 133590294 | Charlie4z | | | | | 328 | 275 | | | | | | |
| **9** | 133593422 | 133593728 | AluSz | | | 382 | 76 | | | | | | | | |
| **9** | 133593728 | 133593794 | Charlie1a | | | 76 | 10 | | | | | | | | |
| **9** | 133593834 | 133593985 | Charlie1a | | | -30 | -181 | | | | | | | | |
| **9** | 133593985 | 133594276 | AluJo | | | -181 | -472 | | | | | | | | |
| **9** | 133648425 | 133648741 | AluSp | 356 | 40 | | | | | | | | | | |
| **9** | 133648741 | 133648826 | **L1MEd** | **40** | **-45** | | | | | | | | | | |
| **9** | 133648839 | 133649166 | AluJb | -58 | -385 | | | | | | | | | | |
| **9** | 133657863 | 133658070 | MIRb | | | | | | | 447 | 240 | | | | |
| **9** | 133658393 | 133658520 | L1MC4 | | | | | | | -83 | -210 | | | | |
| **9** | 133694459 | 133694769 | AluSg | | | | | | | | | 530 | 220 | | |
| **9** | 133694919 | 133695226 | **AluSx3** | | | | | | | | | **70** | **-237** | | |
| **9** | 133696195 | 133696312 | L1ME1 | | | | | | | | | | | | |
| **9** | 133696312 | 133696616 | AluSc8 | | | | | | | | | | | 469 | 165 |
| **9** | 133696616 | 133697051 | **L1ME1** | | | | | | | | | | | **165** | **-270** |
| **9** | 133697051 | 133697227 | AluSx | | | | | | | | | | | -270 | -446 |

Table 37. **BP3 distance to repeated elements**

*The table shows the distance of BP3 (which spots the ABL1-BCR fusion at ABL1) to repeated elements in a region of 800bp (400 upstream and 400 downstream) surrounding the breakpoint. From left to right: the chromosome, start and end of the repeated element, the name of the repeated element, samples column (S indicates the start of the repeated element and E the distance from the end of the repeated element). Bold and underline numbers denote that the BP3 in that sample map within the repeated element. In this case sample 2,8 and 10 have the BP3 that maps within L1MEd, AluSx3 and L1ME1 respectively.*

| Chr | Exon | Start | End | 4 S | 4 E | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 1 | 23523147 | 23524426 | 52101 | **50822** | | | | | | | | |
| 22 | 2 | 23595985 | 23596167 | **-20737** | -20919 | | | | | | | | |

| Chr | Exon | Start | End | 2 S | 2 E | 3 S | 3 E | 9 S | 9 E | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 13 | 23631703 | 23631808 | 207 | **102** | 425 | **320** | 172 | **67** | | | | |
| 22 | 14 | 23632525 | 23632600 | **-615** | -690 | **-397** | -472 | **-650** | -725 | | | | |

| Chr | Exon | Start | End | K562 S | K562 E | 6 S | 6 E | 7 S | 7 E | 8 S | 8 E | 10 S | 10 E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 14 | 23632525 | 23632600 | 217 | **142** | 1508 | **1433** | 2053 | **1978** | 886 | **811** | 2048 | **1973** |
| 22 | 15 | 23634727 | 23634825 | **-1985** | -2083 | **-694** | -792 | **-149** | -247 | **-1316** | -1414 | **-154** | -252 |

| Chr | Exon | Start | End | 5 S | 5 E | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 19 | 23653883 | 23654023 | 882 | **742** | | | | | | | | |
| 22 | 20 | 23655073 | 23655208 | **-308** | -443 | | | | | | | | |

Table 38. **BP1 distance to BCR exons**

*This tables summarize the distances between BP1 and flanking exons. The first 4 columns describe the chromosome, the exon, and the start and end coordinates of the exon. The further columns represent the distance of the BP1 to the flanking exons. The first one describes the patient 4 whose BP1 maps at the m-BCR between 1$^{st}$ and 2$^{nd}$ exon, the second table describes distances of BP1 to flanking exons in patients that carry the b2a2 BCR-ABL1 transcript, in the third the distances in patients that carry the b2a3 isoform and in the last one patient 5 which has the BP1 located in the micro-BCR region between exon 19 and 20.*

## PUBLICATIONS

1. Salvi E, Kuznetsova T, Thijs L, Lupoli S, Stolarz-Skrzypek K, D'Avila F, Tikhonoff V, De Astis S, Barcella M, Seidlerová J, Benaglio P, Malyutina S, Frau F, Velayutham D, Benfante R, Zagato L, Title A, Braga D, Marek D, Kawecka-Jaszcz K, Casiglia E, Filipovsky J, Nikitin Y, Rivolta C, Manunta P, Beckmann JS, Barlassina C, Cusi D, Staessen JA. *Target sequencing, cell experiments, and a population study establish endothelial nitric oxide synthase (eNOS) gene as hypertension susceptibility gene*. Hypertension. 2013, 62(5), 844-52.

2. Ilaria Stefania Pagani, Orietta Spinelli, Cristina Pirrone, Diana Pigini, Sara Lupoli, Francesca D'Avila, Matteo Barcella, Chiara Boroni, Tamara Intermesoli, Ursula Giussani, Federica Bolda, Francesco Pasquali, Francesco Lo Curto, Arnalda Lanfranchi, Fulvio Porta, Cristina Barlassina, Alessandro Rambaldi, Giovanni Porta. *Quantitative DNA Real-Time PCR Compared to mRNA and Cytogenetic assays to Monitor Minimal Residual Disease in Chronic Myeloid Leukemia*. Presented at 55th American Society of Hematology (ASH) Annual Meeting and Exposition, Saturday, December 7, 2013, New Orleans, LA (USA), Abstract 1316.

## ACKNOWLEDGEMENT

I am grateful to Prof. Mario Clerici for the opportunity to perform my PhD work at University of Milan.

I express my deep gratitude to my tutor, Prof. Cristina Barlassina for her continuous support and guidance throughout my research work.

I am very thankful to Prof. Giovanni Porta and his research group for giving me the opportunity to collaborate with them to this research project.

I would also like to thanks PhD Sara Lupoli and PhD Francesca D'Avila, for their friendly help during my work.

I express my sincere thanks to all those with whom I worked at the platform of Genomics and Bioinformatics at Filarete Foundation during these three years, in particular I want to thanks: PhD Dinesh Velayutham and Dr Daniele Braga.

My thanks to Ing. Maurizio Mercurio for his IT support and his enjoyable company every day.

Finally sincerely thanks to my family and to my girlfriend Roberta for their lovely support.

At last my gratitude goes to Prof. Daniele Cusi and the University of Milan for the financial support.